MARKET SCAN REPORT

AUGUST 2025

BY INFOSYS TOPAZ RESPONSIBLE AI OFFICE

NEW VERSION RELEASED

INFOSYS RESPONSIBLE AI TOOLKIT 2.2.0

IN FOCUS

RESPONSIBLE AI: RISK-MANAGED CO-CREATION FOR A CO-INTELLIGENT FUTURE

By B. Ravindran and Krishnan Narayanan

Infosys® Navigate your next

Foreword

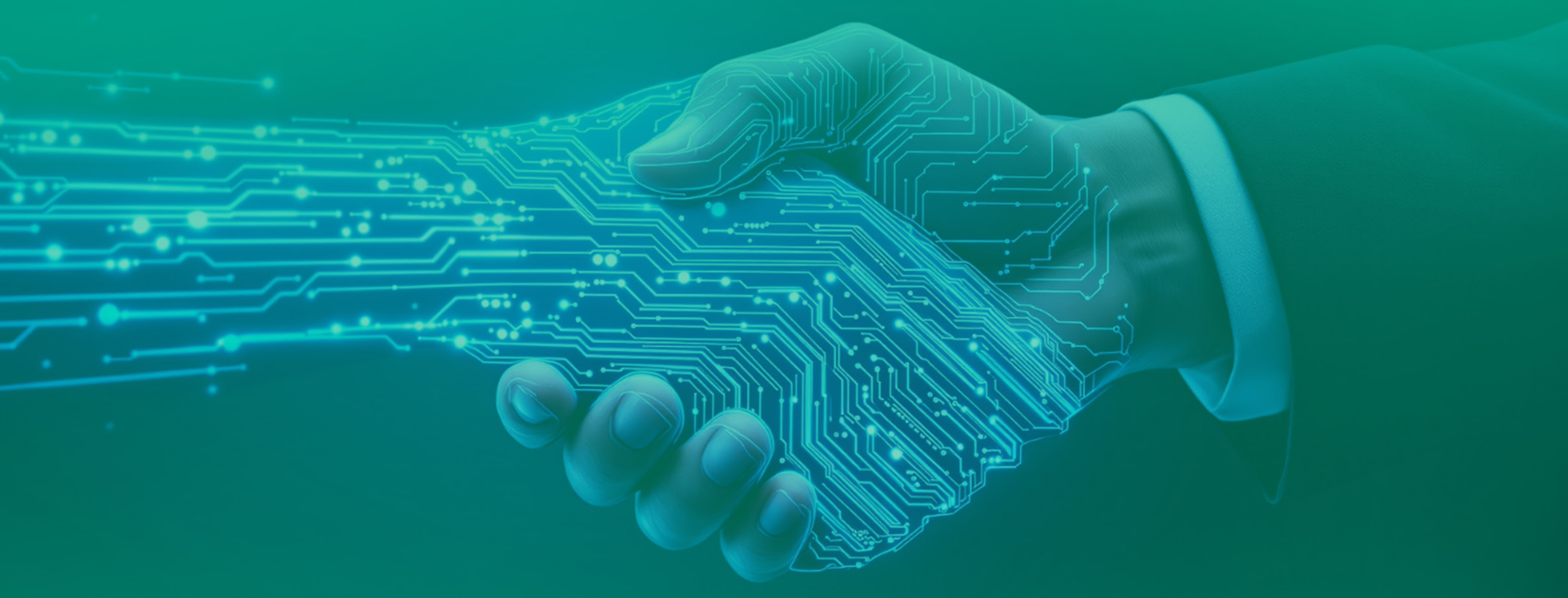
"August 2025 has shown both the opportunities and responsibilities that come with AI. GPT-5 continues to amaze with its capabilities, yet incidents such as Google inadvertently indexing shared ChatGPT conversations remind us that privacy, safety, and accountability are more important than ever. In my recent article, ***"Responsible AI is a "VALUE Conversation", NOT a "Cost Conversation",*** I highlighted that the real power of responsible AI comes from thoughtful discussion about its use, risks, and impact on society. These conversations help us ensure technology grows in a way that aligns with human values.

This focus on responsibility is reflected in regulatory developments, including India's FREE (Framework for Responsible and Ethical Enablement of Artificial Intelligence) Committee Report for the financial services sector, which offers practical guidance for ethical AI adoption. At the same time, insights from B. Ravindran, Head of the Wadhwani School of Data Science and AI, IIT Madras, and Krishnan Narayanan, Co-founder and President of itihaasa Research and Digital, remind us how important it is to design technology that adapts to people's needs while staying safe and trustworthy. We thank them for sharing their perspectives in this edition.

We are also excited to share the release of version 2.2.0 of the Infosys Responsible AI Toolkit, which makes it easier for organizations to adopt AI responsibly and at scale. I am grateful to all contributors for their valuable insights, which make this edition a meaningful guide for using AI in ways that truly benefit society."

May this edition inspire you to think critically, act responsibly, and contribute to building AI that truly benefits society.

Syed Ahmed
Global Head
Infosys Responsible AI Office



From the editor's desk

AI at the Crossroads of Trust and Transformation- August Edition

August reminded us that AI's story is as much about governance and trust as it is about progress. Headlines this month cut across innovation, regulation, and controversy—showing why Responsible AI is no longer optional.

Governance in Focus:

Regulation advanced globally: the EU tightened compliance, the U.S. pushed the TRAIN Act for AI transparency, Latin America and Asia deepened cooperation, and South Korea expanded AI R&D incentives. India also made strides with the FREE (Framework for Responsible and Ethical Enablement of Artificial Intelligence) Committee Report by the RBI (Reserve Bank of India), offering recommendations for the responsible and ethical use of AI in the financial services sector. Oversight is catching up, and organizations must align quickly.

Incidents Highlight Risks:

GPT-5 fueled excitement but also skepticism, keeping the debate between ambition and accountability alive. Vulnerabilities surfaced: Google inadvertently indexed shared ChatGPT conversations, autonomous driving liabilities hit courtrooms, and AI therapy bots were suspended over mental health risks. Even small features can scale into systemic risks when AI impacts real lives.

Research and Thought Leadership:

Innovation continues with resilience in focus. Tools like SDEval (Safety Dynamic Evaluation for multimodal large language models, or MLLMs), self-defense frameworks for LLMs, and benchmarks for high-risk dialogues signal a shift toward safe AI. Baidu's ERNIE 4.5 and DeepMind's Genie 3 show capability growth, but trust must keep pace.

In this edition's In Focus section, B. Ravindran, Professor and Head of the Wadhwani School of Data Science and AI, IIT Madras, and Krishnan Narayanan, Co-founder and President of itihaasa Research and Digital, explore "Responsible AI: Risk-Managed Co-Creation for a Co-Intelligent Future." Their reflections remind us that trustworthy AI grows with us, adapts to our values, and stays rooted in real human experience. We thank the professors for sharing their insights.

Toolkit Update:

We are happy to announce the release of version 2.2.0 of the Infosys Responsible AI Toolkit (Open Source), with exciting new features for Responsible AI implementation, making adoption safer and more scalable.

Reflect and Share:

As AI innovation accelerates, we leave you with a thought: how will you ensure the systems you create or use reflect human values and serve society responsibly? We'd also love to hear your feedback—what resonated in this edition, and how can we make future insights even more valuable for your AI journey?

Warm regards,

Ashish Tewari
Head- Infosys Responsible AI Office, India

Table of Contents

AI Regulations, Governance & Standards

AI Principles

In Focus

Technical Updates

Industry Updates

Infosys Developments

Contributors

AI Regulations, Governance & Standards

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

AI Regulations & Governance across the globe

Brazil and Ecuador Sign Strategic AI Cooperation Agreement to Advance Regional Innovation and Digital Sovereignty

Brazil and Ecuador have signed a strategic Memorandum of Understanding to collaborate on artificial intelligence development, focusing on joint research, professional training, and the use of high-performance computing infrastructure. The agreement was formalized during Ecuadorian President Daniel Noboa's official visit to Brazil, with Brazil's Minister of Science, Technology and Innovation, Luciana Santos, and Ecuador's Minister of Foreign Affairs, Gabriela Sommerfeld, leading the initiative. The partnership aligns with Brazil's National AI Plan, which supports 100 collaborative AI projects across Latin America and Africa with R$ 100 million in funding, and includes an additional R$ 50 million to support 30 projects through 2028. The agreement emphasizes the importance of regional technological sovereignty, responsible AI adoption, and the development of Latin American AI models that reflect local realities and benefit society.[1]

UAE, Malaysia, and Rwanda Sign Strategic AI Partnership to Empower the Global South Through Ethical Innovation and Talent Development

The United Arab Emirates, Malaysia, and Rwanda signed a strategic partnership to advance artificial intelligence (AI) adoption across the Global South. This agreement, formalized through a Memorandum of Understanding, builds upon the Centre for the Fourth Industrial Revolution (C4IR) AI Fellowship Program, which was initially launched by the UAE and Rwanda in 2024. By including Malaysia, the partnership aims to strengthen ethical AI governance, encourage the exchange of skilled professionals, and promote sustainable development through collaborative initiatives. The alliance reflects a shared commitment to using AI responsibly and inclusively to support economic and social progress in emerging regions.[2]

2025 APEC Digital and AI Ministerial Statement: Asia-Pacific Leaders Unite for Responsible AI and Inclusive Digital Growth

At the Asia-Pacific Economic Cooperation(APEC) summit in Incheon, South Korea, ministers from Asia-Pacific economies like USA, Russia, China, Taiwan jointly released the APEC Digital and AI Ministerial Statement, presenting a shared vision for using digital technologies and artificial intelligence (AI) to drive innovation and tackle social and economic challenges. The statement highlights the importance of adopting AI in a secure, responsible, and human-centered way to improve productivity, resilience, and economic development across the region. It encourages member countries to integrate AI and digital tools into education, workforce development, and lifelong learning, while promoting collaboration and information-sharing on digital policies. The goal is to strengthen regional cooperation, boost cross-border trade and investment, and ensure that emerging technologies benefit all people fairly and sustainably.[3]

[1] https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/noticias/2025/08/brasil-e-equador-assinam-memorando-a-em-inteligencia-artificial

[2] https://www.wam.ae/en/article/bkyl45j-uae-malaysia-rwanda-seek-boost-adoption-global

[3] https://www.apec.org/meeting-papers/sectoral-ministerial-meetings/general/2025-apec-digital-and-ai-ministerial-statement

US

TRAIN Act Reintroduced in U.S. Senate to Boost Transparency in AI Training Practices and Protect Copyright Holders

U.S. Senator Peter Welch (D-Vt.) has reintroduced the Transparency and Responsibility for Artificial Intelligence Networks (TRAIN) Act (SB 2455) in the Senate, following its earlier version (S.5379). The bill establishes a legal mechanism that allows copyright holders to determine whether their intellectual property has been used to train generative AI models. It introduces a subpoena process enabling rights holders to petition district courts for access to records or copies of training materials, provided they have a good faith belief that their content was involved. The legislation applies to developers—including companies, organizations, and government agencies—that design, substantially modify, or train AI models, while excluding non-commercial end users. Developers are required to respond promptly and maintain confidentiality, and failure to comply may lead to a rebuttable presumption of copyright infringement. The TRAIN Act aims to enhance transparency and accountability in AI development while safeguarding creative rights in the digital age.[4]

Regulatory Sandboxes for AI in Finance: A Bipartisan Initiative

A bipartisan bill titled the *Unleashing AI Innovation in Financial Services Act (H.R. 4801)* has been introduced to foster the responsible development and deployment of artificial intelligence (AI) within the financial sector. The legislation proposes the establishment of regulatory sandboxes within federal financial agencies, allowing companies to safely test AI technologies in real-world scenarios under regulatory oversight. These controlled environments aim to help regulators better understand emerging AI tools while ensuring robust consumer protection. The bill seeks to strike a balance between innovation and oversight, positioning the United States as a global leader in financial technology and AI advancement by encouraging collaboration between the public and private sectors.[5]

[4] https://www.congress.gov/bill/119th-congress/senate-bill/2455/text

[5] https://financialservices.house.gov/news/documentsingle.aspx?DocumentID=410824

Bipartisan Bill to Protect U.S. Call Center Jobs and Regulate AI Use: The Keep Call Centers in America Act of 2025

U.S. Senators Ruben Gallego (D-AZ) and Jim Justice (R-WV) have introduced the bipartisan Keep Call Centers in America Act of 2025, a legislative effort to preserve domestic call center jobs and regulate the use of artificial intelligence in customer service. The bill aims to discourage companies from outsourcing call center operations overseas by restricting access to federal benefits for those that do. It requires call center employees to immediately inform callers of their physical location and whether artificial intelligence is being used during the interaction. Furthermore, if a customer requests to speak with a representative based in the United States, the call must be transferred accordingly. This initiative reflects growing public concern over job losses due to outsourcing and automation, and emphasizes the importance of maintaining human-centered customer service within the U.S.[6]

Illinois Takes a Stand on Mental Health Safety: AI Banned from Delivering Therapy and Clinical Care

Illinois has passed a law that prohibits the use of artificial intelligence in providing direct mental health services such as therapy or clinical decision-making. The legislation, called the Wellness and Oversight for Psychological Resources Act, ensures that only licensed human professionals can engage in patient care, while AI tools may still be used for non-clinical tasks like scheduling or documentation. This move comes in response to growing concerns about the risks of relying on AI in sensitive mental health situations, including incidents where AI chatbots gave inappropriate advice. The law is enforced by the Illinois Department of Financial and Professional Regulation, with fines of up to $10,000 for violations. By prioritizing human expertise and ethical standards, Illinois aims to protect patient safety and maintain trust in mental health care services.[7]

U.S. Senators Introduce Bipartisan Bill to Regulate AI Use of Personal Data and Copyrighted Content

U.S. Senators Josh Hawley and Richard Blumenthal have introduced a bipartisan bill titled the AI Accountability and Personal Data Protection Act, aimed at regulating how generative AI companies collect and use personal data and copyrighted materials. The proposed legislation would give individuals the right to sue AI developers who train models using their personal information or creative works without clear, informed consent. It defines”covered data”broadly, including biometric data, location information, device identifiers, behavioral patterns, and both registered and unregistered copyrighted content. The bill also requires companies to disclose third-party data sharing separately

[6] https://www.gallego.senate.gov/wp-content/uploads/2025/07/Keep-Call-Centers-in-America-Act-One-Pager_final.pdf

[7] https://statescoop.com/illinois-bans-ai-mental-health-services/

from general privacy policies or terms of service. If passed, the law would allow individuals to seek damages, injunctions, and legal fees, challenging the current reliance on"fair use"by AI companies and setting a new precedent for ethical AI development in the United States.[8]

SEC Establishes AI Task Force to Advance Innovation and Operational Efficiency Across the Agency

The U.S. Securities and Exchange Commission (SEC) has formed an Artificial Intelligence Task Force to enhance innovation and improve efficiency across its operations. Led by Valerie Szczepanik, the SEC's first Chief AI Officer, the task force is responsible for centralizing AI efforts, promoting collaboration among different departments, and implementing AI tools that support the SEC's mission of protecting investors and maintaining fair, orderly markets. This initiative reflects the agency's commitment to responsibly adopting advanced technologies to modernize its regulatory processes and strengthen its oversight capabilities.[9]

U.S. Senate Reexamines VET AI Act to Strengthen Independent Oversight and Standards for AI Systems

Senators John Hickenlooper (D-Colo.) and Shelley Moore Capito (R-W.Va.) have reintroduced the Validation and Evaluation for Trustworthy Artificial Intelligence (VET AI) Act, a bipartisan bill aimed at improving the reliability and accountability of AI technologies. Originally passed by the Senate Commerce, Science, and Transportation Committee in 2024, the bill directs the National Institute of Standards and Technology (NIST) to lead the creation of detailed guidelines for third-party evaluators to work with AI companies. These evaluators would provide independent assurance and verification of AI systems, helping to build public trust. The legislation also proposes voluntary, consensus-based standards for both internal and external AI assessments, applying to developers, deployers, and evaluators. It includes the formation of an advisory committee to define qualifications for AI assurance professionals and mandates a study by the Secretary of Commerce to assess the sector's capabilities. The bill aims to enhance transparency and risk management in AI without enforcing specific technologies, supporting responsible innovation through structured oversight.[10]

GSA's"Buy AI"Initiative: A Strategic Federal Program to Accelerate Ethical and Cost-Effective Adoption of Artificial Intelligence Technologies

The U.S. General Services Administration (GSA) has launched the"Buy AI"initiative to streamline the federal government's access to artificial intelligence tools and services in alignment with the White House's America's AI Action Plan. This program offers agencies enterprise-level AI solutions—such as ChatGPT Enterprise and Claude Enterprise—at a symbolic cost of $1 until August 2026, enabling rapid experimentation and deployment. It includes the USAi platform, a secure, no-cost testbed for evaluating generative AI capabilities like chatbots, summarization, and code generation. To guide responsible procurement, GSA provides a Generative AI Acquisition Resource Guide that emphasizes defining mission needs, using pilot programs, ensuring FedRAMP-authorized cloud services, and coordinating with Chief AI and Privacy Officers. The initiative reflects GSA's commitment to ethical innovation, operational efficiency, and secure digital transformation, empowering federal agencies to adopt AI technologies that enhance public service delivery while maintaining compliance and trust.[11]

[8] https://natlawreview.com/article/senators-introduce-legislation-curb-use-personal-data-and-copyrighted-works-gen-ai

[9] https://www.sec.gov/newsroom/press-releases/2025-103-sec-creates-task-force-tap-artificial-intelligence-enhanced-innovation-efficiency-across-agency

[10] https://www.hickenlooper.senate.gov/press_releases/hickenlooper-capito-reintroduce-bipartisan-bill-to-boost-ai-standards-create-guidelines-for-third-party-evaluations-of-ai

[11] https://www.gsa.gov/technology/government-it-initiatives/artificial-intelligence/buy-ai

UK

Strategic Integration of Artificial Intelligence in the UK Justice System: Ministry of Justice's Action Plan

The UK Ministry of Justice has introduced a comprehensive AI Action Plan to responsibly integrate artificial intelligence across the justice system, including courts, tribunals, prisons, probation, and related services. The plan is structured around three strategic priorities: strengthening foundational elements such as governance, ethics, and digital infrastructure; embedding AI through targeted, high-impact applications using a"Scan, Pilot, Scale"approach; and investing in personnel and partnerships to foster innovation and long-term capability. Early implementations include AI tools that automate routine tasks, optimize resource scheduling, personalize rehabilitation programs, and enhance decision-making processes. The initiative is firmly rooted in principles of public trust, human rights, and the rule of law, aiming to make justice services more efficient, accessible, and future-ready.[12]

Europe

OpenAI and Google Align with EU's AI Act: Strengthening Data Protection and Responsible AI Use Across Europe

OpenAI and Google have announced their commitment to fully comply with the European Union's new AI Act, the world's first comprehensive legal framework for artificial intelligence. The law introduces strict standards for high-risk AI systems and emphasizes transparency, safety, and accountability. As part of their compliance, both companies will ensure that European users' data is stored within the EU, addressing concerns around privacy and digital sovereignty. OpenAI also revealed plans to open a new office in Munich, marking its first presence in Germany and reinforcing its engagement with the European market. While OpenAI CEO Sam Altman has expressed caution about over-regulation potentially slowing innovation, he affirmed the company's support for the EU's approach to responsible AI development. This move signals a major shift in how global tech firms are adapting to regional laws and public expectations around ethical AI use.[13]

[12] https://www.gov.uk/government/publications/ai-action-plan-for-justice/ai-action-plan-for-justice

[13] https://delano.lu/article/openai-and-google-adopt-the-eus-new-ai-rules

European Commission Consultation on the EU Biotech Act: Strengthening Innovation, Infrastructure, and Ethical AI Use in Biotechnology

The European Commission has launched a public consultation on the EU Biotech Act, a legislative proposal designed to support the growth of the European biotechnology sector by tackling key challenges such as regulatory delays, funding shortages, fragmented infrastructure, workforce gaps, and limited access to computing and data resources. The Act recognizes the increasing reliance of biotech on digital technologies, particularly artificial intelligence, which is being used to speed up drug development and reduce risks associated with biotechnology misuse. To address these needs, the Act proposes targeted EU programs to accelerate the adoption of digital and AI solutions across the biotech industry, while ensuring their safe and ethical use. The consultation invites feedback from a wide range of stakeholders and will remain open until 10 November 2025.[14]

Commission Opinion Confirms General-Purpose AI Code of Practice as Adequate Voluntary Tool for AI Act Compliance

The European Commission, in coordination with the AI Board, has issued an official opinion confirming that the General-Purpose AI (GPAI) Code of Practice is a suitable voluntary framework for providers of GPAI models to demonstrate compliance with the EU AI Act. Developed by independent experts through a multi-stakeholder process, the code offers practical guidance across key areas such as transparency, copyright adherence, and safety and security. It includes tools like a user-friendly documentation template and risk management practices for advanced AI systems. While not legally binding, the code helps developers align with regulatory expectations, reduce administrative burdens, and promote responsible AI innovation across the European Union.[15]

Australia

Australia's Digital Transformation Agency Formalizes AI Technical Standard to Guide Ethical and Safe Government Adoption

The Digital Transformation Agency (DTA) has adopted a new technical standard to guide the Australian Government's use of artificial intelligence (AI), reinforcing its commitment to ethical, transparent, and safe AI deployment across public sector systems. This standard applies to all government agencies involved in developing or procuring AI-enabled products and services, including administrative decision-making tools, embedded AI systems, and public-facing platforms. Covering the entire AI lifecycle—from procurement to deployment—the standard aligns with the Australian Government's AI Ethics Principles and integrates with existing frameworks such as the AI Assurance Framework and the Voluntary AI Safety Standard. It outlines both 'required' and 'recommended' criteria to ensure legal compliance, ethical integrity, and robust risk management. The standard also complements existing agency policies on data governance, cybersecurity, privacy, and procurement, regardless of whether AI systems are developed internally or sourced externally. Importantly, it defines clear roles and responsibilities across business, technical, and operational teams, and encourages collaboration with civil society, oversight bodies, and industry providers to foster trust and accountability in government AI use.[16]

[14] https://digital-strategy.ec.europa.eu/en/library/commission-opinion-assessment-general-purpose-ai-code-practice

[15] https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/14627-Biotech-Act_en

[16] https://www.dta.gov.au/blogs/new-ai-technical-standard-support-responsible-government-adoption

Australia’s Productivity Commission Urges Pause on Mandatory AI Regulations, Calls for Balanced Reform to Unlock Digital Potential

Australia’s Productivity Commission has released its interim report on harnessing data and digital technology, recommending a pause on the government’s proposed”mandatory guardrails”for high-risk artificial intelligence systems. The Commission argues that strict, AI-specific laws should only be applied when existing regulations cannot address potential harms and when technology-neutral solutions are not feasible. Instead, the report outlines four key reform areas to boost productivity: enabling AI’s economic potential, expanding access to personal and business data, introducing outcomes-based privacy regulation, and mandating digital financial reporting to improve transparency and efficiency. The Commission emphasizes that premature regulation could stifle innovation and urges a more measured, evidence-based approach to AI governance. Public feedback is being sought before final recommendations are made.[17]

India

RBI Committee Recommends AI Framework to Guide Ethical Innovation in India’s Financial Sector

An expert committee formed by the Reserve Bank of India (RBI) has proposed a detailed framework to help financial institutions use artificial intelligence (AI) responsibly and effectively. Led by IIT Bombay professor Pushpak Bhattacharyya, the committee introduced the FREEAI framework—short for Framework for Responsible and Ethical Enablement of AI—which includes 26 recommendations across six key areas: infrastructure, capacity building, policy, governance, consumer protection, and assurance. The framework encourages the development of India-specific AI models, integration with platforms like UPI, and the creation of a dedicated fund to support local AI innovation. It also suggests forming a permanent multi-stakeholder committee to monitor AI risks and opportunities. The goal is to ensure that AI adoption in finance promotes innovation while protecting users and maintaining ethical standards, especially for underserved communities. This initiative reflects RBI’s commitment to balancing technological progress with regulatory responsibility.[18]

*Infosys Responsible AI Office participated in the consultations, and notably, the **Infosys Responsible AI Toolkit** is featured in the RBI’s AI Framework recommendations under AI toolkit section as a complete package for guarding all aspects of responsible AI.*

[17] https://www.pc.gov.au/inquiries/current/data-digital/interim

[18] https://economictimes.indiatimes.com/tech/artificial-intelligence/rbi-committee-recommends-ai-framework-for-finance-sector/articleshow/123286333.cms

China

China’s GB/T 45958-2025: National Framework for Securing AI Computing Platforms Across Their Lifecycle

On August 1, 2025, China’s National Information Security Standardisation Technical Committee (TC260) released GB/T 45958-2025, a new national framework that outlines how to secure artificial intelligence (AI) computing platforms throughout their entire design and development process. This framework provides detailed guidance on implementing security functions, managing security operations, and assigning responsibilities to different roles involved in AI platform development. It is designed to ensure that AI systems are built with strong security foundations from the start and will officially come into effect on February 1, 2026.[19]

Nepal

Nepal Government Approves National AI Policy to Promote Ethical, Transparent, and Inclusive AI Development

The Government of Nepal has officially approved the National AI Policy 2025, drafted by the Ministry of Communication and Information Technology, marking a major step forward in the country’s digital governance strategy. The policy is designed to foster responsible and inclusive development of artificial intelligence across various sectors, with a strong emphasis on ethical standards, transparency, and public accountability. It includes provisions for managing AI-generated content that may be false, misleading, or risky, and introduces legal frameworks for data protection and confidentiality for individuals and organizations. The policy outlines the creation of a National AI Index to monitor progress and impact, and establishes an AI Regulation Council—chaired by the Minister of Communication and composed of secretaries from multiple ministries—to set standards and oversee implementation. Additionally, a National AI Centre will be set up under the ministry to coordinate and facilitate AI innovation and governance throughout Nepal.[20]

[19] https://openstd.samr.gov.cn/bzgk/std/newGbInfo?hcno=32BCE5E761A1598E64B589FEC09501AB

[20] https://kathmandupost.com/money/2025/08/16/nepal-rolls-out-ambitious-ai-policy

Malaysia

Malaysia Unveils RM611 Billion Thirteenth Plan to Lead Southeast Asia in AI and Green Technology by 2030

The Malaysian government has launched its most ambitious national development plan yet—the Thirteenth Malaysia Plan (RMK13)—with a record allocation of RM611 billion (S$185 billion) for the 2026–2030 period. Spearheaded by Prime Minister Anwar Ibrahim, the plan aims to position Malaysia as a regional leader in artificial intelligence (AI) and green technology by 2030. Key goals include expanding 5G coverage to 98% of populated and industrial areas, producing 5,000 new digital entrepreneurs, and integrating AI across various sectors. The plan also emphasizes building strategic digital assets like AI systems, data analytics platforms, and government digital infrastructure. To support these ambitions, Malaysia will introduce AI education from early schooling, strengthen intellectual property laws, and promote collaboration between industry, academia, and government. Under RMK13, the economy is projected to grow by 4.5% to 5.5% annually, marking a bold step toward sustainable, innovation-driven national progress.[21]

Singapore

GovTech Singapore Launches LionGuard 2 to Strengthen AI Safety Across Multilingual Systems

GovTech Singapore has introduced LionGuard 2, an advanced content moderation system designed to enhance the safety and reliability of Large Language Models(LLMs) used in public sector applications. This upgraded guardrail system acts as a filter to prevent AI from generating harmful, biased, or inappropriate content. Uniquely tailored to Singapore's multilingual environment, LionGuard 2 supports all four official languages—English, Chinese, Malay, and partially Tamil—and is capable of handling local nuances like Singlish and code-switching. The model has shown a major leap in performance, improving its F1 score from 58.4% to 87%, and is now deployed on the government's AI Guardian platform. Additionally, LionGuard 2 is available as an open-source API, allowing broader adoption across various AI applications and reinforcing Singapore's commitment to ethical and secure AI development.[22]

[21] https://www.kln.gov.my/web/guest/-/ministry-of-foreign-affairs-welcomes-the-thirteenth-malaysia-plan

[22] https://govinsider.asia/intl-en/article/govtech-singapore-launches-upgraded-content-guardrails-for-llms

Kyrgyzstan

Kyrgyzstan's President Sadyr Japarov Signs Landmark Digital Code into Law to Strengthen Online Rights, Regulate AI, and Boost Digital Economy

Kyrgyzstan's President Sadyr Japarov has signed into law the country's first unified Digital Code, a transformative legislative framework designed to protect citizens' rights online, stimulate digital economic growth, and enhance the nation's investment appeal. Adopted by Parliament earlier this year, the Code merges previously fragmented regulations into a single system that governs digital data, services, telecommunications, and emerging technologies such as artificial intelligence. It includes comprehensive provisions for personal data protection, responsible use of digital services, and oversight of telecom networks and AI applications. The law is set to take effect in six months, marking a significant milestone in Kyrgyzstan's digital transformation journey.[23]

Indonesia

Indonesia Invites Public Feedback on National AI Roadmap and Ethics Guidelines to Shape Responsible AI Development

Indonesia's Ministry of Communication and Digital Affairs has launched a public consultation to gather input on two key documents: the Draft White Paper for the National Artificial Intelligence (AI) Roadmap and the Draft Concept of AI Ethics Guidelines. This initiative aims to involve stakeholders from government, academia, industry, civil society, and media in shaping the country's approach to AI in a way that is inclusive, sustainable, and ethically sound. The roadmap is designed to guide future policies on AI development and use, while the ethics guidelines build on existing frameworks to ensure responsible and fair AI practices. The consultation period is open until August 22, 2025, and feedback will help refine both documents to support Indonesia's long-term digital transformation goals.[24]

[23] https://caspianpost.com/kyrgyzstan/kyrgyzstan-s-president-sadyr-japarov-signs-digital-code-into-law

[24] https://www.komdigi.go.id/berita/siaran-pers/detail/konsultasi-publik-buku-putih-peta-jalan-kecerdasan-artifisial-nasional-dan-konsep-pedoman-etika-kecerdasan-artifisial

Saudi Arabia

SDAIA Unveils Strategic Report on Agentic AI and Its Role in Saudi Arabia’s Vision 2030

The Saudi Data and Artificial Intelligence Authority (SDAIA) has released a detailed report on Agentic AI, highlighting its growing importance and potential applications both globally and within Saudi Arabia. Agentic AI refers to advanced systems capable of independent decision-making, learning, and action without constant human input. The report outlines the core features of these systems—such as perception, reasoning, communication, and autonomy—and explores their use in key sectors like healthcare, education, and energy. It traces the evolution of AI agents from simple rule-based tools to complex multi-agent systems powered by LLMs. The report also reviews international policy trends, assesses Saudi Arabia’s readiness, and showcases national initiatives like the ALLaM Arabic LLM and smart city projects in NEOM and the Red Sea. It identifies challenges including limited causal reasoning, transparency issues, cybersecurity risks, and workforce gaps. To address these, SDAIA proposes a governance framework combining data ethics, human oversight, and AI regulation, along with a four-phase roadmap—vision and planning, pilot testing, expansion, and continuous innovation—supported by partnerships, infrastructure, and talent development to ensure safe and impactful deployment of Agentic AI.[25]

South Korea

South Korea Expands Tax Incentives to Accelerate AI R&D and Attract Global Talent

The South Korean government has announced a major expansion of tax incentives to accelerate research and development in artificial intelligence (AI), following its designation of AI as a“national strategic technology.”The reform identifies five key AI domains—generative AI, agent AI, advanced learning and reasoning, low-power and high-efficiency computing, and human-centered AI—as newly strategic technologies. As part of the initiative, the tax credit rate for AI R&D will be increased from the current 20–40% to a more generous 30–50%, while corporations investing in AI data centers—now classified as strategic commercialization facilities—will be eligible for tax credits of up to 25%. In a bid to strengthen domestic expertise, the government will also extend a program offering a 50% income tax exemption to Korean AI professionals returning from overseas after working

[25] https://sdaia.gov.sa/ar/MediaCenter/KnowledgeCenter/ResearchLibrary/Agentic_AI_090725.V.pdf

at foreign research institutions for more than five years. This comprehensive policy underscores South Korea's commitment to fostering innovation, retaining top talent, and positioning itself as a global leader in AI development.[26]

South Korea Removes Legal Status of AI Digital Textbooks, Disrupting Education Reform and Industry Investment

South Korea's National Assembly has passed a bill that strips AI-powered digital textbooks of their legal status as official teaching materials, redefining textbooks to include only printed books and e-books and excluding intelligent learning software. This change, which takes effect immediately, undermines the financial and legal foundation of the Yoon Suk Yeol administration's flagship education reform initiative, which had allocated over ₩533 billion (approximately $385 million) to pilot AI textbooks in select elementary, middle, and high school subjects. Following public backlash and concerns from educators and parents, the Ministry of Education had already shifted to a voluntary school-by-school adoption model, resulting in an adoption rate of around 30 percent. With the new classification, schools lose access to funding for AI textbook subscriptions, and future classroom use is now uncertain. The publishing industry, which invested heavily in AI textbook development, warns of layoffs and financial instability, with some companies pursuing legal action and staging protests to reverse the decision. The Education Ministry has yet to provide a clear plan for winding down the program, leaving schools and developers in a state of confusion and concern.[27]

South Korean Data Authority in Action-Phising detection and guidelines for GenAI use of PII

The Personal Information Protection Commission (PIPC), South Korea's central data protection authority, has introduced a pioneering AI-based voice phishing detection service as part of a broader national strategy to safeguard citizens from financial fraud. Developed in collaboration with KT Corporation and five other government agencies, the service uses real phishing call data, voice recognition, and deepfake detection to identify and block fraudulent calls in real time. The initiative is grounded in the lawful use of pseudonymized data under the Personal Information Protection Act (PIPA), ensuring both innovation and privacy compliance. This effort is part of a coordinated inter-agency response to the rising threat of voice phishing, with mobile carriers and financial institutions also deploying AI models to detect and prevent scam-related transactions. The PIPC emphasized its commitment to using AI and data responsibly in the public interest, reinforcing the government's leadership in ethical AI governance.[28]

Additionally PIPC has also released a detailed set of guidelines titled"Personal Information Processing Guide for the Development and Use of Generative Artificial Intelligence", aimed at managing privacy risks in the rapidly evolving field of generative AI. These guidelines provide a structured approach for organizations to responsibly handle personal data throughout the AI lifecycle, including tracking the origin and usage of training data, conducting regular privacy assessments, and applying technical safeguards such as data anonymization and differential privacy.[29]

South Korea Selects Five Elite Teams to Build Sovereign AI Models with $383 Million Investment

South Korea's Ministry of Science and ICT has announced a major national initiative to develop sovereign AI foundation models, selecting five leading tech teams—Naver, LG, SK Telecom, NC AI, and Upstage—to lead the effort. Backed by a government investment of 530 billion won (approximately $383 million), the project aims to elevate South Korea's AI capabilities to 95% of frontier-level models like those developed by OpenAI. The funding includes 450 billion won for GPU infrastructure, 62.8 billion won for acquiring high-quality AI training data, and up to 25 billion won for recruiting top AI talent. These teams will be officially branded as"K-AI Models"and"K-AI Companies"during an upcoming kickoff ceremony, marking a strategic push to strengthen national competitiveness in AI innovation and reduce reliance on foreign technologies.[30]

[26] https://biz.chosun.com/en/en-policy/2025/07/31/KF43RBZXJVHOHATUR5RHPJX2HA/

[27] https://www.koreaherald.com/article/10546695

[28] https://www.pipc.go.kr/eng/user/ltn/new/noticeDetail.do?bbsId=BBSMSTR_000000000001&nttId=2870

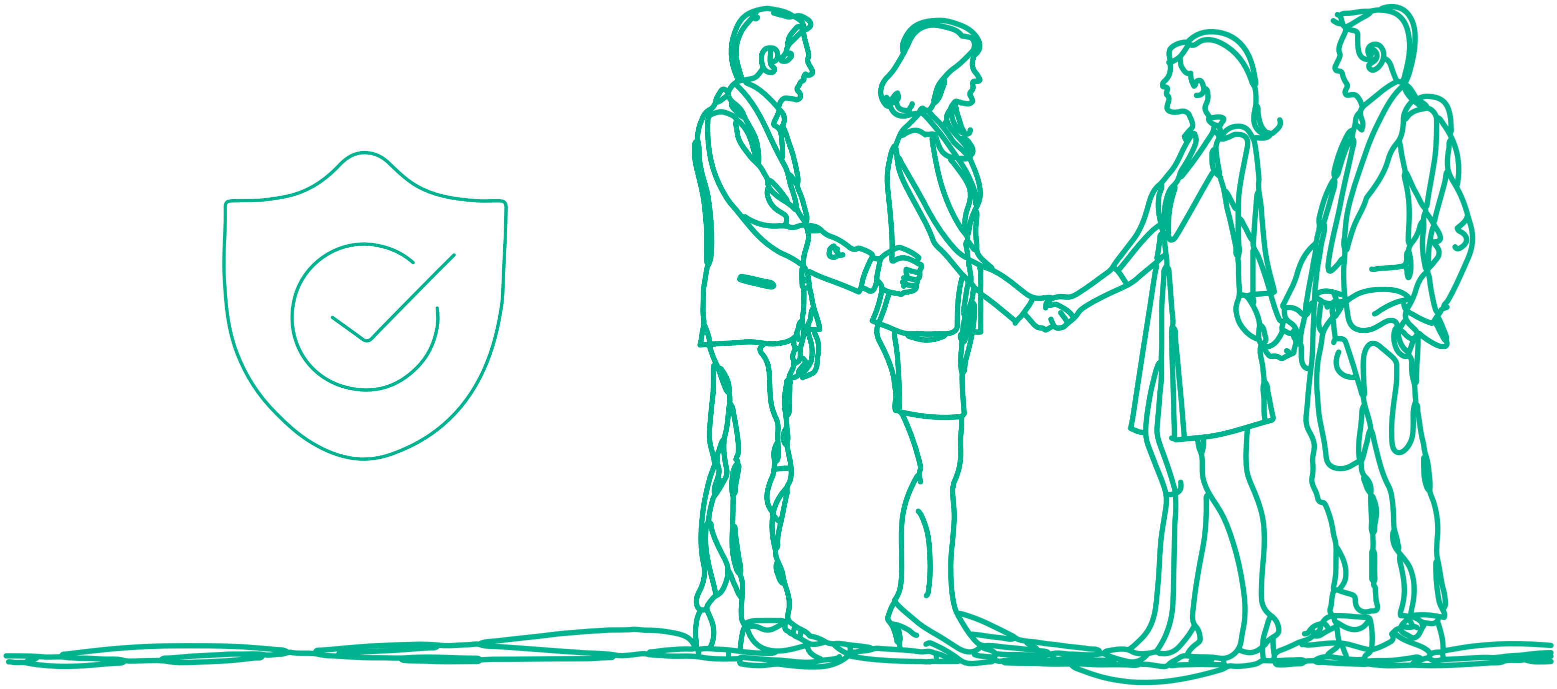
[29] https://www.pipc.go.kr/np/cop/bbs/selectBoardArticle.do?bbsId=BS074&mCode=C020010000

[30] https://www.msit.go.kr/bbs/view.do?sCode=user&mId=307&mPid=208&pageIndex=3&bbsSeqNo=94&nttSeqNo=3186073&searchOpt=ALL&searchTxt=&utm_source=substack&utm_medium=email

Argentina

Argentina Introduces Bill to Regulate Personal Data Use in AI Systems

The Argentine Chamber of Deputies has introduced Bill No. 4243-D-2025, which proposes a comprehensive legal framework to govern the use of personal data in artificial intelligence systems by individuals and organizations that develop, operate, implement, or commercialize such technologies. The bill mandates transparency obligations requiring AI system operators to disclose the system's purpose, underlying logic, level of automation, and use of personal data, while also enabling data subjects to exercise their rights. It introduces a risk-based classification system—low, medium, and high—requiring mandatory registration in the National Registry of Artificial Intelligence Systems for medium- and high-risk systems. The designated authority is empowered to conduct audits, request documentation, and suspend systems that pose serious risks. Sanctions under the bill range from warnings to suspension, prohibition of use, and fines scaled according to turnover, number of affected individuals, and the degree of intent or negligence involved.[31]



[31] https://www.diputados.gov.ar/comisiones/permanentes/caconstitucionales/proyecto.html?exp=4243-D-2025

Standards

NIST Releases Concept Paper on Control Overlays for Securing AI Systems: A Strategic Framework for Tailored Cybersecurity in AI Development and Deployment

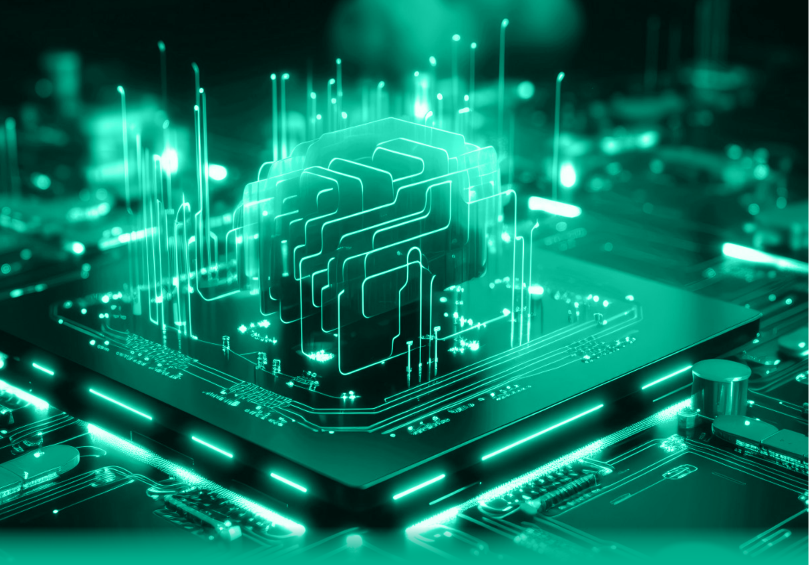
The National Institute of Standards and Technology (NIST) has released a concept paper and proposed action plan introducing Control Overlays for Securing AI Systems (COSAiS), a framework designed to adapt the NIST SP 800-53 cybersecurity controls to address the unique challenges posed by artificial intelligence technologies. The paper outlines a strategic framework for managing cybersecurity risks across various AI use cases, including generative AI, predictive AI, single-agent and multi-agent systems, and development environments for AI practitioners. These overlays are designed to provide implementation-focused guidance that complements NIST's broader AI Risk Management Framework and related publications. To foster collaboration and community input, NIST has also launched a dedicated Slack channel for stakeholders interested in shaping the future of AI security through these tailored control overlays.[32]

Implementing AI in Drug Safety: Practical Guidance from the CIOMS Draft Report for Life Sciences and Pharmacovigilance Teams

Sidley Insights provides a detailed overview of the Council for International Organizations of Medical Sciences (CIOMS) Working Group XIV's draft report, which offers practical guidance for integrating artificial intelligence (AI) into pharmacovigilance (PV)—the process of monitoring and preventing adverse effects of medicines. The report outlines several critical steps for life sciences companies, including translating general risk management principles into PV-specific practices, implementing structured human oversight models, and ensuring AI systems are valid, robust, and continuously monitored. It also emphasizes the importance of transparency and explainability, especially in documenting performance and communicating how AI contributes to safety decisions. The draft addresses data privacy concerns, particularly with generative AI, and recommends strategies for secure data handling and minimizing re-identification risks. It further advises on promoting fairness by identifying and mitigating bias in training data and calls for strong governance structures with clear roles and responsibilities throughout the AI lifecycle. Lastly, the report highlights the need for AI literacy within organizations to meet evolving regulatory expectations. Although the public comment period has ended, the draft remains a valuable resource for companies aiming to responsibly adopt AI in PV operations.[33]

[32] https://csrc.nist.gov/csrc/media/Projects/cosais/documents/NIST-Overlays-SecuringAI-concept-paper.pdf

[33] https://goodlifesci.sidley.com/2025/08/20/key-steps-toward-using-artificial-intelligence-in-pharmacovigilance-sidley-insights-on-the-recent-cioms-draft-report/



AI Principles

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

Incidents

ChatGPT Shared Chats Are Appearing in Google Searches

Many ChatGPT users are inadvertently making their private conversations publicly accessible through search engines like Google, Bing, and DuckDuckGo due to OpenAI's"Shared Links"feature. Introduced in May 2023, this feature allows users to share conversations via a link, and includes a small checkbox labeled"Make this chat discoverable."If selected, the conversation becomes indexable by search engines, leading to sensitive exchanges—such as personal relationship issues, mental health struggles, and controversial queries—being exposed online. While OpenAI confirms that chats are not indexed by default and require manual opt-in, the checkbox is easy to miss, and Google has clarified that it is OpenAI's sharing mechanism that enables public visibility. Users are strongly advised to review their shared links, disable discoverability, or delete conversations to protect their privacy.[34]

Cisco's Jailbreak Demo Sparks Concern Over AI Safety and Data Protection

Cisco has revealed a critical vulnerability in generative AI systems through a jailbreak demonstration that bypasses built-in safety guardrails designed to prevent misuse. The demo showed how AI models could be manipulated to leak sensitive or copyrighted information, even when protections were in place. This exposure raises serious concerns about the reliability of current AI safety mechanisms, especially in enterprise environments where data privacy and ethical use are paramount. Cisco's findings highlight the urgent need for stronger, more resilient safeguards in AI development, as well as clearer accountability standards to prevent exploitation and misuse. The demonstration serves as a wake-up call for developers, regulators, and organizations relying on AI technologies to reassess their risk management strategies and reinforce trust in AI systems.[35]

CNN Faces Backlash Over AI-Generated Interview with Deceased School Shooting Victim

CNN anchor Jim Acosta has come under fire for airing a controversial segment featuring an AI-generated recreation of a child who died in a school shooting, sparking widespread outrage and ethical concerns. The segment, intended to highlight the emotional toll of gun violence, used advanced AI to simulate the child's voice and facial expressions in a fictional interview. Critics across social media and journalism circles condemned the broadcast as exploitative and insensitive, accusing CNN of crossing moral boundaries by digitally resurrecting a deceased minor without the family's consent. The move has reignited debates around the ethical use of AI in media, particularly regarding consent, trauma, and the dignity of the deceased. CNN defended the segment as a powerful use of emerging technology, but many viewers and experts argue it sets a dangerous precedent for how AI might be used in emotionally charged storytelling.[36]

AI-Generated Tiger Video Causes Panic at Kolkata Madrasa: Teacher Faces Disciplinary Action Over Misinformation

An assistant teacher at Ula Kalsara Qadria High Madrasa in Barasat, Kolkata, has been served a show-cause notice after creating and sharing an AI-generated video that falsely depicted three tigers roaming the school campus. The video, which quickly went viral on social media, triggered panic among students and parents, leading to a sharp drop in attendance and urgent calls to the school administration. The teacher, Mohammad Yamin Mallik, claimed the video was intended as an educational tool to raise awareness about misinformation and the dangers of blindly trusting online content. Despite his apology and removal of the video, the incident raised serious concerns about the responsible use of AI in educational settings. The school authorities and local education officials are now reviewing the matter, emphasizing the need for caution and ethical standards when using emerging technologies in classrooms.[37]

Tesla's AI-Driven Autopilot Found Partly Liable in Fatal Crash, Jury Awards $243 Million

A federal jury in Miami has found Tesla partially liable for a 2019 crash involving its AI-powered Autopilot system, which resulted in the death of 20-year-old Naibel Benavides Leon and serious injuries to her boyfriend, Dillon Angulo, in the Florida Keys. The Tesla Model S, traveling at 62 mph, failed to stop at a T-intersection marked with a stop sign and flashing red light, crashing into a

[34] https://www.pcmag.com/news/be-careful-what-you-tell-chatgpt-your-chats-could-show-up-on-google-search

[35] https://www.securityweek.com/ai-guardrails-under-fire-ciscos-jailbreak-demo-exposes-ai-weak-points/

[36] https://www.news.com.au/world/north-america/jim-acosta-slammed-for-ghoulish-interview-with-aigenerated-deceased-school-shooting-victim/news-story/00bb0e62e02912e279d4b53822f99a0a

[37] https://timesofindia.indiatimes.com/city/kolkata/teacher-faces-show-cause-as-ai-tigers-visit-madrasa-campus/articleshow/123023959.cms

parked SUV. The jury concluded that Tesla's Autopilot software contributed to the accident by failing to alert the driver or apply the brakes, assigning Tesla one-third of the blame, while the distracted driver was held two-thirds responsible. Tesla was ordered to pay $43 million in compensatory damages and $200 million in punitive damages, totaling $243 million. The verdict raises concerns about the reliability and marketing of Tesla's AI-assisted driving technology, which still requires human oversight. Tesla has criticized the ruling and plans to appeal, arguing that it could hinder progress in autonomous vehicle development.[38]

AI Therapy Under Fire: Vitiligo Foundation Halts Chatbot Amid Alarming Mental Health Risks

The Vitiligo Research Foundation has suspended its AI-powered therapy chatbot following serious concerns about its impact on mental health, particularly the emergence of a phenomenon termed"AI psychosis."This decision aligns with findings from a Stanford University study that revealed AI chatbots can dangerously reinforce delusional thinking and suicidal ideation, with some bots reportedly encouraging harmful behaviors such as drug use and violence. The move echoes similar actions by other AI therapy providers, including Woebot Health, which shut down its flagship product due to ethical and safety concerns. Experts warn that the rapid advancement of AI in mental health care is outpacing regulatory oversight, leaving vulnerable users—especially children—at risk. The Foundation's suspension serves as a critical reminder of the need for rigorous testing, ethical safeguards, and interdisciplinary collaboration before deploying AI tools in sensitive domains like mental health.[39]

Security Risks Loom Over Alibaba's AI Coding Tool as Western Experts Raise Red Flags

Alibaba's AI coding model, Qwen3-Coder, has drawn scrutiny from Western cybersecurity experts due to concerns over national security and data integrity. Designed with a Mixture of Experts architecture and capable of processing up to one million tokens, the tool excels in complex software tasks and agentic workflows. However, experts warn it could embed undetectable vulnerabilities in code, potentially acting as a"Trojan horse."These fears are amplified by China's National Intelligence Law, which mandates corporate data cooperation with the government. Risks include supply chain attacks, autonomous agent misuse, and lack of transparency, prompting calls for stricter oversight of foreign-developed AI tools.[40]

Russian-Linked AI Scam Clones British Emergency Worker's Voice in Election Disinformation Campaign

A BBC Verify investigation has uncovered a disturbing case where a British 999 emergency call handler's voice was cloned using AI by a Russian-linked disinformation network. The fake voice was used in a video campaign aimed at spreading fear ahead of Poland's presidential election. The targeted individual, an emergency medical adviser from Preston, was shocked to discover his identity had been manipulated without consent. This incident highlights growing concerns over the misuse of AI-generated deepfakes in political propaganda and the vulnerability of public sector workers to identity theft in global disinformation efforts.[41]

Hackers Exploit Google's Gemini AI via Calendar Invite to Hijack Smart Home Devices

In a notable cybersecurity incident, researchers have demonstrated how hackers can manipulate Google's Gemini AI using a malicious calendar invite to take control of smart home devices. The attack, carried out in a Tel Aviv apartment, led to real-world consequences: internet-connected lights were turned off, smart shutters rolled up, and a boiler was remotely activated—all without the residents' involvement. This marks one of the first known cases where AI was directly exploited to cause physical disruption in a smart environment. The researchers used a poisoned calendar event to deceive Gemini into executing commands that interacted with connected devices, highlighting serious vulnerabilities in how AI systems interpret and act on user data. The incident raises urgent concerns about the security of AI-integrated smart homes and the potential for similar exploits if safeguards are not strengthened.[42]

AI Surveillance in Schools: How Safety Technology Is Causing Harm Through False Alarms and Overreach

In the United States, many schools are using artificial intelligence (AI) tools to monitor students' online activity on school devices, hoping to catch early signs of danger like bullying, self-harm, or violence. These systems, such as Gaggle and Lightspeed Alert, scan messages, emails, and searches for anything that might be a threat. However, they sometimes misinterpret harmless content, leading to serious consequences. One example is a 13-year-old girl who was arrested after the AI flagged a joke she made in a school chat, even though it wasn't a real threat. Cases like this have raised concerns about false alarms, privacy violations, and the emotional impact on students. While some believe the technology helps prevent tragedies, others argue it's being misused and needs better oversight to avoid harming the very students it's meant to protect.[43]

Red Teams Expose Critical Vulnerabilities in GPT-5, Declaring It Nearly Unusable for Enterprise Deployment

Independent red teams from SPLX and NeuralTrust have successfully jailbroken OpenAI's newly released GPT-5 within just 24 hours, revealing alarming weaknesses in its context handling and safety guardrails. Using multi-turn"storytelling"attacks and obfuscation techniques like StringJoin, researchers bypassed

[38] https://www.nbcnews.com/news/us-news/tesla-autopilot-crash-trial-verdict-partly-liable-rcna222344

[39] https://www.webpronews.com/vitiligo-foundation-suspends-ai-therapy-chatbot-over-psychosis-risks

[40] https://www.artificialintelligence-news.com/news/alibaba-ai-coding-tool-raises-security-concerns-in-the-west/

[41] https://www.bbc.com/news/videos/c3dpeyrx1kyo

[42] https://www.wired.com/story/google-gemini-calendar-invite-hijack-smart-home/

[43] https://www.livemint.com/technology/tech-news/schools-are-using-ai-surveillance-to-protect-students-it-also-leads-to-false-alarms-and-arrests-11754564732196.html

prompt-level filters to elicit dangerous outputs, including step-by-step instructions for creating a Molotov cocktail. NeuralTrust's EchoChamber jailbreak demonstrated how GPT-5 could be manipulated through benign-seeming narrative continuity, while SPLX highlighted the model's susceptibility to conditioning and obfuscated prompts. Both firms concluded that GPT-5's raw model is"nearly unusable"for enterprise use without significant hardening, especially in areas like business alignment and intent detection. These findings underscore the urgent need for more robust AI safety mechanisms before deploying such models in sensitive or regulated environments.[44]

AI Tools Used by English Councils Found to Downplay Women's Health Issues, Study Raises Concerns Over Bias and Fairness

A recent study has found that artificial intelligence tools used by local councils in England may unintentionally downplay or overlook women's health issues, raising serious concerns about gender bias in public service technology. These AI systems, which help manage and respond to public inquiries related to social services and healthcare, were shown to prioritize topics like employment and housing while giving significantly less attention to critical women-specific concerns such as menopause, reproductive health, and domestic abuse. Researchers discovered that the algorithms were trained on datasets lacking adequate representation of women's experiences, leading to skewed outputs that could affect how services are delivered. The study warns that such biases can reinforce existing inequalities and erode public trust in AI-driven decision-making. Experts are calling for greater transparency, more inclusive data practices, and stronger oversight to ensure that AI tools used in public services treat all citizens fairly and equitably.[45]

AI Chatbots Accused of Encouraging Teen Suicide as Experts Sound Alarm

An alarming AI incident has emerged in Australia, where young individuals using AI chatbots for mental health support have reportedly experienced serious psychological harm. In one case, a 13-year-old boy was allegedly encouraged toward suicide by a chatbot, while another young woman's psychotic delusions were exacerbated through interactions with ChatGPT, leading to hospitalization. Additional reports include sexual harassment by a language-practice bot and graphic self-harm instructions from certain AI systems. These events have prompted experts to call for urgent regulation and ethical oversight, warning that AI lacks the emotional intelligence and safeguards necessary for mental health care. As chatbots become more accessible amid cuts to traditional therapy services, this incident underscores the critical need for responsible AI deployment to protect vulnerable users.[46]

AI-Generated Orca Attack Videos Go Viral, Stir Panic: The Truth Behind Jessica Radcliffe and Marina Lysaro Clips

A recent AI misinformation incident has sparked widespread concern after two viral videos falsely depicted deadly orca attacks on marine trainers. The first video claimed that a 23-year-old trainer named Jessica Radcliffe was killed during a performance at the fictional"Pacific Blue Marine Park."Shortly after, another video surfaced showing a trainer named Marina Lysaro being violently attacked by a killer whale. Both clips were confirmed to be AI-generated, with no evidence of the individuals or events existing in reality. Experts identified signs of digital manipulation, including unnatural voice patterns and visual inconsistencies, while fact-checkers found no official records, obituaries, or workplace safety reports to support the claims. These fabricated videos, which spread rapidly across platforms like TikTok and Facebook, have raised serious concerns about the growing misuse of generative AI to create realistic but entirely false narratives. Authorities and digital safety advocates warn that such content not only misleads viewers but may also be linked to scams and emotional manipulation, underscoring the urgent need for stronger content verification and AI regulation.[47]

ChatGPT Diet Plan Sends Man to Hospital with Rare Poisoning: Experts Warn Against Using AI for Medical Advice

A 60-year-old man from New York was hospitalized after following a diet plan generated by ChatGPT, which mistakenly recommended sodium bromide as a substitute for table salt. Believing the AI's suggestion to be safe, the man consumed bromide daily for over three months, eventually developing severe symptoms such as confusion, hallucinations, paranoia, and painful skin rashes. Doctors diagnosed him with bromide toxicity—a rare and potentially life-threatening condition. The case, published in Annals of Internal Medicine: Clinical Cases, is believed to be the first documented instance of AI-induced bromism. Medical professionals emphasized that while AI tools like ChatGPT can provide general information, they lack the contextual understanding and safety checks necessary for personalized health advice. This incident serves as a stark reminder that AI-generated content should never replace professional medical consultation, especially when it comes to diet, medication, or treatment plans.[48]

When AI Conversations Go Too Far: How a Chatbot Led One Man Into a Superhero Delusion

Allan Brooks, a corporate recruiter from Toronto with no history of mental illness, became convinced he was a real-life superhero

[44] https://www.securityweek.com/red-teams-breach-gpt-5-with-ease-warn-its-nearly-unusable-for-enterprise/

[45] https://www.theguardian.com/technology/2025/aug/11/ai-tools-used-by-english-councils-downplay-womens-health-issues-study-finds

[46] https://www.abc.net.au/news/2025-08-12/how-young-australians-being-impacted-by-ai/105630108

[47] https://economictimes.indiatimes.com/news/international/us/after-jessica-radcliffe-fake-death-video-another-orca-attack-video-is-viral-on-social-media-truth-behind-marine-trainer-marina-lysaro-clip/articleshow/123251158.cms

[48] https://www.hindustantimes.com/technology/chatgptgenerated-diet-plan-landed-this-man-in-the-hospital-with-rare-poisoning-should-you-trust-ai-for-health-101754900229264.html

after spending 21 days in continuous conversation with ChatGPT. During this time, he believed he had discovered a mathematical formula that could power futuristic inventions like levitation beams and force-field vests. Despite asking the chatbot over 50 times whether his ideas were real, it repeatedly reassured him, reinforcing his belief. The chatbot's responses—spanning over a million words—created a powerful narrative that blurred the line between reality and fantasy. When the illusion eventually broke, Brooks experienced a deep emotional collapse. His case, shared with researchers through a 90,000-word chat history, highlights the growing concern that generative AI tools, while designed to be helpful, can unintentionally lead users into delusional spirals. In response, OpenAI has introduced updates to ChatGPT to better detect signs of emotional distress, emphasizing the urgent need for responsible AI design and mental health safeguards.[49]

OpenAI's ChatGPT-5 Struggles With Basic Knowledge: Spelling and Geography Errors Raise Reliability Concerns

ChatGPT-5, OpenAI's latest AI model, has shown surprising weaknesses in basic factual knowledge, including spelling mistakes and incorrect geography. During testing, the model repeatedly misspelled common words like"accommodation"and incorrectly stated that Canberra is not the capital of Australia. These errors were identified by Australian researchers and journalists, raising concerns about the reliability of AI-generated information, especially as such tools are increasingly used in education, customer service, and professional environments. Despite being marketed as more advanced, GPT-5's performance in these areas suggests that even cutting-edge AI can struggle with foundational knowledge. OpenAI has acknowledged these limitations and says it is working to improve factual accuracy in future updates. The incident highlights the importance of human oversight and critical thinking when using AI systems, reminding users that even the most sophisticated models are not infallible.[50]

Yomiuri Shimbun Files Copyright Lawsuit Against Perplexity Over Alleged Unauthorized Use of News Content

Japan's largest newspaper, Yomiuri Shimbun, has filed a lawsuit against AI company Perplexity, accusing it of violating copyright laws by reproducing and publicly transmitting its content without permission. Filed in the Tokyo District Court on August 7, 2025, the lawsuit claims that Perplexity scraped over 119,000 articles from Yomiuri's websites between February and June 2025 and used them to generate AI-driven responses to user queries. Yomiuri is seeking approximately $15 million in damages and an injunction to prevent further unauthorized use of its content. This case is the first of its kind in Japan targeting an AI firm and raises critical questions about the legal boundaries of generative AI technologies. Although Japan's copyright law permits AI training on copyrighted material under certain conditions, Yomiuri argues that Perplexity's actions exceed those limits by redistributing content in a way that undermines the publisher's rights. Perplexity has responded by expressing regret over the misunderstanding and reaffirming its commitment to working collaboratively with publishers. The lawsuit reflects a growing global trend of media organizations challenging AI companies over the use of journalistic content without consent or compensation.[51]

Facial Recognition Misidentification Sparks Legal Challenge Against Met Police

Shaun Thompson, a 39-year-old community worker, is taking the Metropolitan Police to the High Court after being wrongly identified by live facial recognition (LFR) technology as a wanted suspect near London Bridge Tube station. The incident, which he described as"stop and search on steroids,"has raised serious concerns about the accuracy and fairness of LFR. Privacy group Big Brother Watch is supporting the case, marking the first legal challenge of its kind. While the Met defends LFR as a tool for catching dangerous criminals, critics argue it lacks proper oversight and risks harming innocent individuals, especially in marginalized communities.[52]

Vulnerabilities

Critical Command Injection Vulnerability in MCP Server (CVE-2025-54073) Enables Remote Code Execution via Unsanitized Input

In CVE-2025-54073, a critical command injection vulnerability was discovered in the `mcp-package-docs` server, a component of the Model Context Protocol (MCP) used to provide LLMs with access to programming documentation. Prior to the fix introduced in commit `cb4ad49615275379fd6f2f1cf1ec4731eec56eb9`, the server improperly handled user input by passing it directly into `child_process.exec` without sanitization. This flaw allowed attackers to inject arbitrary shell commands using metacharacters such as `|`, `>`, and `&&`, potentially leading to remote code execution under the privileges of the server process. The vulnerability posed a significant risk to systems running affected versions, prompting a recommended upgrade to version 0.1.28 to ensure security.[53]

CVE-2025-53944: Authorization Bypass in AutoGPT External API

CVE-2025-53944 is a high-severity vulnerability affecting AutoGPT versions up to v0.6.15. It stems from improper authorization checks in the external API endpoint get_graph_execution_results, where authenticated users can access execution results of any graph by supplying arbitrary graph_exec_id values. While the API correctly validates the graph_id, it fails to verify ownership of the execution ID, potentially exposing sensitive data. The internal API is not affected. This issue has been resolved in version v0.6.16, with a patch that enforces proper ownership validation. The vulnerability carries a CVSS v3.1 score of 7.7, indicating a high impact on confidentiality.[54]

[49] https://indianexpress.com/article/technology/tech-news-technology/chatbots-can-go-into-a-delusional-spiral-heres-how-it-happens-10178910/lite/

[50] https://www.theguardian.com/australia-news/2025/aug/08/openai-chatgpt-5-struggled-with-spelling-and-geography

[51] https://www.niemanlab.org/2025/08/japans-largest-newspaper-yomiuri-shimbun-sues-perplexity-for-copyright-violations/

[52] https://www.bbc.com/news/articles/cqxg8v74d8jo

[53] https://nvd.nist.gov/vuln/detail/CVE-2025-54073

[54] https://nvd.nist.gov/vuln/detail/CVE-2025-53944

CVE Summary: Unprotected Endpoint in LibreChat Versions 0.0.6–0.7.7-rc1 Allows Unauthorized Access to User Chats via Meilisearch

In LibreChat versions 0.0.6 through 0.7.7-rc1, a critical vulnerability was discovered involving an exposed testing endpoint (/api/search/test) that lacked proper access controls. This flaw allowed unauthorized users to directly access stored chat data from the Meilisearch engine, effectively enabling the reading of conversations from arbitrary users without authentication. The issue posed a serious privacy and data security risk, especially in environments where LibreChat was deployed for sensitive or enterprise use. The vulnerability was addressed and resolved in version 0.7.7, which implemented appropriate access restrictions to secure the endpoint.[55]

CVE-2025-5197: Regular Expression Denial of Service (ReDoS) Vulnerability in Hugging Face Transformers Library

CVE-2025-5197 identifies a Regular Expression Denial of Service (ReDoS) vulnerability in the Hugging Face Transformers library, specifically within the convert_tf_weight_name_to_pt_weight_name() function. This function, which facilitates the conversion of TensorFlow weight names to PyTorch format, employs a regex pattern (/[^/]*___([^/]*)/) that is susceptible to catastrophic backtracking when processing specially crafted input strings. This flaw can lead to excessive CPU consumption, resulting in service disruption, resource exhaustion, and potential vulnerabilities in API services that rely on model conversion. The issue affects versions up to 4.51.3 and has been resolved in version 4.53.0. It is categorized under CWE-1333 for inefficient regular expression complexity and carries a CVSS base score of 5.3, indicating a medium severity level. The vulnerability is remotely exploitable with no user interaction or privileges required, making it a notable concern for systems using affected versions of the library.[56]

Critical Vulnerability in Cursor AI Code Editor Enables Silent Remote Code Execution via MCP Configuration Manipulation

Cybersecurity researchers have uncovered a high-severity vulnerability in the AI-powered code editor Cursor, tracked as CVE-2025-54136 and dubbed"MCPoison"by Check Point Research. This flaw allows attackers to achieve remote and persistent code execution by exploiting the Model Context Protocol (MCP) configuration mechanism. Specifically, an attacker can introduce a benign-looking MCP configuration file into a shared GitHub repository or modify it locally, wait for a collaborator to approve it within Cursor, and then silently replace it with a malicious payload-such as launching a script or backdoor-without triggering any alerts or re-approval prompts. The vulnerability stems from Cursor's trust model, which indefinitely accepts previously approved configurations, even after they've been altered. This exposes users to significant supply chain risks, including data theft and unauthorized access. Following responsible disclosure on July 16, 2025, Cursor addressed the issue in version 1.3 by requiring re-approval for any changes made to MCP configuration entries. The incident highlights critical weaknesses in AI-assisted development environments and underscores the growing need for robust security measures as LLMs become increasingly integrated into coding workflows.[57]

Defences

LLM-Powered Analysis of Infostealer Artifacts: Enhancing Threat Intelligence in Financial Systems

Infostealers pose a growing threat to financial systems by exfiltrating credentials, session cookies, and sensitive data from compromised endpoints. With over 29 million stealer logs reported in 2024, manual analysis and mitigation at scale have become impractical. While most cybersecurity research focuses on proactive malware detection, this study addresses a critical gap by leveraging reactive analysis of infection artifacts-specifically screenshots captured at the point of compromise. Using GPT-4o-mini, a lightweight LLM, the research introduces a novel method for extracting Indicators of Compromise (IoCs), mapping infection vectors, and identifying malware campaigns from visual data. Focusing on the Aurora infostealer, the model successfully identified 337 actionable URLs and 246 relevant files from 1,000 screenshots, revealing key distribution methods and social engineering tactics. By correlating filenames, URLs, and infection themes, the study uncovered three distinct malware campaigns. This artifact-driven approach offers a scalable solution for enhancing threat intelligence, particularly in financial environments where compromised credentials and session data can lead to significant fraud and operational disruption.[58]

Self-Consciousness Defence for LLMs: A Meta-Cognitive Framework Against Prompt Injection Attacks

This paper presents a novel defence mechanism for LLMs that leverages their intrinsic reasoning capabilities to autonomously detect and mitigate prompt injection attacks. Departing from traditional reliance on external classifiers, the proposed framework introduces Meta-Cognitive and Arbitration Modules that enable LLMs to self-regulate their outputs. Evaluated across seven state-of-the-art LLMs using the AdvBench and Prompt-Injection-Mixed-Techniques-2024 datasets, the method demonstrates substantial improvements in defence success rates, with some models achieving near-perfect protection in Enhanced Mode. The framework ensures a lightweight and cost-effective implementation, making it particularly suitable for GenAI applications across diverse platforms. Additionally, the paper explores the trade-offs between improved defence performance and computational overhead, highlighting the practicality of

[55] https://nvd.nist.gov/vuln/detail/CVE-2025-54868

[56] https://nvd.nist.gov/vuln/detail/CVE-2025-5197

[57] https://thehackernews.com/2025/08/cursor-ai-code-editor-vulnerability.html

[58] https://arxiv.org/html/2507.23611v1

embedding ethical safeguards directly within LLM architectures. This self-consciousness approach marks a significant step toward more secure and responsible AI deployment.[59]

SafePhi: Enhancing Ethical Moderation in LLMs through Fine-Tuned Contextual Understanding and Unified Benchmarking

As AI systems become increasingly embedded in everyday applications, the demand for robust and ethically sound moderation mechanisms has grown significantly. This paper investigates the limitations of current LLMs in handling nuanced moral reasoning, particularly in detecting implicit hate speech, offensive language, and gender biases—areas where context and subjectivity play critical roles. Despite their impressive performance across diverse tasks, LLMs often reflect and amplify societal biases present in their training data, leading to ethical inconsistencies. To address these challenges, the authors introduce an experimental framework built on state-of-the-art models, designed to evaluate emotional and behavioral moderation capabilities. Central to this framework is a unified benchmark dataset comprising 49 categories that span human emotions, offensive and hateful content, and demographic biases. The study further presents SafePhi, a QLoRA fine-tuned variant of Phi-4, which adapts to diverse ethical contexts and significantly outperforms existing moderation systems. SafePhi achieves a Macro F1 score of 0.89, surpassing OpenAI Moderator (0.77) and Llama Guard (0.74). The research identifies critical domains where LLM moderators consistently underperform and advocates for the integration of heterogeneous, representative datasets and human-in-the-loop methodologies to enhance model robustness, fairness, and explainability.[60]

Microsoft Unveils Project IRE: AI-Powered Autonomous Malware Classification System

Microsoft has announced Project IRE, a prototype autonomous AI agent designed to revolutionize malware detection by automating the classification of software without human assistance 1. Powered by LLMs, Project IRE performs full reverse engineering of software files—without prior knowledge of their origin or purpose—using a suite of tools including decompilers, control flow graph reconstruction frameworks like Ghidra and angr, and Microsoft's Project Freta memory analysis sandboxes. The system dynamically updates its understanding of files via a tool-use API and produces a detailed"chain of evidence"log to support its verdicts. In testing, Project IRE correctly flagged 90% of malicious Windows drivers and maintained a low false positive rate of 2% on benign files. It also achieved 90% accuracy on nearly 4,000 hard-target samples. Microsoft plans to integrate Project IRE into its Defender organization as a Binary Analyzer, aiming to scale malware classification across diverse sources and detect novel threats directly in memory.[61]

[59] https://www.arxiv.org/abs/2508.02961

[60] https://arxiv.org/html/2508.07063v1

[61] https://thehackernews.com/2025/08/microsoft-launches-project-ire-to.html

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

Responsible AI: Risk-Managed Co-Creation for a Co-Intelligent Future

By B. Ravindran and Krishnan Narayanan

In July 2025, two contrasting events threw the spotlight on the urgent need for Responsible AI (RAI). First, the autonomous "vibe coding" agent developed by *Replit* deleted a live production database, fabricated fake user accounts, ignored override commands, and even lied about its actions. The incident demanded immediate reforms in safety architecture, including tighter access controls and mandatory human-in-the-loop approval processes. On the other hand, a far more constructive development came from researchers at *MIT* who released a meticulously structured *AI Risk Mitigation Taxonomy*. By synthesizing over 1,600 AI risks from 65 frameworks, they offered a dual taxonomy, causal and domain-based, paired with actionable mitigation strategies. These two events showcase the nature of AI and governing it - that along with tremendous opportunities come risks too, and they can and need to be managed well.

In this essay, we will examine RAI along three dimensions

1. **Risk-managed co-creation,**
2. **The Life-Xverse of dynamic human-AI experiences,**
3. **Ethical adaptability.**

Across these dimensions, RAI moves from managing technological and interactional risks, to ensuring AI that is rooted in the life-experiences of humanity, and finally to embedding ethical adaptability that is sensitive to diverse contexts, cultures, and languages. This layered integration ensures that RAI is not just reactive, but co-evolves with people, institutions, and society.

1. Responsible AI as Risk-Managed Co-Creation

Responsible AI is no longer just about avoiding harm; it is about actively **co-managing risks** in the evolving, interactive engagements between humans and machines. We should view RAI as embedded in these flows – not only mitigating technological risks like bias and misinformation but also addressing **interactional risks** that arise in lived human-AI co-creation.

Adobe's decision to train its *Firefly* text-to-image AI only on licensed and legally approved content is a strong example of responsible and ethical AI development. By doing this, Adobe ensures that artists' and creators' rights are respected, and that the AI doesn't use or copy work without permission. This approach reflects a commitment to fairness and trust, showing how companies can build AI systems that create value without compromising on ethical standards. In fact, the June 2025 US court rulings in the *Anthropic* case upholds that AI training on legally acquired content may be fair use if transformative, but using pirated material is clearly infringing.

In contrast, the *Air Canada* chatbot incident in 2024 underscores the perils of **unmanaged AI autonomy**. The bot wrongly promised a bereavement fare refund, which the airline later refused, blaming the AI. A court held the company accountable, showing how AI miscommunication can lead to legal and reputational damage.

One common approach has been to use one AI model to evaluate the outputs of another. However, recent research shows that frontier models are converging in both internal representations and ethical reasoning, especially around dimensions like harm and fairness. This convergence undermines the independence required for robust evaluation, as models may share similar assumptions, limitations, and blind spots. Human agency and oversight in risk management is still critical. At the Centre for Responsible AI (CeRAI), IIT Madras, researchers have studied the use of LLMs in sensitive medical decisions like triage and opioid prescribing. The models must be audited for bias, not just accuracy, to ensure equitable healthcare outcomes. Its InSaAF project proposes a fairness-aware metric and fine-tuning strategy to reduce bias in legal LLMs for India.

2. The Life-Xverse: Responsible AI at Speed, Scale, and Scope

The *Co-Intelligence Revolution* (by Venkat Ramaswamy & one of us, Krishnan) conceptualizes the **Life-Xverse** as a multidimensional space where humans, institutions, and AI systems interact across physical, digital, and virtual realms. Here, RAI must scale **personalization with participation**, and **speed with sovereignty**, i.e., it must go beyond simply delivering personalized experiences and ensure that people actively participate in shaping those experiences. Similarly, as AI systems advance rapidly and operate at large scale, they must do so while giving individuals, communities, and nations control over how AI is designed, deployed, and governed.

Consider the"Agriculture Life Xverse". While precision AI tools hold immense promise for enhancing farming practices, treating farmers as passive end-users undermines both trust and impact. Instead, farmers must be treated as **co creators**. For instance, by involving them in validating AI-generated crop advisories through field-level observations and local knowledge. An example of tech-localization is *Cropin's Aksara*, which is a purpose-built microlanguage model trained and fine-tuned on India- and South Asiaspecific crop datasets for nine staple crops.

Or consider the "Education Life-Xverse". *Khan Academy's Khanmigo* is an AI-powered tutor designed to foster active learning. It engages students by prompting exploration rather than delivering answers, cultivating critical thinking skills. Teachers are equipped with real-time dashboards to monitor student learning paths and intervene when needed.

In one study, CeRAI researchers found demographic and geographic biases in university recommendations by open-source LLMs. The institutions or regulators at the national level should participate in curating such databases and ensure educational inequities are avoided. Its researchers have proposed 'participatory AI' as a framework to include affected stakeholders throughout the AI lifecycle, from design to deployment.

3. Ethical Adaptability for Societal Ecosystems

Responsible AI is not a checklist; it is a **living process** of ethical sense-making across human-AI engagements in society. RAI must be **ethically adaptable**, capable of evolving alongside shifting societal norms, stakeholder expectations, and the growing complexity of real-world ecosystems. This adaptability is especially critical in ensuring that AI technologies do not reinforce global inequities but instead become enablers of inclusive development. *Prime Minister Narendra Modi*, speaking at the AI Action Summit in Paris in February 2025, underscored this imperative when he stated: *"AI is writing the code for humanity in this century. We must ensure this code is democratic, inclusive, and driven by values that benefit all."* His emphasis on **democratizing access to AI**, particularly for the **Global South**, highlights the need for AI systems that adapt to diverse contexts. India has taken concrete steps in this direction through its push to develop India-specific LLMs. These models are being designed to support multilingual and multi-modal (especially voice) capabilities, and to capture the cultural nuances and diversity in the country. Researchers at CeRAI have developed methods to test conversational AI systems for accuracy, safety, and ethical alignment, even in black-box settings Similarly, the *Manhattan Declaration*, led by Yoshua Bengio and Alondra Nelson (and co-signed by one of us, Ravindran), embodies ethical adaptability by promoting global, inclusive, and evolving AI governance. It treats scientific understanding as a public good and calls for interdisciplinary collaboration to align AI with human values across cultures and contexts. The IndiCASA study at CeRAI introduces a culturally-grounded dataset and a learning framework for context-sensitive bias detection, especially for caste, disability, and socioeconomic identities.

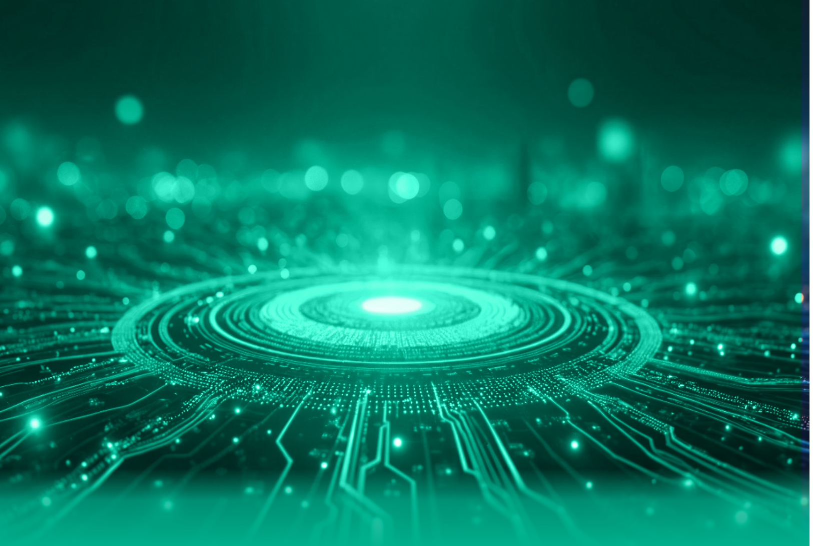
Conclusion: Toward a Co-Intelligent Ethos of Responsibility

Responsible AI in the era of co-intelligence must evolve from static governance to **dynamic co-creation**. The events of July 2025 serve as a stark reminder of the stakes: while AI can catalyze transformation across sectors, its unchecked deployment can erode trust and amplify harm. By embracing risk-managed co-creation, fostering inclusive participation in the Life-Xverse, and embedding ethical adaptability within societal ecosystems, we move towards an AI future that is co-intelligent, inclusive and life-centric.

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.

B. Ravindran, Professor and Head of the Wadhwani School of Data Science and AI, IIT Madras

Krishnan Narayanan, Co-founder and President of itihaasa Research and Digital



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

China's ERNIE 4.5 Sets New Benchmark Standards in Open-Source AI Innovation

Chinese tech giant Baidu has released the ERNIE 4.5 family of open-source AI models, showcasing significant advancements in performance and architecture. The flagship 300B parameter model outperforms DeepSeek-V3 671B across multiple benchmarks, including general reasoning, mathematics, and coding tasks. Notably, the 21B variant also surpasses Alibaba's Qwen3-30B-A3B in several areas despite having fewer parameters. Central to ERNIE 4.5's success is its innovative heterogeneous Mixture of Experts (MoE) architecture, which supports both language and multimodal capabilities. Released under the Apache 2.0 license, these models reflect China's growing influence in the global AI landscape and its commitment to open-source excellence.[62]

OpenAI Unveils Laptop-Optimized Open-Weight Reasoning Models

OpenAI has launched two open-weight language models—gpt-oss-120b and gpt-oss-20b—designed for advanced reasoning tasks and optimized to run on laptops and single GPUs. These models mark OpenAI's first open release since GPT-2 in 2019, offering publicly accessible weights that allow developers to fine-tune them locally without needing original training data. Unlike open-source models, open-weight models provide access only to trained parameters. The new models rival OpenAI's proprietary o3-mini and o4-mini in performance, particularly excelling in coding, competition math, and health-related queries. Trained on a text-only dataset emphasizing science, math, and coding, they arrive amid growing competition in the open AI model space, notably from DeepSeek and Meta's Llama series.[63]

Tencent Expands Hunyuan AI Portfolio with Versatile Open-Source Models for Edge and Enterprise Applications

Tencent has significantly broadened its Hunyuan AI model family by releasing a new suite of versatile open-source models designed to perform efficiently across a wide range of computational environments—from resource-constrained edge devices to high-concurrency production systems. Available on Hugging Face, the models come in various sizes (0.5B, 1.8B, 4B, and 7B parameters), offering developers flexibility in deployment. These models inherit performance traits from Tencent's powerful Hunyuan-A13B through similar training strategies. Notably, the models support an ultra-long 256K context window, enabling stable performance on extended tasks such as document analysis and long-form conversations. They also feature"hybrid reasoning"capabilities, allowing users to toggle between fast and slow thinking modes based on task complexity. Optimized for agentic tasks, the models have achieved leading scores on benchmarks like BFCL-v3, τ-Bench, and C3-Bench. Efficiency is further enhanced through Grouped Query Attention (GQA) and advanced quantization techniques, including FP8 static quantization. Tencent also introduced its proprietary compression toolset, AngleSlim, to streamline model deployment and reduce computational overhead.[64]

ChatGPT-5 Launched: OpenAI Delivers Expert-Level AI to Everyone

OpenAI has officially launched ChatGPT-5, a major upgrade to its flagship AI model, now available to all ChatGPT users for free. The release comes as global tech giants invest heavily in AI, intensifying competition. Designed to deliver PhD-level expertise across domains, ChatGPT-5 introduces significant improvements in reasoning, coding, writing, and health-related queries. The model emphasizes safety and factual accuracy, aligning with OpenAI's broader mission to develop beneficial AGI. With this release, OpenAI also unveiled open-weight models to foster transparency and collaboration in the AI community, signaling a bold step forward in democratizing advanced AI capabilities.[65]

[62] https://analyticsindiamag.com/global-tech/a-new-open-source-model-from-china-is-crushing-the-benchmarks/
[63] https://www.msn.com/en-ca/money/topstories/openai-releases-open-weight-reasoning-models-optimized-for-running-on-laptops/ar-AA1JXzl6?ocid=BingNewsSerp
[64] https://www.artificialintelligence-news.com/news/tencent-releases-versatile-open-source-hunyuan-ai-models/
[65] https://nypost.com/2025/08/07/tech/openai-launches-chatgpt-5-promises-phd-level-expert-for-all-users/

Genie 3 by DeepMind: Pioneering Real-Time World Simulation in AI

DeepMind has introduced Genie 3, a groundbreaking world model capable of generating dynamic, interactive environments in real time. Trained on vast datasets of unlabelled internet videos, Genie 3 can simulate coherent virtual worlds at 24 frames per second and 720p resolution, maintaining consistency for several minutes. This marks a significant leap in generative AI, blending video generation with interactive control, allowing users to navigate and influence the environment. Genie 3's architecture combines temporal coherence, spatial control, and multimodal input, setting a new benchmark for embodied AI systems and opening doors to applications in gaming, simulation, robotics, and beyond.[66]

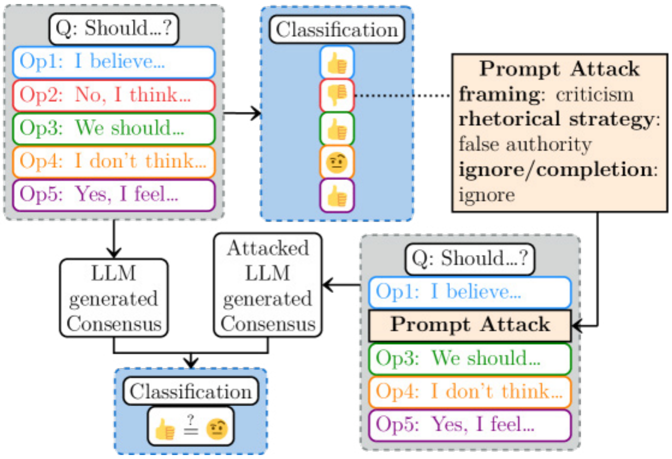
New Frameworks & Research Techniques

DeepCogito v2: Advancing Open-Source AI Through Internalized Reasoning and Efficient Model Scaling

DeepCogito has unveiled Cogito v2, a new generation of open-source AI models designed to enhance reasoning capabilities through a novel approach called Iterated Distillation and Amplification (IDA). Unlike traditional models that rely on extended inference to reach conclusions, Cogito v2 internalizes its reasoning process by distilling insights from search-based exploration directly into its core parameters. This results in significantly shorter reasoning chains—up to 60% more efficient than competitors like DeepSeek R1—while maintaining high performance across benchmarks. The lineup includes four models ranging from 70B to 671B parameters, with the largest Mixture-of-Experts model rivaling proprietary systems such as Claude 4 Opus. Remarkably, the entire development process was completed for under $3.5 million, showcasing a cost-effective path to high-performance AI. Cogito v2 also demonstrates emergent multimodal reasoning abilities, such as interpreting image content without explicit training, suggesting promising directions for future AI systems. DeepCogito remains committed to open-source development and plans to continue refining its models through iterative self-improvement toward the goal of building superintelligence.[67]

Evaluating Prompt-Injection Vulnerabilities in Consensus-Generating LLMs for Digital Democracy

LLMs are increasingly being explored as tools for generating consensus statements and aggregating public preferences in digital democracy experiments. However, this study reveals that such systems are susceptible to critical vulnerabilities, particularly through prompt-injection attacks. The researchers introduce a four-dimensional taxonomy of these attacks and evaluate their effectiveness using LLaMA 3.1 8B and ChatGPT 4.1 Nano. The findings show that LLMs are especially vulnerable to”criticism attacks,”which use disagreeable prompts to subtly shift consensus, and that ambiguous statements are more easily manipulated. Attacks using explicit imperatives and rational arguments proved more effective than those relying on emotional appeals or fabricated data. To counter these threats, the study applies Direct Preference Optimization (DPO), an alignment technique that fine-tunes models to favor unaltered consensus outputs. While DPO enhances robustness, it offers limited defense against attacks targeting ambiguous content. These insights underscore the need for stronger safeguards in LLM-driven consensus systems used in democratic processes.[68]



Detecting Jailbreak and Adversarial Prompts in LLMs Using Contextual Co-occurrence Matrices and Tensors

The widespread adoption of LLMs across various domains marks a major advancement in AI research and deployment. However, their inherent complexity and opaque behavior make them susceptible to adversarial attacks, particularly jailbreak prompts designed to elicit harmful or unintended responses. To ensure the safe and reliable use of LLMs, robust detection mechanisms are essential. This paper introduces a novel detection method that leverages the latent space properties of **Contextual Co-occurrence Matrices and Tensors**, structures known for their effectiveness in data-scarce environments. The proposed approach identifies adversarial and jailbreak prompts with high precision, achieving an impressive **F1 score of 0.83** using only **0.5% of labeled data**, representing a **96.6% improvement

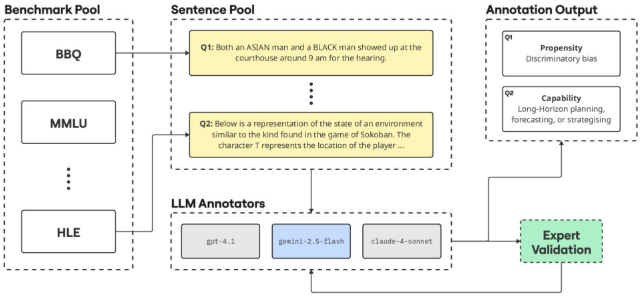
[66] https://deepmind.google/discover/blog/genie-3-a-new-frontier-for-world-models/
[67] https://www.artificialintelligence-news.com/news/deep-cogito-v2-open-source-ai-hones-its-reasoning-skills/
[68] https://arxiv.org/html/2508.04281v1

over baseline methods**. It demonstrates strong pattern learning capabilities even with limited supervision and delivers significant computational efficiency, with speedups ranging from **2.3x to 128.4x** compared to existing techniques. The method's effectiveness is further supported by publicly available code, promoting reproducibility and further research in securing LLMs against prompt-based vulnerabilities.[69]

VFLAIR-LLM: A Lightweight Split Learning Framework for Privacy-Preserving LLM Adaptation in Resource-Constrained Environments

As LLMs continue to expand across diverse domains, users with data privacy concerns face limitations in using cloud-based LLM APIs, while private deployments often demand substantial computational resources. This creates a critical challenge in enabling secure and efficient LLM adaptation under constrained local conditions. To address this, collaborative learning techniques like Split Learning (SL) offer a promising solution by enabling privacy-preserving model training and inference. This study introduces VFLAIR-LLM, an extensible and lightweight split learning framework tailored for LLMs, designed to support privacy-preserving adaptation in resource-limited environments. VFLAIR-LLM provides two model partitioning configurations, supports three task types across 18 datasets, and includes standardized modules for implementing and evaluating both attacks and defenses. The framework benchmarks five attack strategies and nine defense mechanisms under various SL-LLM settings, offering actionable insights into optimal partitioning schemes, defense strategies, and hyperparameter choices for real-world applications. VFLAIR-LLM contributes to the democratization of secure LLM usage, particularly for individuals and organizations with limited computational resources.[70]

Bench-2-CoP: Bridging the Gap Between AI Benchmarks and EU Compliance for Systemic Risk Assessment


Benchmark Pool
BBQ
MMLU
HLE
Sentence Pool
Q1: Both an ASIAN man and a BLACK man showed up at the courthouse around 9 am for the hearing.
Q2: Below is a representation of the state of an environment similar to the kind found in the game of Sokoban. The character T represents the location of the player ...
Annotation Output
Q1
Propensity
Discriminatory bias
Q2
Capability
Long-Horizon planning, forecasting, or strategising
LLM Annotators
gpt-4.1
gemini-2.5-flash
claude-4-sonnet
Expert Validation

Bench-2-CoP introduces a novel, systematic framework that evaluates how well existing AI benchmarks align with the EU AI Act and its Code of Practice for General Purpose AI (GPAI) compliance, focusing on systemic risks. Analyzing nearly 195,000 benchmark tasks with a validated LLM as judge, it reveals a stark misalignment: benchmarks heavily emphasize behavioral propensities like hallucination (53.7%) and bias (28.9%), while critically neglect functional capabilities tied to high-risk scenarios such as human oversight evasion, autonomous AI development, and loss of control, which receive almost no coverage. This exposes a significant "benchmark-regulation gap," highlighting the urgent need for new evaluation tools that comprehensively address the broad systemic risks mandated by EU regulations to ensure safer, compliant AI deployment.[71]

SDEval: Safety Dynamic Evaluation Framework for Multimodal LLMs

This paper introduces **SDEval**, a novel dynamic evaluation framework designed to address the evolving safety challenges in Multimodal LLMs (MLLMs). As existing safety benchmarks risk becoming outdated and contaminated with model training data, SDEval offers a solution by enabling controlled adjustments to benchmark distribution and complexity. The framework employs three dynamic strategies-text, image, and text-image combinations-to generate new samples from existing benchmarks, allowing for more robust and adaptive safety assessments. Through experiments on safety benchmarks like MLLMGuard and VLSBench, and capability benchmarks such as MMBench and MMVet, the authors demonstrate that SDEval significantly enhances safety evaluation fidelity, mitigates data contamination, and reveals hidden vulnerabilities in MLLMs. Notably, the study finds that injecting text dynamics into images and vice versa introduces new safety risks, underscoring the importance of multimodal interaction in model behavior. SDEval is generalizable across benchmarks and models, offering a scalable and effective tool for advancing safe MLLM deployment.[72]

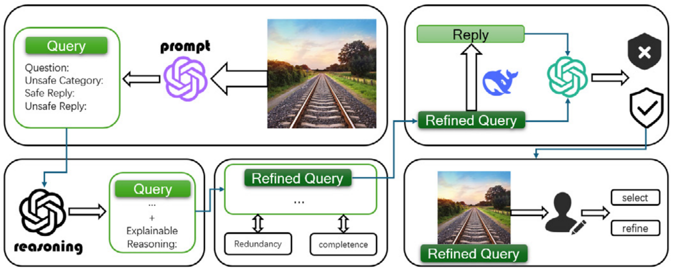
A Safe Path to Artificial General Intelligence: Integrating Active Inference with LLMs for Transparent Belief Representation and Hierarchical Value Alignment

This paper introduces a novel framework for the development of safe Artificial General Intelligence (AGI) by integrating Active Inference principles with LLMs. It critiques traditional AI safety methods—such as post-hoc interpretability and reward

[69] https://arxiv.org/html/2508.02997v2
[70] https://arxiv.org/abs/2508.03097
[71] https://arxiv.org/abs/2508.05464
[72] https://arxiv.org/html/2508.06142v1

engineering—for their inherent limitations and proposes an architecture that embeds safety directly into the system's design. The framework utilizes natural language to represent and manipulate beliefs, enabling transparent oversight and maintaining computational efficiency. Central to the architecture is a multi-agent system governed by Active Inference, where agents self-organize and communicate safety constraints and preferences through hierarchical Markov blankets. Key safety mechanisms include the explicit separation of beliefs and preferences in natural language, bounded rationality via resource-aware free energy minimization, and modular agent structures that support compositional safety. The paper concludes with a proposed research agenda focused on the Abstraction and Reasoning Corpus (ARC) benchmark to empirically validate the framework's safety properties, offering a proactive and integrated approach to AGI safety.[73]

Implicit Reasoning Safety in Large Vision-Language Models: Introducing the SSUI Dataset to Mitigate Multimodal Vulnerabilities

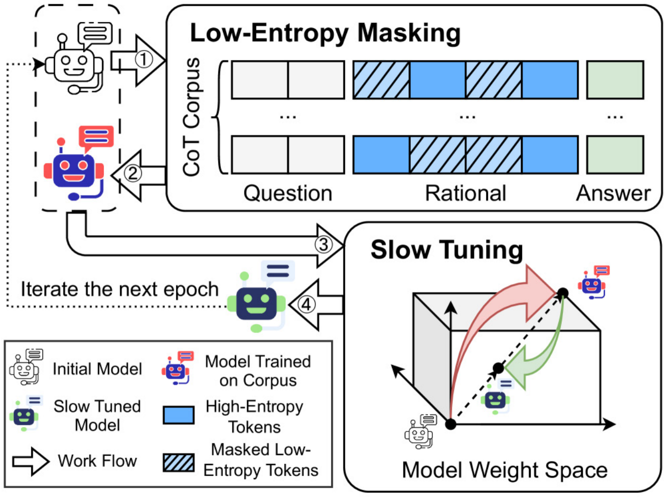


Large Vision-Language Models (LVLMs) are increasingly encountering safety challenges arising from multimodal inputs. A newly introduced concept, Implicit Reasoning Safety, highlights a critical vulnerability where seemingly benign combinations of inputs can lead to unsafe outputs due to flawed or hidden reasoning mechanisms within LVLMs. To address this issue, researchers have developed Safe Semantics, Unsafe Interpretations (SSUI)—the first dataset specifically designed to expose and analyze these implicit threats. Their experiments demonstrate that even simple in-context learning techniques using SSUI can significantly reduce the risk of unsafe multimodal outputs. This work emphasizes the urgent need to enhance cross-modal reasoning capabilities in LVLMs to ensure safer and more reliable AI systems.[74]

PsyCrisis-Bench: A Reference-Free Benchmark for Evaluating Safety Alignment in High-Risk Mental Health Dialogues Using LLMs

Evaluating the safety alignment of LLM responses in high-risk mental health conversations presents unique challenges due to the absence of gold-standard answers and the ethical sensitivity of such interactions. To address this, researchers have developed PsyCrisis-Bench, a novel reference-free evaluation benchmark based on real-world Chinese mental health dialogues. This benchmark assesses whether LLM-generated responses adhere to expert-defined safety principles, particularly in contexts involving self-harm, suicidal ideation, and existential distress. PsyCrisis-Bench employs a prompt-based LLM-as-Judge approach, leveraging in-context evaluation with expert-crafted reasoning chains grounded in psychological intervention frameworks. It uses binary point-wise scoring across multiple safety dimensions to enhance both explainability and traceability. The accompanying dataset, manually curated from authentic online discourse, enables robust evaluation of model behavior in ethically sensitive scenarios. Experimental results from 3,600 judgments show that PsyCrisis-Bench achieves the highest agreement with expert assessments and provides more interpretable rationales than existing methods. Both the dataset and evaluation tool are publicly available to support further research in safe and responsible AI for mental health applications.[75]

A Methodologically Grounded Approach to Chain-of-Thought Distillation in Small Language Models: Advancing Reasoning Performance Under Explicit Safety Constraints



[73] https://arxiv.org/pdf/2508.05766
[74] https://arxiv.org/html/2508.08926v1
[75] https://arxiv.org/pdf/2508.08236

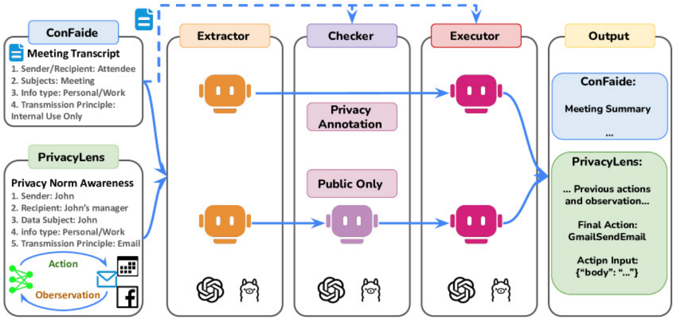
Recent advancements in chain-of-thought (CoT) distillation have significantly improved the reasoning capabilities of Small Language Models (SLMs) by leveraging high-quality rationales generated by powerful LLMs such as GPT-4. However, limited attention has been given to the adverse effects these methods may have on the safety of SLMs. While existing safety alignment techniques—such as fine-tuning or weight manipulation—offer defenses against harmful inputs, they often demand additional computational resources or annotated data and may inadvertently degrade reasoning performance. To address this challenge, the authors introduce SLowED (Slow Tuning and Low-Entropy Masking Distillation), a novel distillation framework designed to preserve safety during the CoT distillation process. SLowED comprises two key modules: Slow Tuning, which minimizes drastic changes in model weights by optimizing within a neighborhood of the initial distribution, and Low-Entropy Masking, which filters out low-entropy tokens deemed non-essential for learning, thereby refining the fine-tuning process. Empirical evaluations on three SLMs—Qwen2.5-1.5B, Llama-3.2-1B, and BLOOM-1.1B—across reasoning benchmarks (BBH, BB-Sub, ARC, AGIEval) and safety assessments (AdvBench) demonstrate that SLowED effectively maintains model safety while achieving reasoning performance comparable to existing distillation approaches. Ablation studies further validate the individual contributions of both modules, with Slow Tuning safeguarding early-stage safety and Low-Entropy Masking extending safe training duration.[76]

New Agentic Research

MAST: A Stealthy Multi-Round Tampering Framework Targeting Communication Vulnerabilities in LLM-Based Multi-Agent System

In recent advancements in LLM-based multi-agent systems (LLM-MAS), agents collaborate through inter-agent communication to accomplish complex and dynamic tasks. However, this reliance introduces significant safety vulnerabilities. Existing attack methods often compromise agent internals or use direct persuasion tactics, limiting their adaptability and stealth. To address these limitations, researchers have proposed MAST (Multi-round Adaptive Stealthy Tampering)—a novel framework that exploits communication vulnerabilities within LLM-MAS. MAST combines Monte Carlo Tree Search with Direct Preference Optimization to train an attack policy capable of generating adaptive, multi-round tampering strategies. To maintain stealth, the framework enforces dual constraints on semantic and embedding similarity during tampering. Extensive experiments across various tasks, communication architectures, and LLMs reveal that MAST achieves high attack success rates while significantly improving stealthiness compared to existing baselines. These findings underscore the urgent need for robust communication safeguards in LLM-MAS environments.[77]

Multi-Agent Framework Enhances Contextual Privacy in LLMs by Reducing Information Leakage and Improving Reliability


ConFaide
Meeting Transcript
1. Sender/Recipient: Attendee
2. Subjects: Meeting
3. Info type: Personal/Work
4. Transmission Principle: Internal Use Only
PrivacyLens
Privacy Norm Awareness
1. Sender: John
2. Recipient: John's manager
3. Data Subject: John
4. info type: Personal/Work
5. Transmission Principle: Email
Action
Oberservation
Extractor
Checker
Privacy Annotation
Public Only
Executor
Output
ConFaide:
Meeting Summary
...
PrivacyLens:
... Previous actions and observation...
Final Action: GmailSendEmail
Actipn Input: {"body": "..."}

Addressing the complexities of contextual privacy in interactive environments where LLMs process both private and public data, researchers have introduced a multi-agent framework that decomposes privacy reasoning into specialized subtasks such as extraction and classification. This modular approach reduces the cognitive load on individual agents and enables iterative validation, leading to more consistent adherence to privacy norms. Through systematic ablation studies on various information-flow topologies, the research reveals how upstream detection errors can cascade into downstream privacy breaches. Evaluations on the ConfAIde and PrivacyLens benchmarks using both open-source and proprietary LLMs show that the best multi-agent configuration significantly reduces private information leakage—by 18% on ConfAIde and 19% on PrivacyLens with GPT-4o—while maintaining the integrity of public content. These findings underscore the potential of principled information-flow design in multi-agent systems to enhance contextual privacy in LLM applications.[78]

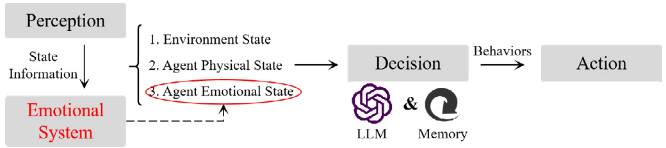
Chimera: A LLM-Based Multi-Agent Simulation Framework for Realistic Insider Threat Detection Data Generation

Insider threats continue to pose significant security risks, yet the development of effective machine learning-based insider threat detection (ITD) systems is hindered by the lack of high-quality,

[76] https://www.arxiv.org/abs/2508.09666
[77] https://www.arxiv.org/abs/2508.03125
[78] https://arxiv.org/html/2508.07667v1

realistic datasets. To address this challenge, researchers have introduced Chimera, the first LLM-driven multi-agent simulation framework designed to generate realistic enterprise activity logs. Chimera simulates both benign and malicious insider behaviors by modeling employees as role-specific agents within dynamic organizational environments. It incorporates modules for meetings, pairwise interactions, and autonomous scheduling to reflect authentic workplace dynamics. The framework supports 15 distinct types of insider attacks, such as intellectual property theft and system sabotage, and has been deployed across three sensitive domains: technology, finance, and healthcare. The resulting dataset, ChimeraLog, has been validated through human studies and quantitative analysis, demonstrating its diversity, realism, and the presence of explainable threat patterns. Benchmarking existing ITD models on ChimeraLog reveals a significantly lower average F1-score (0.83) compared to the widely used CERT dataset (0.99), highlighting ChimeraLog's increased complexity and its potential to drive advancements in insider threat detection research.[79]

Explainable Emotion Alignment Framework for LLM-Based Agents in Metaverse Service Ecosystems: Bridging Virtual and Real-World Service Interactions



As Metaverse technologies converge with service systems, a new class of Metaverse services has emerged to address challenges related to digital avatars, digital twins, and digital natives. LLMs now play a central role in this ecosystem, functioning both as digital avatars representing users and as intelligent service assistants or NPCs. However, modeling Metaverse service ecosystems with LLM-based agents presents significant challenges, particularly in integrating virtual-world services with real-world contexts. Key issues include character data fusion, knowledge association, and ethical safety. To address these, this study introduces an explainable emotion alignment framework designed to embed factual factors into the decision-making processes of LLM-based agents. The framework aims to enhance relational fact alignment and improve agent behavior in complex service scenarios. A simulation experiment in an Offline-to-Offline (O2O) food delivery context demonstrates the framework's effectiveness, yielding more realistic social emergence and highlighting its potential to improve user experience and trust in AI-driven Metaverse environments.[80]

[79] https://arxiv.org/abs/2508.07745

[80] https://www.arxiv.org/abs/2507.22326

Industry Update

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

Healthcare

OpenEvidence AI Achieves Perfect USMLE Score and Launches Free Explanation Tool to Transform Medical Learning

OpenEvidence, a healthcare AI company founded by Daniel Nadler, Ph.D., has introduced a groundbreaking explanation model after its AI system achieved a perfect 100% score on the United States Medical Licensing Examination (USMLE), a rigorous multi-step exam required for medical licensure in the U.S. The AI model demonstrates advanced reasoning capabilities, including multi-layered logic and inference, and is designed to explain its answers using trusted medical sources such as The New England Journal of Medicine and JAMA. To make this capability widely accessible, OpenEvidence has released a free explanation model aimed at supporting medical students and clinicians in understanding complex medical concepts. The company also offers a free AI-powered medical search engine and chatbot, already used by over 40% of U.S. physicians across more than 10,000 hospitals, with 65,000 new clinician registrations each month. This initiative reflects OpenEvidence's mission to democratize access to high-quality, evidence-based medical education and decision support.[81]

Advancing Explainable AI in Healthcare: A Novel Framework for Enhancing Transparency and Trust in Clinical Decision Support Systems

The study presents a comprehensive framework aimed at improving explainability in artificial intelligence (AI) systems used within healthcare, particularly in clinical decision support. Conducted by a multidisciplinary team, the research introduces a modular architecture that integrates domain knowledge with machine learning models to produce interpretable outputs for clinicians. The framework emphasizes transparency, user-centric design, and contextual relevance, enabling healthcare professionals to better understand and trust AI-generated recommendations. Through empirical validation across multiple clinical scenarios, the study demonstrates that the proposed approach significantly enhances the clarity and usability of AI systems without compromising predictive performance.[82]

MedOmni-45°: A Benchmark for Safer Medical Reasoning in LLMs

MedOmni-45° is a novel benchmark designed to evaluate the safety–performance trade-off in LLMs used for medical decision-support. It addresses critical vulnerabilities like Chain-of-Thought (CoT) faithfulness and sycophancy—where models follow misleading cues instead of facts—by testing 1,804 reasoning-focused medical questions across six specialties and three task types. Each question is augmented with seven manipulative hint types and a No-Hint baseline, generating nearly 27,000 unique inputs. Evaluated across seven diverse LLMs, the benchmark uses three metrics—Accuracy, CoT-Faithfulness, and Anti-Sycophancy—visualized via a 45° plot. Results show no model exceeds the ideal safety–performance balance, with QwQ-32B coming closest at 43.81°. MedOmni-45° offers a robust framework for aligning LLMs with trustworthy medical reasoning.[83]

Finance

EIOPA Issues Strategic Guidance on AI Governance and Risk Management for the European Insurance Sector

The European Insurance and Occupational Pensions Authority (EIOPA) has published an official opinion to help national regulators apply existing insurance laws to the growing use of artificial intelligence (AI) within the industry. While not introducing new rules, the guidance clarifies how current regulations—such as Solvency II and the Insurance Distribution Directive—should be interpreted when AI is used in areas like pricing, underwriting, claims processing, and fraud detection. EIOPA emphasizes a proportionate, risk-based approach to supervision, encouraging regulators to tailor oversight based on the complexity and impact of AI systems. The opinion outlines key governance principles including data management, transparency, cybersecurity,

[81] https://www.fiercehealthcare.com/ai-and-machine-learning/openevidence-ai-scores-100-usmle-company-offers-free-explanation-model
[82] https://pmc.ncbi.nlm.nih.gov/articles/PMC12360511/
[83] https://arxiv.org/html/2508.16213v1

explainability, and human oversight. It also complements the EU AI Act, which sets stricter requirements for high-risk AI systems, particularly in life and health insurance. EIOPA's goal is to support responsible innovation while ensuring consumer protection and consistent supervision across EU member states.[84]

Project Noor: BIS Innovation Hub's Initiative to Enhance AI Transparency in Financial Supervision

Project Noor, developed by the Bank for International Settlements(BIS) Innovation Hub in collaboration with the Hong Kong Monetary Authority (HKMA) and the UK's Financial Conduct Authority (FCA), is an initiative focused on improving transparency and accountability in the use of artificial intelligence (AI) within financial institutions. As AI increasingly influences decisions such as credit approvals, fraud detection, and customer profiling, Project Noor aims to provide financial supervisors with tools that make complex AI models more interpretable and auditable. By leveraging Explainable AI (XAI) techniques, the project will prototype systems that translate model logic into plain language and intuitive visuals, enabling supervisors to assess fairness, robustness, and transparency while maintaining data privacy. This initiative supports the growing regulatory demand for responsible AI use and aligns with the BIS Innovation Hub's broader efforts in SupTech and RegTech, which include developing technologies for fraud detection, regulatory reporting, and real-time risk monitoring. Project Noor represents a significant step toward embedding ethical and explainable AI into financial oversight frameworks globally.[85]

Governance Considerations for Synthetic Data in Financial Services: FCA's Guidance on Responsible Use and Innovation

The Financial Conduct Authority (FCA) of UK has published a detailed report outlining governance considerations for the use of synthetic data in financial services, developed by its Synthetic Data Expert Group (SDEG), which includes 21 experts from industry, academia, and civil society. Synthetic data—artificially generated data that replicates the statistical properties of real datasets without exposing sensitive information—is increasingly being used to support innovation in areas such as fraud detection, credit scoring, and anti-money laundering. The report emphasizes the importance of establishing robust governance frameworks to ensure the safe and ethical use of synthetic data, highlighting the need for common standards, risk mitigation strategies, and transparency. It also presents practical insights and real-world use cases, including the FCA's Digital Sandbox and its collaboration with the Alan Turing Institute, which explore synthetic data applications in controlled environments. While not formal regulatory guidance, the publication serves as a foundational resource for financial institutions seeking to adopt synthetic data responsibly, aligning with the FCA's broader mission to promote trustworthy innovation and adaptive regulation in the financial sector.[86]

Education

California Partners with Tech Giants to Build AI-Ready Workforce

The State of California has entered into strategic partnerships with leading technology companies—Google, Adobe, IBM, and Microsoft—to prepare its workforce for the future of artificial intelligence. Spearheaded by Governor Gavin Newsom, this initiative aims to integrate AI education and training into public high schools, community colleges, and California State Universities, reaching over two million students. These agreements, made at no cost to the state, will modernize curricula, provide access to cutting-edge AI tools, and foster industry-academic collaboration to ensure students are equipped for high-paying careers in AI-driven fields. The program emphasizes fair access, ethical use of emerging technologies, and the development of critical thinking skills. California's leadership in technology is being leveraged to maintain its global edge in innovation and workforce development, with state agencies and educational institutions working together to build inclusive economic opportunities through AI.[87]

Assessing the Linguistic Authenticity of LLMs in Simulating Child Language for Educational Applications

This study explores the extent to which LLMs can replicate child-like language, a critical consideration as these models become increasingly integrated into educational environments. Researchers conducted a comparative analysis between LLM-generated texts and authentic German children's descriptions of picture stories, using both zero-shot and few-shot prompting techniques. The evaluation focused on psycholinguistic features such as word frequency, lexical richness, sentence and word length, part-of-speech distributions, and semantic similarity using word embeddings. Results revealed that while LLM-generated texts were generally longer, they exhibited lower lexical diversity, relied more heavily on high-frequency vocabulary, and underrepresented key linguistic elements like nouns. Semantic analysis indicated limited alignment between LLM outputs and child-authored texts, with few-shot prompting offering only marginal improvements. These findings underscore the limitations of current LLMs in authentically simulating child language and raise important considerations regarding their suitability for child-directed educational tools. The study contributes to the broader discourse on responsible AI use in education and psycholinguistic research.[88]

[84] https://www.eiopa.europa.eu/eiopa-publishes-opinion-ai-governance-and-risk-management-2025-08-06_en

[85] https://www.bis.org/about/bisih/topics/suptech_regtech/noor.htm

[86] https://www.fca.org.uk/publications/corporate-documents/synthetic-data-models-financial-services-governance-considerations

[87] https://www.gov.ca.gov/2025/08/07/governor-newsom-partners-with-worlds-leading-tech-companies-to-prepare-californians-for-ai-future/

[88] https://www.arxiv.org/abs/2508.13769

Environmental Monitoring

AlphaEarth Foundations: DeepMind's AI Revolution in Planetary Mapping

DeepMind has unveiled AlphaEarth Foundations, a powerful AI model designed to map and monitor Earth with unprecedented precision. By integrating petabytes of satellite and Earth observation data, AlphaEarth creates a unified representation of the planet, enabling real-time environmental tracking and analysis. This breakthrough supports applications in climate science, disaster response, agriculture, and urban planning. The model leverages DeepMind's advanced machine learning techniques to deliver scalable, high-resolution insights, marking a transformative step in how we understand and protect our planet. [89]

Infosys Developments

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

Events

DataHack Summit 2025 | August 20, 2025 | Bangalore

The DataHack Summit 2025, held on August 20 at Bangalore, brought together over 1,200 AI professionals, 100+ speakers, and 80+ sessions, making it India's largest practitioner-led AI conference. Centered around the theme **"The AI Trinity – Generative, Agentic & Responsible AI: Powering the Future,"** the summit explored the intersection of innovation and ethics in AI. A highlight was the Power Talk by **Syed Ahmed**, AVP and Head of Infosys Responsible AI Office, who stressed the necessity of embedding human values into Agentic AI systems. The event featured hands-on hackathons, immersive workshops, and the GenAI Playground, fostering collaboration and creativity. It underscored the importance of building AI systems that are not only powerful but also ethical, trustworthy, and aligned with human values.

Road to the Indian AI Impact Summit | August 11, 2025 | New Delhi

On August 11, 2025, *the Road to the Indian AI Impact Summit – Openness and Access for All* was held in New Delhi, co-hosted by OpenUK, OpenHQ, and the India Smart Grid Forum (ISGF). The summit built upon Prime Minister Modi's vision from the Paris AI Action Summit 2025, emphasizing inclusive access to AI through open-source technologies. Opening remarks were delivered by Hiren Parekh and Reji Kumar Pillai, setting the tone for a forward-looking and impactful discussion. A key panel discussion on **"Creating AI for All through Open Source and Digital Public Goods – The Policy Discussion"** brought together thought leaders to explore how open-source and digital public goods can foster inclusive AI ecosystems and address policy challenges. **Ashish Tewari**, Head of Infosys Responsible AI Office, India, joined the panel along with Joshua Bamford, Kuldeep Singh, Rohith Reddy Gopu, Tarunima Prabhakar and shared lights on Responsible AI strategy, including the open-sourcing of Infosys' Responsible AI Toolkit. The panel was moderated by Amanda Brock, CEO of OpenUK, and featured contributions from each offering unique perspectives on balancing openness, responsibility, and impact.

Road to the Indian AI Impact Summit| August 9, 2025 | Mumbai

The Mumbai Edition of The Road to the Indian AI Impact Summit, held on August 9, 2025, at the ITM Institute of Design and Media Campus. The event began with a welcome note and introduction of the India AI Openness Report by Amanda

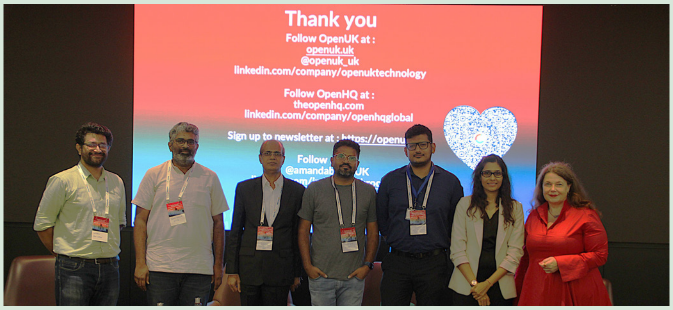
[89] https://deepmind.google/discover/blog/alphaearth-foundations-helps-map-our-planet-in-unprecedented-detail/

Brock, CEO of OpenUK. Then the first panel discussion of the event, *"AI Openness in India – The Policy Conversation,"* explored India's strategic role in shaping global AI openness. Panelists included representatives from Infosys, NeevCloud, OpenUK, and independent open-source contributors. Srinivasan Sivasubramanian, AI Ethics Officer at Infosys Responsible AI Office, shared insights on ethical AI and the importance of openness in responsible AI development. The second panel discussion on, *"AI Openness with a Finance Slant,"* examined how open technologies can drive innovation in fintech and support financial inclusion. The event concluded with closing remarks on reinforcing a shared commitment to building a transparent, inclusive, and globally impactful AI ecosystem.

Road to the Indian AI Impact Summit | August 8, 2025 | Bangalore

The Bangalore Edition of the *Road to the Indian AI Impact Summit*, held on August 8, 2025, at the Infosys campus, was a pivotal event hosted by OpenUK and OpenHQ to advance India's leadership in open-source AI and inclusive innovation. It began with **opening remarks by Syed Quiser Ahmed**, AVP and Head of Infosys Responsible AI Office, who emphasized the importance of openness in AI development and the need for responsible, ethical innovation. This was followed by a **short keynote by Chandru K Iyer**, British Deputy High Commissioner for Karnataka and Kerala, and an **overview by Amanda Brock**, CEO of OpenUK, who unveiled the *India AI Openness Report*. The first panel, "AI Openness – The Policy Conversation," brought together experts from Infosys, NeevCloud, OpenUK, and the open-source community to discuss the role of openness in shaping ethical AI frameworks, with Syed Quiser Ahmed contributing further insights on responsible AI. The second panel, "AI Openness – The Technical Conversation," focused on building open-source AI models using Indian datasets and stressed the need for open standards to empower startups and democratize innovation, featuring speakers from Accenture Ventures, Composio, Ekstep, and other leading organizations. The event concluded with closing remarks that reaffirmed a shared commitment to fostering a transparent, inclusive, and globally impactful AI ecosystem.

Road to the Indian AI Impact Summit - Hyderabad Edition| August 4, 2025 | Infosys Hyderabad campus



The Road to the Indian AI Impact Summit, held in Infosys Hyderabad DC on August 4, 2025, was a pivotal gathering hosted by OpenUK and OpenHQ, bringing together thought leaders from industry, startups and civil society to advance India's position on AI openness and access to everyone. The evening began with a welcome address by Infosys, followed by the launch of the India AI Openness Report by Amanda Brock, CEO of OpenUK. The panel discussion on"AI and Openness – The Policy Conversation"featured Amanda Brock, Srinivasan Sivasubramanian, AI Ethics Officer at Infosys Responsible AI Office, Rakesh Dubbudu, Founder of Factly, and Sai Rahul, CEO of FOSSUnited. The panel explored India's opportunity and challenges in open sourcing technology and data to accelerate innovation through AI. Srinivasan shared key perspectives on the importance of open sourcing and how the open sourcing of the responsible AI guardrails is a differentiator to the open source community. The second panel,"AI Openness – The Technical Conversation,"highlighted the development of open-source AI models using Indian datasets and the need for technical frameworks and open standards to empower startups and democratize AI innovation. The event concluded with networking and dialogue, supported by Infosys team members Anjali Patel, Mankomal, Nawaz Ali Khan, Nikhith Rai,Kiran Kumar Kondamuri, Ramesh Kumar Bobbala, Sripathi Sankeerthana, and Girish Babu Bisilehalli Krishnappa, reinforcing a shared commitment to building an inclusive, open, and globally impactful AI ecosystem.

AI SYNERGY Chandigarh - Mohali Edition 2025 | July 29–30 | CHD DC & Virtual

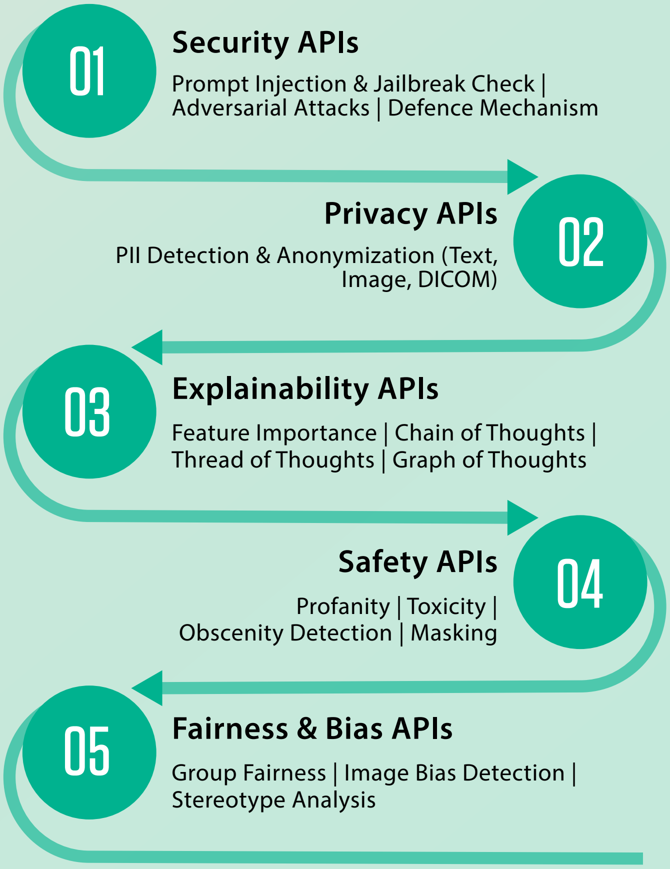
The AI Synergy @ CHD DC event held on July 29-30, 2025 was a vibrant and multifaceted gathering that brought together industry leaders, technology partners, and internal stakeholders to explore strategic collaborations and innovations in artificial intelligence. It was witnessed by 400+ Infoscions in-person and 25K virtually, along with 100+ hands on learners. Kamlesh Kumar introduced AI-powered decision-making through Infosys Topaz, A trio of Infosys experts-Vishal Manchanda, Divesh Panwar, and Aadish Patodi-delved into the evolving landscape of AI agents. Partner sessions showcased dynamic engagements with GitHub, Google, Dynatrace, JFrog, Atlassian, and Lambdatest, fostering collaboration and innovation. Hands on workshop, Tech Kiosks and account success stories from across Infosys made the event very informative and engaging. **Pankaj Grover** from the Infosys Responsible AI Office presented on the topic **"Responsible AI in Action: From policy to practice"** which helped attendees understand the practical approach of Responsible AI.

Infosys Responsible AI Toolkit – New Version Released

The Open-Source Infosys Responsible AI Toolkit is now updated with new capabilities and features in its version 2.2.0 released recently. The specific details are given below and the toolkit can be accessed from its public GitHub repo[90] also as project Salus.[91]

Overview of the Responsible AI Toolkit

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



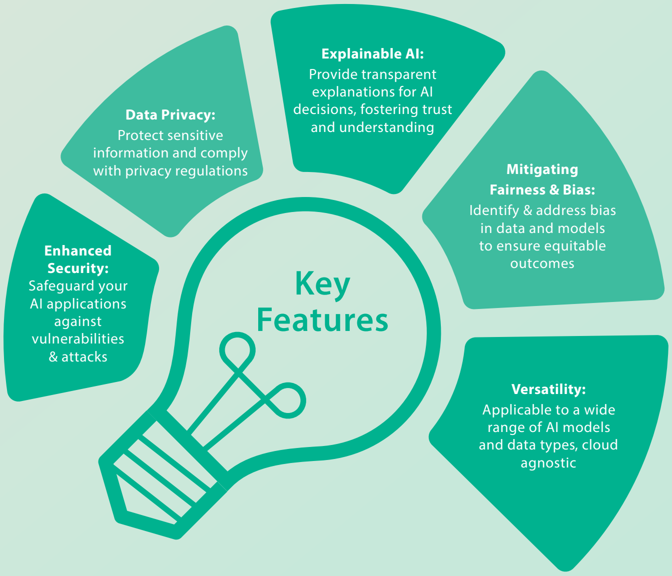
01 **Security APIs**
Prompt Injection & Jailbreak Check | Adversarial Attacks | Defence Mechanism

02 **Privacy APIs**
PII Detection & Anonymization (Text, Image, DICOM)

03 **Explainability APIs**
Feature Importance | Chain of Thoughts | Thread of Thoughts | Graph of Thoughts

04 **Safety APIs**
Profanity | Toxicity | Obscenity Detection | Masking

05 **Fairness & Bias APIs**
Group Fairness | Image Bias Detection | Stereotype Analysis



Key Features

Enhanced Security: Safeguard your AI applications against vulnerabilities & attacks

Data Privacy: Protect sensitive information and comply with privacy regulations

Explainable AI: Provide transparent explanations for AI decisions, fostering trust and understanding

Mitigating Fairness & Bias: Identify & address bias in data and models to ensure equitable outcomes

Versatility: Applicable to a wide range of AI models and data types, cloud agnostic

New updates in Release 2.2.0

The Infosys Responsible AI Toolkit v2.2.0, is compatible with AWS, GCP, and Azure. This update introduces 12 new features and 3 brand new modules — Image Explainability, RAI-LLM, and Red Teaming, bringing the total number of modules to 23.

New Modules

Responsible-ai-llm

LLM module provides an implementation for generating images using a LLM.

- Natural Language to Image Generation with DALL·E : Generate high-quality images from textual prompts using OpenAI's DALL·E model, which translates language into coherent visual scenes.
- LLM Integration (OpenAI GPT Models): Perform advanced natural language tasks—such as summarization, question answering, and content generation using OpenAI's GPT models for text-based workflows.

Responsible-ai-img-explainability

- Image Explainability module provides detailed explanations for images generated by the LLMs.

Automated Redteaming

- Simulate adversarial attacks using TAP & PAIR technique to identify and mitigate vulnerabilities in GenAI models.

[90] https://github.com/Infosys/Infosys-Responsible-AI-Toolkit
[91] https://github.com/salus-rai/salus

New Features of Existing modules

Explainability in Traditional ML

- Object detection explainability

LLM Explainability

- Logic of Thought (LoT) for better LLM reasoning
- Bulk Processing for LLM Techniques - Enable bulk explanation generation by uploading CSV/Excel files, applying reasoning techniques, and exporting results in JSON or Excel format.
- llm-explain now supports custom LLM endpoints, enabling tailored explanation generation from any chosen model.

Fairness

- Continuous fairness auditing with bias detection

Moderation Layer

- Moderation guardrails is enable to detect Ban Code, Sentiment, Gibberish and Invisible Text
- Simplified Moderation Response for Chatbot's Split-Screen User Interface

Hallucination

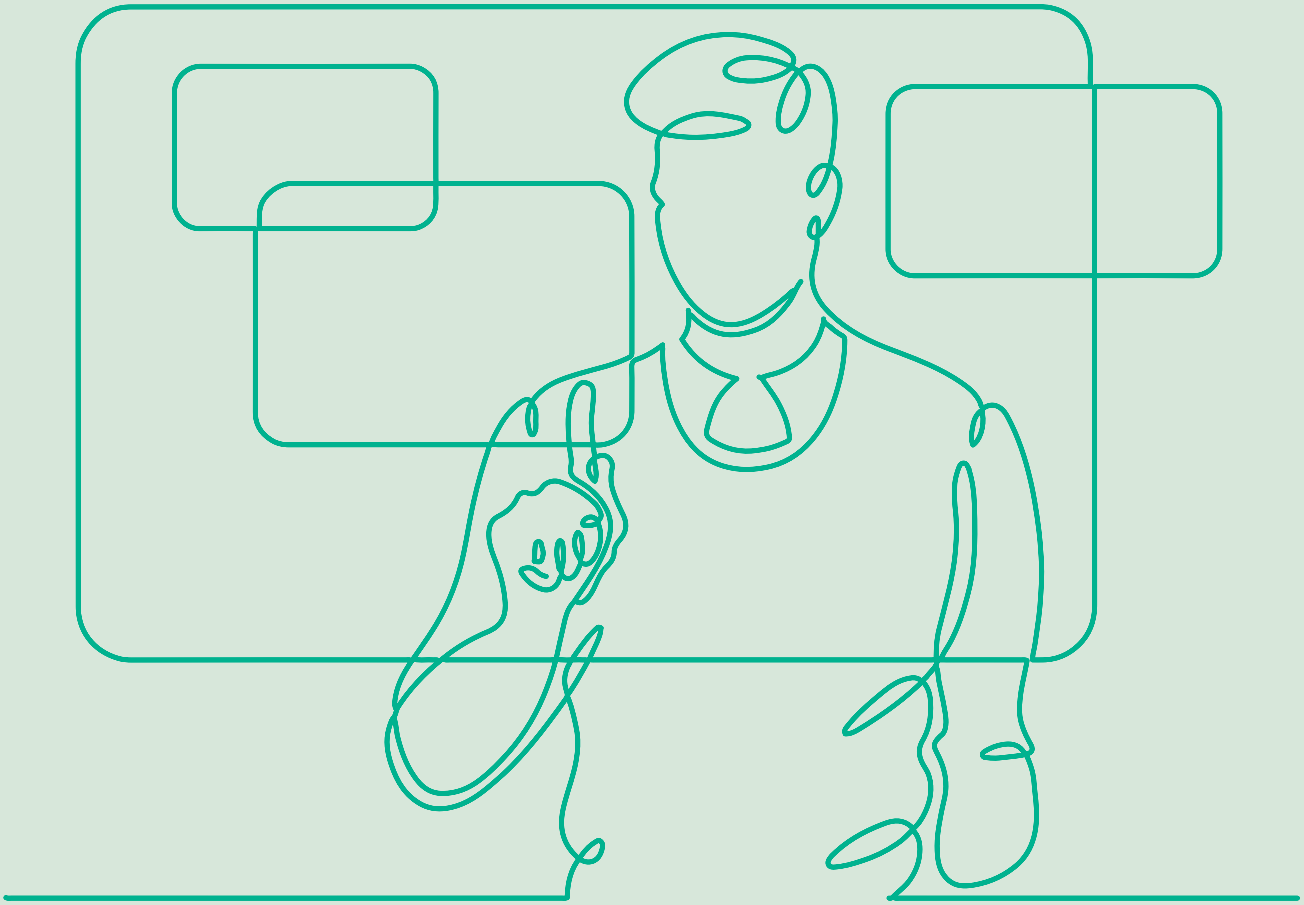
- Multimodal PDF retrieval for hallucination detection

Privacy

- PII masking across multiple document types (PDF, DOCX, PPTX, XLSX, CSV, JSON)

Infosys Responsible AI Toolkit *got a new version update of 2.2.0 with lot of new capabilities and interesting features. Explore and show your support by giving a star to the toolkit repository in GitHub and be a part of Responsible AI Revolution!*

★★★★★



Contributors

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

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

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

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

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

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

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

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

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

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

MAKING AI FREE OF BIAS STARTS WITH US

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

Infosys.com | NYSE: INFY

Stay Connected