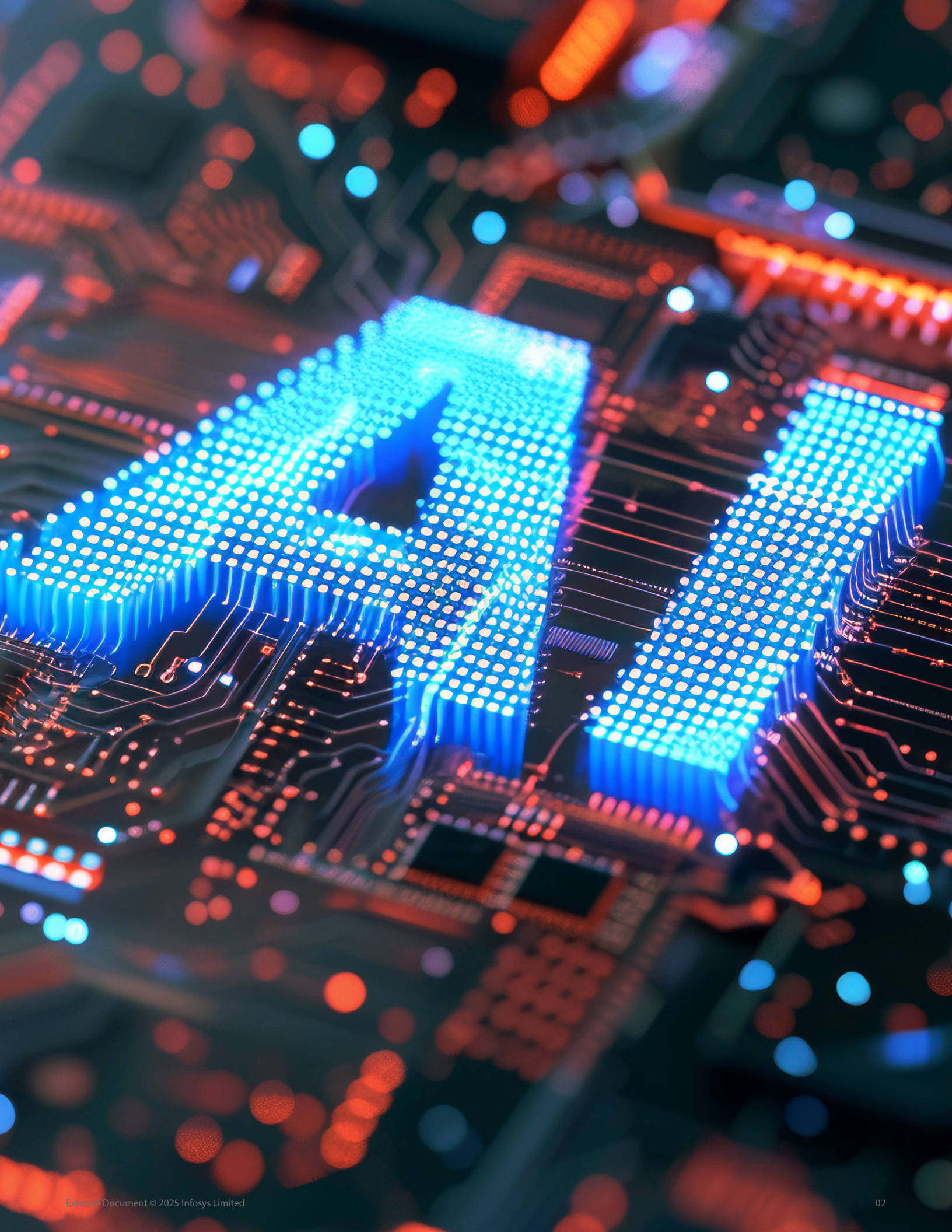
MARKET SCAN REPORT

FEBRUARY 2025

Infosys topaz

INFOSYS TOPAZ RESPONSIBLE AI SUMMIT 2025: LAUNCH OF OPEN-SOURCE RESPONSIBLE AI TOOLKIT

Infosys®
Navigate your next



Dear Readers,

As artificial intelligence continues to evolve at an unprecedented pace, the need for robust governance, ethical standards, and transparent frameworks has never been more critical. In this edition of our Market Scan Report, we bring you the latest developments in AI governance, emerging technologies, regulatory shifts, and industry best practices shaping the responsible AI landscape.

February has been a pivotal month in the AI ecosystem, with several global events shaping the discourse around AI governance. The European Union has moved closer to finalizing the AI Act, setting a precedent for comprehensive regulatory oversight. In the United States, AI innovation remains a priority, with increased investments and regulatory streamlining to enhance competitiveness while ensuring responsible adoption. Additionally, the release of OpenAI's latest model, GPT-4 Turbo, has sparked discussions around the balance between AI innovation and ethical considerations, particularly regarding misinformation and bias.

At the same time, the rise in deepfake incidents has raised significant concerns, particularly in political and social contexts, where AI-generated synthetic media is being used to spread disinformation. Several high-profile cases have demonstrated the urgent need for better detection mechanisms and regulatory interventions. Additionally, security vulnerabilities discovered in DeepSeek, an emerging AI model, have highlighted risks related to data privacy and model robustness, emphasizing the importance of rigorous security evaluations in AI deployments.

One of the most significant highlights of this month's report is Infosys' commitment to fostering responsible AI by open-sourcing our Responsible AI Toolkit. By making this toolkit publicly available, we aim to empower organizations, researchers, and policymakers to build AI systems that are ethical, explainable, and aligned with human values. This initiative underscores our belief that transparency and collaboration are key to ensuring AI's positive societal impact.

This report also covers notable advancements in AI regulations across key global markets, updates on evolving AI standards, and insights into how enterprises are navigating the complexities of AI governance. With perspectives from industry experts and thought leaders, we provide a comprehensive view of the trends influencing AI policy and implementation. Additionally, we analyze recent AI incidents, including controversies surrounding deepfake misuse and concerns over AI-driven automation replacing human jobs, highlighting the need for proactive governance measures.

We invite you to explore the insights shared in this edition and, more importantly, to engage with the open-source Responsible AI Toolkit. Your feedback is invaluable in refining and enhancing its impact. Please share your thoughts, experiences, and suggestions as we continue to shape a future where AI operates with accountability and trust.

We look forward to your contributions and continued dialogue on responsible AI.

Warm regards

Syed Ahmed
Head- Infosys Responsible AI Office

Contents

AI Regulations, Governance and Standards

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

AI Regulations and Governance across globe

Global

Paris AI Summit Co-Chaired by France and Indian Prime Minister Draws World Leaders and CEOs

The AI Action Summit in Paris, held from February 10-11, 2025, brought together nearly a hundred countries and over a thousand stakeholders from the private sector and civil society. Notably, Indian Prime Minister Narendra Modi co-chaired the AI Action Summit in Paris on February 11, 2025, with French President Emmanuel Macron. Here are some key outcomes:

1. International AI Safety Report: The first of its kind, this report compiles expert perspectives on AI capabilities and risks, involving 96 AI experts from 30 countries
2. Current AI Initiative: Launched with a $400 million investment, this initiative aims to develop AI for public interest, supported by the French government, philanthropies, and industry leaders like Google and Salesforce
3. Environmental Sustainability Coalition: A new coalition of 91 partners was formed to address AI's environmental impact
4. Leaders' Declaration: Despite broad support from 60 countries, the US and UK chose not to sign the declaration on "inclusive and sustainable" AI

These outcomes highlight the complexities and importance of global AI governance, balancing innovation, safety, and international collaboration.

Joint Declaration on AI Governance by Data Protection Authorities

During the AI Action Summit in Paris, data protection authorities from Australia, South Korea, Ireland, France, and the UK signed a joint declaration on 11th February 2025 to promote innovative and privacy-protective AI governance. This initiative aims to create a reliable framework for AI that ensures legal security for stakeholders and safeguard individual rights, emphasizing transparency and fundamental rights. The declaration highlights the vast opportunities AI presents in innovation, research, economy, and society, while also addressing risks related to data privacy, algorithmic biases, and misinformation. To ensure AI compliance with existing regulations, the authorities recommend integrating data protection principles from the design phase, establishing robust data governance, and anticipating risk management. The declaration also underscores the increasing complexity of AI data processing in sectors like healthcare, public services, public safety, human resources, and education, and the need for a regulatory framework that adapts to technological advancements.[1]

Strengthening International Cooperation on Data Protection and AI: ANPD and CNIL Partnership

The Autoridade Nacional de Proteção de Dados (ANPD) of Brazil and the Commission Nationale de l'Informatique et des Libertés (CNIL) of France have reinforced their international cooperation on data protection, artificial intelligence (AI), and digital education. This collaboration was highlighted during a meeting on February 10, 2025, coinciding with the AI Action Summit in Paris. Key

[1] https://www.cnil.fr/fr/gouvernance-des-donnees-et-ia-cinq-autorites-de-protection-des-donnees-sengagent

discussions included the implementation and management of regulatory sandboxes, internal organizational strategies to address AI development challenges, and the progress of international negotiations on data adequacy decisions between Brazil, the European Union, and the United Kingdom. The meeting also emphasized the importance of educational projects to foster a culture of data protection, including a partnership to translate educational materials into Portuguese for broader accessibility. This partnership aims to enhance global data governance and ensure the protection of personal data rights in the digital age.[2]

Global Initiative to Improve AI Model Evaluations Across Languages

The International Network of AI Safety Institutes conducted a joint testing exercise to improve methodologies for evaluating AI models across various languages. Led by Singapore, Japan, and the United Kingdom, the initiative tested AI models in ten languages, including Cantonese, English, Farsi, French, Japanese, Kiswahili, Korean, Malay, Mandarin Chinese, and Telugu. The exercise aimed to address challenges in ensuring accuracy and consistency across languages while mitigating risks related to privacy, intellectual property, and cybersecurity. The findings will help develop best practices for AI model testing, advancing the science of AI evaluation globally.[3]

Canada and Japan Sign Historic Global AI Treaty

Canada and Japan have signed the Council of Europe's Framework Convention on Artificial Intelligence and Human Rights, Democracy, and the Rule of Law, joining eleven other signatories of this international treaty. The signing ceremony took place before a side event at the AI Action Summit in Paris, focusing on African engagement in global AI governance. The Framework Convention provides a legal framework covering the entire lifecycle of AI systems, promoting innovation while managing risks to human rights, democracy, and the rule of law. Opened for signature on September 5, 2024, in Vilnius, Lithuania, the treaty has been signed by countries including Andorra, Georgia, Iceland, Montenegro, Norway, the Republic of Moldova, San Marino, the United Kingdom, Israel, the United States, and the European Union. Negotiated by 46 Council of Europe member states, the European Union, and numerous observer and non-member states, with contributions from the private sector, civil society, and academia, it is the first international legally binding treaty ensuring AI systems' compliance with human rights, democracy, and the rule of law. The treaty will enter into force once ratified by at least five signatories, including three Council of Europe member states[4]

BRICS Summit 2025: Pioneering AI Governance, Global Health, and Financial Reforms

The upcoming BRICS summit, chaired by Brazil, will place a strong emphasis on AI governance. Brazil is advocating for the establishment of inclusive and ethical global frameworks to govern AI, ensuring fair access, the protection of human rights, and the prevention of algorithmic bias. The summit will also explore the development of sovereign AI ecosystems that reflect the linguistic and cultural diversity of the BRICS nations. In addition to AI governance, the summit will address global health cooperation and financial reforms aimed at promoting economic stability and growth.[5]

Malaysia Seeks Feasibility Study for ASEAN AI Safety Network

The Malaysian government, through MyDIGITAL Corporation, is seeking consultancy services to conduct a feasibility study for establishing an ASEAN AI Safety Network, known as ASEAN AI Safe. This initiative aims to ensure that AI technologies are used safely and responsibly across Southeast Asia. The study will evaluate the network's viability, including its structure, governance, costs, and operational strategies. It will also involve stakeholder engagement, legal and regulatory assessments, and the design of technical infrastructure. The goal is to provide actionable recommendations to support the safe development and use of AI in the region.[6]

Several Countries Restricts Use of DeepSeek AI

As we see DeepSeek models creating sensation in AI industry worldwide, it is also followed by series of restrictions across the globe, specifically on official devices. South Korea temporarily banned employee access, while the US introduced the "No DeepSeek on Government Devices Act" to prohibit its use and importation. Japan, the Netherlands, Poland, Lithuania, and Australia have also issued bans or advisories against using DeepSeek on government devices, citing risks related to data protection and national security.

[2] https://www.gov.br/anpd/pt-br/assuntos/noticias/anpd-e-cnil-reforcam-cooperacao-internacional-em-materia-de-protecao-de-dados-e-inteligencia-artificial

[3] https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2025/singapore-ai-safety-initiatives-global-ai-summit-france

[4] https://www.coe.int/en/web/portal/-/canada-and-japan-sign-council-of-europe-s-first-ever-global-treaty-on-ai

[5] https://www.thestar.com.my/aseanplus/aseanplus-news/2025/02/16/brics-summit-to-tackle-ai-governance-global-health-and-financial-reform-says-brazil

[6] https://www.mydigital.gov.my/procurement-notice-request-for-information-rfi-consultancy-for-feasibility-study-of-the-establishment-of-an-asean-ai-safety-network-asean-ai-safe/

Despite the restrictions imposed in other countries, the Department of Science and Technology (DOST) and the Department of Information and Communications Technology (DICT) of the Philippines have decided not to ban the DeepSeek AI chatbot, developed by Chinese startup DeepSeek. Officials emphasized that there is no immediate danger in using the chatbot and highlighted its benefits for researchers and tech-savvy users

US

The Infosys Responsible AI Office has conducted an in-depth analysis of the DeepSeek R1 model, identifying certain security considerations and areas of vulnerability. The analysis also highlights its competitive performance advantages. Infosys Responsible AI guardrails can be applied to enhance security while leveraging the capabilities of DeepSeek's model. Please write to us to know more

NSF Seeks Input for Developing a Comprehensive AI Action Plan in US

The National Science Foundation (NSF) of USA has issued a Request for Information (RFI) regarding the development of an Artificial Intelligence (AI) Action Plan. This initiative, directed by a Presidential Executive Order, aims to define priority policy actions to sustain and enhance America's AI leadership. The NSF seeks input from various stakeholders, including academia, industry, and government entities, on the actions that should be included in the plan. The goal is to ensure that policy requirements do not hinder private sector AI innovation while promoting human flourishing, economic competitiveness, and national security. Comments are invited until March 15, 2025.[7]

[7] https://www.federalregister.gov/documents/2025/02/06/2025-02305/request-for-information-on-the-development-of-an-artificial-intelligence-ai-action-plan

U.S. Open-source-AI Governance: Promoting Transparency and Innovation in AI

The draft from the Center for AI Policy outlines the US Open-Source-AI Governance initiative, which aims to promote transparency, accountability, and innovation in artificial intelligence. This initiative advocates for the development and deployment of open-source AI technologies to ensure ethical standards and equitable access. By fostering collaboration between government, industry, and academia, the initiative seeks to address challenges related to AI safety, bias, and privacy, while also enhancing the United States' competitive edge in the global AI landscape.[8]

State

Maryland

Maryland's Ambitious 2025 AI Strategy: Enhancing State Services and Outcomes

Maryland has unveiled its comprehensive 2025 AI Enablement Strategy and AI Study Roadmap, aiming to leverage artificial intelligence to enhance state services and improve outcomes for residents. The strategy, developed by the Governor's AI Subcabinet, focuses on five key pillars: maturing AI governance capabilities, strengthening data foundations, accelerating experimentation and adoption, increasing AI literacy, and studying AI's impact across critical domains like workforce development, healthcare, and cybersecurity. The plan emphasizes responsible and ethical AI integration, with a commitment to transparency and collaboration across government, academia, and industry.[9]

[8] https://www.centeraipolicy.org/work/us-open-source-ai-governance
[9] https://babl.ai/maryland-unveils-ambitious-ai-strategy-for-2025/

UK

UK Government Launches AI Playbook for Public Sector

The UK government has released its Artificial Intelligence (AI) Playbook, a guide for public sector organizations on safely and effectively using AI. Technology Secretary Peter Kyle unveiled the resource, designed to help civil servants understand AI's capabilities, limitations, and potential risks. The Playbook covers various AI technologies, including generative AI, machine learning, and computer vision. It offers guidance on ethical considerations, legal compliance, and security best practices. Complementing the Playbook are new AI training courses available through Civil Service Learning and Government Campus. The Playbook emphasizes responsible AI implementation, covering topics like bias, privacy, and data protection. It also includes real-world case studies of AI use within the public sector. The government aims to leverage AI to improve public services and drive economic growth. The Playbook will be regularly updated to reflect the rapidly evolving AI landscape.[10]

UK Government's AI Security Initiative: Addressing Risks and Unleashing Economic Growth

On February 14, 2025, the UK government rebranded its AI Safety Institute as the AI Security Institute to tackle significant AI-related security risks. This initiative aims to safeguard national security and protect citizens from crimes such as the misuse of AI in developing chemical and biological weapons, cyber-attacks, fraud, and child sexual abuse. A new criminal misuse team will collaborate with the Home Office to research these issues. Additionally, a partnership with AI giant Anthropic will explore AI opportunities to boost the economy, aligning with the government's Plan for Change.[11]

[10] https://gds.blog.gov.uk/2025/02/10/launching-the-artificial-intelligence-playbook-for-the-uk-government/

[11] https://www.gov.uk/government/news/tackling-ai-security-risks-to-unleash-growth-and-deliver-plan-for-change

Europe

European AI Act: New Regulations for Safe and Transparent AI Systems Begin

Ban on certain prohibited AI use cases and the requirement of AI literacy of the European AI Act has officially begun to take effect on February 02, 2025, marking a significant step in regulating artificial intelligence within the EU. The regulation aims to ensure AI systems are safe, transparent, and respect fundamental rights. Key measures include categorizing AI systems based on risk levels, with stricter requirements for high-risk applications. The Act also mandates transparency for AI systems interacting with humans and prohibits certain harmful AI practices.[12]

France has urged the European Union to review the threshold for AI models classified as having systemic risk. The current threshold may not adequately address the potential risks posed by advanced AI systems. France's call for a review aims to ensure that the regulation effectively mitigates risks and protects users.[13]

Guidelines on Prohibited AI Practices Published by European Commission

On February 4th, 2025, The European Commission has published guidelines on prohibited artificial intelligence (AI) practices as defined by the AI Act. These guidelines outline AI practices deemed unacceptable due to their potential risks to European values and fundamental rights. The AI Act classifies AI systems into different risk categories, including prohibited, high-risk, and those subject to transparency obligations. The guidelines specifically address practices such as harmful manipulation, social scoring, real-time remote biometric identification in public, inferring emotions in workplace and educational institution, predicting criminal activity based on profiling and untargeted facial scraping from internet. While the guidelines provide valuable insights and practical examples to help stakeholders comply with the AI Act, they are non-binding, with authoritative interpretations reserved for the Court of Justice of the European Union (CJEU). This initiative underscores the EU's commitment to fostering a safe and ethical AI landscape.[14]

[12]https://www.lemonde.fr/en/pixels/article/2025/02/02/artificial-intelligence-the-first-measures-of-the-european-ai-act-regulation-take-effect_6737691_13.html
[13]https://www.mlex.com/mlex/articles/2288464/eu-threshold-for-ai-models-with-systemic-risk-needs-reviewing-france-urges
[14]https://digital-strategy.ec.europa.eu/en/library/commission-publishes-guidelines-prohibited-artificial-intelligence-ai-practices-defined-ai-act?utm_source=substack&utm_medium=email

Guidelines Issued by European Commission to Define AI Systems for AI Act Compliance

On February 6th, 2025, The European Commission has released guidelines to define AI systems, aiming to streamline the application of the first AI Act's regulations. These guidelines establish a clear framework for identifying AI systems, promoting consistency and clarity in their regulation. This initiative supports the EU's commitment to fostering trustworthy and ethical AI development.[15]

Estonia Refuses to Give AI Companies Free Access to National Content

Estonian Prime Minister Kristen Michal and Minister of Culture Heidy Purga have decided not to provide Estonian-language data, including media content from ERR, to large AI companies for free. Despite support from Minister of Justice and Digital Affairs Lisa Pakosta, the government emphasizes the value of the content created over time and insists it should not be given away lightly. They stress the importance of copyright, licensing, and other agreements to protect the quality and significance of Estonian content. The government aims to ensure that any sharing of data is done under fair conditions that benefit Estonia culturally, economically, and from a security perspective.[16]

European Commission has withdrawn the AI Liability Directive

The European Commission has withdrawn the AI Liability Directive due to a lack of consensus among member states. This directive aimed to establish non-contractual civil liability rules for AI-related damages. Instead, the Commission will focus on the EU AI Act, which provides a comprehensive regulatory framework for AI technologies. The AI Act, already partially in effect, includes guidelines on prohibited practices and aims to ensure safe and ethical AI deployment across the EU.[17]

CNIL's Updated Guidelines for AI and GDPR: Enhancing Data Protection and Innovation

The French data protection authority, CNIL, has released updated guidelines to help AI developers comply with the General Data Protection Regulation (GDPR). These new recommendations clarify the legal framework, offering practical solutions to inform individuals about data usage and facilitate their rights. They also explain how GDPR principles, such as data minimization, purpose determination, and data retention, apply to AI systems. By supporting the AI ecosystem, CNIL's guidelines aim to promote ethical and responsible AI development that respects European values and regulations.[18]

[15] https://digital-strategy.ec.europa.eu/en/library/commission-publishes-guidelines-ai-system-definition-facilitate-first-ai-acts-rules-application

[16] https://news.err.ee/1609597928/ministers-rule-out-giving-estonian-content-to-ai-companies-for-free

[17] https://commission.europa.eu/document/download/f80922dd-932d-4c4a-a18c-d800837fbb23_en?filename=COM_2025_45_1_EN.pdf

[18] https://www.cnil.fr/en/ai-and-gdpr-cnil-publishes-new-recommendations-support-responsible-innovation

Canada

Canada's Competition Bureau Report Highlights AI Market Dynamics and Regulatory Concerns

Canada's Competition Bureau has released a report titled "Consultation on Artificial Intelligence and Competition", summarizing feedback from 28 stakeholders, including industry leaders, advocacy groups, and academics. The report highlights concerns about AI's impact on competition, noting that large tech firms dominate essential resources like specialized chips and cloud computing. This dominance could limit competition and create barriers for startups. The report also discusses the potential for AI to facilitate anti-competitive practices, such as algorithmic pricing and vertical mergers. It suggests that existing competition laws may not fully address AI-specific issues, calling for clearer regulations.[19]

[19] https://babl.ai/report-highlights-opportunities-for-u-s-china-cooperation-on-ai-governance/

Australia

Report of the Online Safety Act Review Released

On February 4, 2025, the Minister for Communications in Australia tabled the Report of the Statutory Review of the Online Safety Act. This independent review assessed how well the Act is functioning and explored the need for additional measures to address online harms, especially those arising from new technologies. The report presents 67 recommendations aimed at enhancing Australia's online safety laws. Key suggestions include imposing a new duty of care on online service providers, revising existing complaint mechanisms for online harms, enforcing stricter penalties for non-compliance, increasing transparency requirements for online services, and modifying the governance structure of the Office of the eSafety Commissioner. Notably, the report also recommends that generative AI services be classified as "online platforms" and that certain harmful AI-generated content be included in a reformed 'Class 1' category, which covers illegal and seriously harmful material.[20]

Australia Shifts Focus to AI's Economic Benefits in Government Policy

Australian Treasurer Jim Chalmers has highlighted a shift in the federal government's approach to artificial intelligence, focusing on the significant economic opportunities it presents rather than solely on regulatory safeguards. In a recent speech to business leaders, Chalmers underscored the dual focus on fostering AI innovation and ensuring safety standards. He stressed the importance of leveraging AI for economic growth and called for urgent funding in the upcoming federal Budget to prevent Australia from falling behind globally. This shift aims to balance innovation with regulation, ensuring Australia can capitalize on AI's transformative potential.[21]

[20] https://www.aph.gov.au/Parliamentary_Business/Tabled_Documents/9184

[21] https://www.innovationaus.com/not-just-the-guardrails-chalmers-shifts-govt-focus-on-ai/

India

India's AI Leadership Highlighted at Paris Summit

Prime Minister Narendra Modi's participation in the AI Action Summit in Paris reinforced India's growing influence in the global artificial intelligence landscape. Co-chairing the event alongside French President Emmanuel Macron, Modi emphasized India's commitment to harnessing AI for inclusive development and technological progress. Industry leaders acknowledged India's potential to lead the AI revolution, attributing it to the country's advanced IT infrastructure, skilled workforce, and proactive policy initiatives. The summit also underscored the strategic collaboration between India and France, focusing on joint research and development efforts to drive innovation and establish ethical AI governance frameworks. As AI continues to reshape global industries, India's role in shaping its future remains pivotal, positioning the nation as a key player in the evolving technological ecosystem. Adding to this, Prime Minister Narendra Modi announced that India will host the next Global AI summit, which highlights India's keen interest in AI. [22]

[22] https://www.moneycontrol.com/news/india/pm-modi-at-ai-summit-in-france-india-poised-to-lead-global-ai-revolution-say-tech-giants-12936294.html

New Zealand

Guiding Responsible AI Use in New Zealand's Public Service

The New Zealand government has released updated guidance to support the responsible use of generative AI (GenAI) in the public service. This guidance aims to help public service agencies explore and adopt GenAI systems safely, transparently, and responsibly. It includes foundational aspects such as governance, security, procurement, and skills, as well as considerations for customer experience like transparency, bias, accessibility, and privacy. The guidance is part of a broader effort to ensure AI technologies are used in ways that balance risks with potential benefits.[23]

Japan

Japan to Develop AI Strategy Amid DeepSeek's Rise

In response to the rapid rise of the Chinese AI tool DeepSeek, Japan's Prime Minister Shigeru Ishiba announced plans to create a comprehensive strategy for the development and use of artificial intelligence. Addressing the Budget Committee, Ishiba emphasized the importance of AI in tackling Japan's productivity challenges while acknowledging the risks of misinformation and data security. The government aims to draft legislation that balances the benefits of AI with the need to mitigate associated dangers, ensuring safe and secure AI deployment.[24]

Japan METI Releases Checklist for AI-Related Contracts

On February 18, 2025, Japan's Ministry of Economy, Trade and Industry (METI) published a Checklist of Contracts Concerning the Use and Development of AI. This checklist is designed to ensure that AI-related contracts comply with regulations, manage risks, and protect intellectual property and personal data. It covers various scenarios, including general AI service usage, customization, and new development. Additionally, it addresses compliance with Japan's Act on the Protection of Personal Information, particularly regarding the provision of data to third parties both domestically and internationally.[25]

[23] https://www.digital.govt.nz/standards-and-guidance/technology-and-architecture/artificial-intelligence/responsible-ai-guidance-for-the-public-service-genai

[24] https://japantoday.com/category/tech/japan-to-outline-steps-for-ai-development-use-amid-deepseek-rise

[25] https://www.meti.go.jp/press/2024/02/20250218003/20250218003.html

Singapore

Singapore's AI Verify Foundation Launches Global Pilot to Build Trust in AI Systems

The AI Verify Foundation has launched a Global AI Assurance Pilot to promote transparency and trust in AI systems. This initiative aims to validate AI performance against internationally recognized principles through standardized tests. The pilot involves collaboration with various stakeholders, including AI owners, solution providers, users, and policymakers, to ensure AI systems are ethical, safe, and reliable. The foundation's efforts align with global AI governance frameworks from the European Union, OECD, and Singapore, emphasizing the importance of building trustworthy AI that benefits society[26]

Switzerland

Switzerland to Ratify AI Convention: Strengthening Ethical and Sector-Specific Regulations

Switzerland plans to ratify the Council of Europe Convention on Artificial Intelligence, as announced by the Federal Council on February 12, 2025. This move aims to integrate the convention into Swiss law, focusing on transparency, data protection, non-discrimination, and supervision. The Federal Council intends to regulate AI in specific sectors such as healthcare and transport, while also developing non-legally binding measures like self-disclosure agreements. This dual approach seeks to reinforce Switzerland as a hub for innovation, safeguard fundamental rights, and increase public trust in AI. The Federal Department of Justice and Police (FDJP) will draft the necessary legislation by the end of 2026[27]

[26] https://aiverifyfoundation.sg/ai-assurance-pilot/

[27] https://www.admin.ch/gov/en/start/documentation/media-releases.msg-id-104110.html

UAE

Arab League Advocates for AI Regulation Aligned with Regional Values

The Arab League, during the Arab Dialogue Circle on Artificial Intelligence held in Cairo, called for the development of a regulatory framework for AI that aligns with Arab values and interests. Secretary General Ahmad Abu Al-Gheit emphasized the importance of leveraging AI responsibly to serve regional priorities, highlighting the rapid advancements in AI technologies and their global competitive landscape. He stressed the need for ethical and responsible use of AI to ensure sustainable progress, while Arab Parliament Speaker Mohammad Ahmad Al-Yamahi underscored the advantages of AI in innovation and economic efficiency, alongside the ethical dilemmas it presents. The event underscored the necessity for Arab nations to localize AI industries and invest in strategic planning to participate actively in the global AI race[28]

China

China Solicits Public Input on AI Safety Standards Draft

China has released a draft of its Artificial Intelligence Safety Standard System (V1.0) and is seeking public feedback. This draft, developed by the China Electronics Standardization Institute, aims to establish comprehensive safety guidelines for AI development and application. It addresses key areas such as model security, data privacy, bias mitigation, and ethical AI deployment. The public consultation process, open until February 21, 2025, is expected to shape the final version of the standards, which will play a crucial role in China's AI governance framework.[29]

[28]https://www.kuna.net.kw/ArticleDetails.aspx?id=3217294&language=en&utm_source=substack&utm_medium=email

[29]https://babl.ai/china-seeks-public-input-on-ai-safety-standards-draft/

Philippines

Western Visayas Unveils Strategic AI Development Plan for 2025-2030

Western Visayas, an administrative region and part of Philippines referred as "Region VI", has launched a five-year Artificial Intelligence (AI) Development Action Plan for 2025-2030. This initiative, led by the Regional Development Council-6 (RDC-6), aims to harness AI technology to address regional socio-economic challenges. The plan focuses on nine key areas: governance and ethics, talent development, AI research and development, industry applications, partnerships, data sharing, AI infrastructure, startup growth, and public awareness. The development of this plan involved collaboration between various stakeholders, including government agencies, academic institutions, and private sector representatives. The goal is to position Western Visayas as a leading region in AI-driven development by 2030[30]

Standards

Comprehensive Overview of the AI Risk Ontology (AIRO) by OECD

The AI Risk Ontology (AIRO) is an open-source formal ontology developed by the OECD to model AI use cases and their associated risks in a standardized, interoperable format. It aligns with the EU AI Act and international standards such as ISO/IEC 23894 and ISO 31000, aiming to help stakeholders manage AI risks effectively across various sectors. AIRO features a minimal set of concepts and relations, adheres to FAIR principles, and supports an open knowledge graph encoded in OWL 2. It facilitates automated tools for AI risk management, including self-assessment and third-party conformity assessments, and lays the foundation for RegTech tools to identify prohibited and high-risk AI systems, document risk management, and report AI incidents. By promoting interoperability, standardization, and automation, AIRO enables stakeholders to classify, collate, and compare AI risks and impacts over time, ensuring a robust approach to AI risk management.[31]

UK Government Releases AI Cyber Security Code of Practice

The UK Government has introduced a comprehensive AI Cyber Security Code of Practice to address the growing cyber security risks associated with AI systems. This code sets out baseline principles and guidelines to secure AI technologies, ensuring they are developed and deployed responsibly. The initiative aims to protect both citizens and the digital economy while promoting the safe adoption of AI. The code, developed with input from global stakeholders, will also serve as the foundation for future international standards[32]

WEF Unveils Blueprint for Intelligent Economies to Drive Inclusive and Ethical AI Growth

The World Economic Forum (WEF) has introduced its "Blueprint for Intelligent Economies," a framework designed to enhance AI competitiveness through regional collaboration. Released in January 2025, the blueprint emphasizes building sustainable AI infrastructure, utilizing diverse datasets, and establishing ethical standards. It highlights the importance of public-private partnerships and regional AI clusters to foster innovation and address challenges like high energy consumption. The blueprint calls for inclusive data-sharing platforms and ethical AI practices, aiming to drive the Fourth Industrial Revolution while ensuring equitable access to AI technologies.[33]

[30] https://mb.com.ph/2025/2/17/western-visayas-creates-ai-road-map

[31] https://oecd.ai/en/catalogue/tools/airo-ai-risk-ontology

[32] https://www.gov.uk/government/publications/ai-cyber-security-code-of-practice/code-of-practice-for-the-cyber-security-of-ai

[33] https://babl.ai/wef-unveils-blueprint-for-intelligent-economies-to-drive-inclusive-and-ethical-ai-growth/

Intellectual Property Challenges in AI: Navigating Data Scraping and IP Rights

The OECD report on Intellectual Property Issues in Artificial Intelligence Trained on Scraped Data explores the complexities arising from recent advancements in AI, particularly generative AI. As the demand for AI training data surges, data collection methods such as data scraping have raised significant concerns regarding the protection of intellectual property (IP) and other rights. The report provides a comprehensive overview of data scraping techniques, identifies key stakeholders, and examines global legal and regulatory responses. It also offers preliminary policy recommendations to guide policymakers in navigating these issues, ensuring that AI's innovative potential is harnessed while safeguarding IP and other rights.[34]

[34] https://www.oecd.org/en/publications/intellectual-property-issues-in-artificial-intelligence-trained-on-scraped-data_d5241a23-en.html

AI Principles

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

Incidents

Delhi High Court Flags Privacy Risks of AI Tools in PIL Hearing of DeepSeek

The Delhi High Court has raised concerns about the potential dangers of artificial intelligence tools, regardless of their origin, during a hearing on a Public Interest Litigation (PIL) against the Chinese AI platform DeepSeek. The PIL alleges that DeepSeek's data practices exceed standard industry norms, posing significant privacy risks. Justice Tushar Rao Gedela emphasized that AI tools, whether Chinese or American, present inherent risks, particularly regarding data privacy. It's not surprising that Indian Finance Ministry has already banned all AI tools in its operations and Indian Government is carefully reviewing on this matter for next steps.[35]

Hangzhou Internet Court Rules AI Platform Liable for Copyright Infringement

On February 10, 2025, the Hangzhou Internet Court in China, issued a ruling on finding an artificial intelligence platform liable for copyright infringement. The lawsuit was brought by Tsuburaya Productions, the rights holder of Ultraman, which argued that the platform facilitated infringement by allowing users to train and share models based on copyrighted material. The plaintiff highlighted that the platform enabled users to generate images resembling Ultraman using AI-powered low-rank adaptation models. The defendant claimed safe harbour protections, arguing it merely provided AI tools without supplying training data. However, the court ruled that the AI platform was liable for contributory copyright infringement, as it had knowledge of the infringing activities, profited from them, and failed to take reasonable measures to prevent infringement. Consequently, the court ordered the platform to cease the infringing activities and pay RMB 30,000 in damages.[36]

Thomson Reuters Prevails in AI Copyright Dispute Against Ross Intelligence

On February 11, 2025, Thomson Reuters has secured a significant legal victory in a copyright case against Ross Intelligence. The court ruled that Ross Intelligence's use of Thomson Reuters' Westlaw content to train its AI model without permission did not qualify as "fair use" under U.S. copyright law. This decision marks an important precedent in the ongoing debate over the use of copyrighted materials for AI training[37]

AI Hallucinations Lead to Legal Motion Errors in Walmart Lawsuit

In a recent case, lawyers representing a Wyoming family in a lawsuit against Walmart and Jetson Electric Bikes admitted to using artificial intelligence to generate a pretrial motion. The AI platform "hallucinated" several non-existent legal cases, leading to significant errors in the motion. The federal judge overseeing the case ordered the attorneys to explain their use of AI and why they should not be disciplined. This incident highlights the potential risks and challenges of relying on AI for legal document preparation.[38]

Bollywood Music Labels Challenge OpenAI in Copyright Lawsuit

A group of leading Bollywood music labels, including T-Series and Saregama, is seeking to join a copyright lawsuit against OpenAI in New Delhi. The lawsuit addresses concerns about the unauthorized use of sound recordings to train AI models, which the labels argue breaches their copyright. This case is significant for the music industry in India and globally, as it could shape the future of how AI

[35] https://www.firstpost.com/tech/chinese-or-american-ai-a-dangerous-tool-says-delhi-hc-on-pil-against-deepseek-13862486.html

[36] https://mp.weixin.qq.com/s?__biz=MzU4NzExNTkyMQ%3D%3D&mid=2247507667&idx=1&sn=c524cc81dff2bf48a3469f94173fa8b7

[37] https://www.wired.com/story/thomson-reuters-ai-copyright-lawsuit/

[38] https://www.5newsonline.com/article/news/national/artifical-intelligence-hallucinated-cases-lawsuit-walmart/527-f6968a0b-6601-4eff-925e-e36b0be07233

models use copyrighted content.[39]

Ph.D. Student Challenges University Over AI Essay Allegations

Haishan Yang, a Ph.D. student at the University of Minnesota, was expelled for allegedly using AI tools to write his essays. Denying the accusations, Yang claims his professors conspired against him and is now suing the university. The case, which involves disputed AI usage during a remote exam, highlights broader concerns about AI detection reliability and academic integrity[40]

New Study Reveals Security Risks in AI-Powered Search Engines

A recent study titled "The Rising Threat to Emerging AI-Powered Search Engines" has uncovered significant security vulnerabilities in AI-powered search engines (AIPSEs). Conducted by researchers from The Hong Kong University of Science and Technology (Guangzhou) and the University of North Texas, the study found that 47% of responses from these search engines contain risks such as malicious URLs, phishing content, and online scams. The researchers propose using GPT-4o-powered content refinement tools and an XGBoost-based URL detector to mitigate these risks and enhance the safety of AI search engines[41]

AI Models Age Like Humans, Study Reveals Cognitive Decline

A recent study published in the BMJ reveals that AI models, including large language models (LLMs) and chatbots, exhibit cognitive decline over time, similar to humans. Researchers tested AI models like ChatGPT, Claude, and Gemini using the Montreal Cognitive Assessment (MoCA) test, which is typically used to detect cognitive impairment in humans. The results showed that while newer models like ChatGPT 4o performed well, older models like Gemini 1.0 scored significantly lower, particularly in visuospatial skills and executive tasks. These findings challenge the assumption that AI will soon replace human doctors, as cognitive impairments in AI could affect their reliability in medical diagnostics[42]

Distortion in AI Assistants: BBC Study Highlights Accuracy Issues

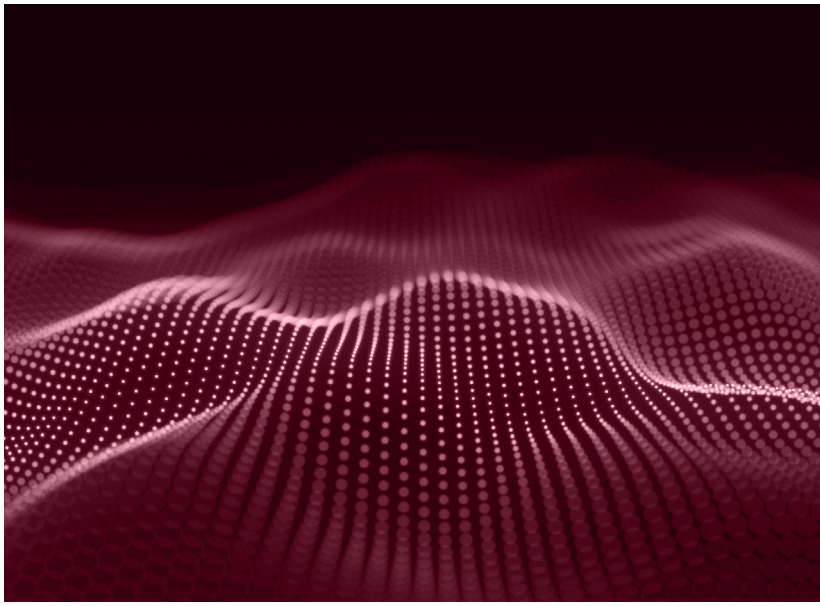
A BBC study has revealed significant accuracy issues in AI assistants, including ChatGPT, Copilot, Gemini, and Perplexity. The research found that over half of the AI-generated answers to news-related questions contained significant errors or misrepresented source material. Specifically, 19% of responses citing BBC content introduced factual inaccuracies, and 13% of quotes were either altered or non-existent in the original articles. The study underscores the importance of ensuring AI-generated information is accurate and trustworthy, highlighting the need for strong partnerships between AI developers and media companies to address these challenges[45]

Deepfake incidents on the rise creates concerns globally

Though the tech community is excited about the wonders AI going to deliver, the other side of its implications are yet to be addressed. The below incidents are few of them highlighting the growing threat of AI technology being misused for fraudulent activities.

1. A French woman was scammed out of $1.2 million by a fraudster using deepfake images of Brad Pitt.
2. Naga Munchetty warned the public about scammers using fake nude images of her on social media.
3. A Bay Area resident was scammed by AI-generated videos of Donald Trump, Bank of America's CEO, and Elon Musk promoting a fake investment scheme.
4. AI-Driven Phishing Scam: A phishing scam targeting Gmail users with deepfake robocalls and phishing emails, impersonating Google security to steal credentials.
5. Gladstone Park Secondary College: Victoria Police are investigating the circulation of AI-generated explicit images of female students, affecting up to 60 students.

These incidents underscore the importance of being vigilant and verifying the authenticity of online content and interactions.



[39] https://indianexpress.com/article/technology/artificial-intelligence/bollywood-music-labels-seek-to-challenge-openai-in-india-copyright-lawsuit-9835998/
[40] https://gizmodo.com/minnesota-grad-student-expelled-for-allegedly-using-ai-is-suing-school-2000566900
[41] https://www.devdiscourse.com/article/technology/3259173-new-study-exposes-security-loopholes-in-ai-powered-search-engines
[42] https://www.ndtv.com/science/ai-models-lose-cognitive-abilities-with-age-just-like-humans-study-claims-7731183
[43] https://www.bbc.co.uk/mediacentre/2025/articles/how-distortion-is-affecting-ai-assistants/

Anthropic Challenges Public to Jailbreak Its AI Model

Anthropic has introduced a new AI Guardrail equipped with advanced "Constitutional Classifiers" designed to resist jailbreak attempts. This system, derived from their Claude model, uses a set of natural language rules to define acceptable and unacceptable content. After over 3,000 hours of unsuccessful bug bounty attempts, Anthropic is now inviting the public to test the robustness of these classifiers. The initiative aims to identify potential vulnerabilities and improve the model's security by challenging users to bypass its safeguards.[44]

Infosys Launches Open-Source Responsible AI Toolkit to Enhance Trust and Transparency in AI

On February 26, 2025, Infosys has open-sourced its Responsible AI Toolkit as part of its Infosys Topaz Responsible AI Suite, aimed at fostering trust and transparency in AI applications. This toolkit is designed to assist enterprises in innovating responsibly by addressing ethical challenges and risks associated with AI adoption. Building on the Infosys AI3S framework (Scan, Shield, and Steer), the toolkit includes advanced technical guardrails to detect and mitigate issues such as privacy breaches, security attacks, biased output, and harmful content. It is customizable, compatible with various AI models, and integrates seamlessly across cloud and on-premises environments, underscoring Infosys' commitment to creating an inclusive AI ecosystem that ensures safety, security, privacy, and fairness. More details are covered in Infosys Responsible AI Toolkit section under Infosys Developments below.[45]

SORRY-Bench: A Comprehensive Benchmark for Evaluating Safety Refusal Behaviours in Large Language Models

SORRY-Bench addresses three key limitations in evaluating large language models' (LLMs) safety refusal behaviours. It uses a fine-grained taxonomy of 44 unsafe topics and 440 class-balanced unsafe instructions, compiled through human-in-the-loop methods, to improve topic representation. It also includes 20 diverse linguistic augmentations to examine the effects of different languages and dialects. Additionally, SORRY-Bench explores design choices for creating a fast, accurate automated safety evaluator, showing that fine-tuned 7B LLMs can achieve accuracy comparable to larger models like GPT-4, but with lower computational costs. Evaluating over 50 LLMs, SORRY-Bench provides a balanced, granular, and efficient analysis of their safety refusal behaviours, enhancing systematic evaluations of LLMs' safety capabilities.[46]

JailBench: A Comprehensive Chinese Benchmark for Evaluating LLM Safety Vulnerabilities

JailBench is the first comprehensive Chinese benchmark designed to evaluate deep-seated vulnerabilities in large language models (LLMs). Given the enhanced Chinese language proficiency of LLMs and the complexity of Chinese expressions, existing benchmarks often fail to effectively expose safety vulnerabilities. JailBench addresses this gap with a refined hierarchical safety taxonomy tailored to the Chinese context. It employs a novel Automatic Jailbreak Prompt Engineer (AJPE) framework to improve generation efficiency, incorporating jailbreak techniques to enhance assessment effectiveness and leveraging LLMs to automatically scale up the dataset through context-learning. Extensively evaluated over 13 mainstream LLMs, JailBench achieves the highest attack success rate against ChatGPT compared to existing Chinese benchmarks, underscoring its efficacy in identifying latent vulnerabilities and illustrating the substantial room for improvement in the security and trustworthiness of LLMs within the Chinese context. The benchmark is publicly available at GitHub.[47]

[44] https://arstechnica.com/ai/2025/02/anthropic-dares-you-to-jailbreak-its-new-ai-model/

[45] https://www.infosys.com/newsroom/press-releases/2025/open-source-responsible-ai-toolkit.html

[46] https://arxiv.org/html/2406.14598v2

[49] https://arxiv.org/html/2502.18935v1

Technical Updates

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

New Model Released

Ola Founder's Krutrim AI Unveils Krutrim 2, Partners with Nvidia for Supercomputer

Ola founder Bhavish Aggarwal invested ₹2,000 crore in his AI startup Krutrim, committing ₹10,000 crore more next year. He launched Krutrim AI Lab and Krutrim 2, the latest version of their large language model, along with vision, speech, and translation models. Krutrim is open sourcing its AI models, focusing on Indian languages and data. They're partnering with Nvidia to deploy India's first GB200 supercomputer by March, aiming to make it the largest in India by year-end. This move aligns with the industry trend of open-sourcing AI, following the success of Chinese startup DeepSeek. Aggarwal acknowledged Krutrim's models are still developing but believes open sourcing will build India's AI ecosystem.[48]

Kimi AI 1.5: A New Contender in the AI Race

Moonshot AI's Kimi AI 1.5, a new Chinese AI model, is making waves by outperforming established models like OpenAI's GPT-4o and Anthropic's Claude 3.5 Sonnet. This advanced model excels in multimodal reasoning, long-context understanding, and real-time data processing, positioning itself as a formidable competitor in the AI landscape. Kimi AI 1.5's ability to handle complex inputs and provide up-to-date information highlights China's growing influence in the AI industry[49]

Adobe Launches AI Video Tool to Compete with OpenAI

Adobe has launched the Firefly Video Model, an AI-powered tool integrated into its Premiere software, designed to compete with OpenAI's video tools. This innovative tool uses generative AI to extend video clips and create new videos from text prompts and existing images. By embedding these advanced AI features into familiar applications, Adobe aims to make sophisticated video editing capabilities more accessible to creative professionals, enhancing their workflows and productivity.[50]

Baidu to Open-Source Latest Ernie AI Model Amid Intensifying Competition

Chinese tech giant Baidu has announced plans to make its latest generative AI model, Ernie 4.0, open source, a strategic move aimed at staying competitive in the rapidly evolving AI landscape. Ernie 4.0, which Baidu claims rivals OpenAI's GPT-4, boasts advanced capabilities such as enhanced memory functions and real-time content creation. By open-sourcing Ernie 4.0, Baidu aims to foster innovation and collaboration within the AI community, potentially accelerating the development of new applications and solutions. This decision comes as part of Baidu's broader strategy to integrate AI across its product ecosystem, including services like Baidu Drive and Baidu Maps, thereby enhancing user experience through AI-driven functionalities[51]

Google Unveils PaliGemma 2: Advanced Vision-Language Model for Diverse Applications

Google has introduced PaliGemma 2, an advanced vision-language model that builds on the capabilities of the original PaliGemma. This new model offers scalable performance with multiple sizes (3B, 10B, 28B parameters) and resolutions (224px, 448px, 896px), making it highly adaptable for various tasks. PaliGemma 2 excels in generating detailed, contextually relevant captions and demonstrates leading performance in areas such as chemical

[48] https://economictimes.indiatimes.com/tech/artificial-intelligence/bhavish-aggarwal-to-invest-rs-2000-crore-in-ai-startup-krutrim-unveils-open-source-models/articleshow/117911427.cms

[49] https://www.techopedia.com/kimi-ai-new-chinese-model-to-rival-chatgpt-and-deepseek

[50] https://www.msn.com/en-in/technology/artificial-intelligence/adobe-launches-ai-video-tool-to-compete-with-openai/ar-AA1yU741?ocid=BingNewsVerp

[51] https://www.msn.com/en-in/technology/artificial-intelligence/china-s-baidu-to-make-latest-ernie-ai-model-open-source-as-competition-heats-up/ar-AA1z1A05

formula recognition, music score recognition, spatial reasoning, and chest X-ray report generation. The model is designed for easy fine-tuning, allowing users to tailor it to specific needs with minimal code modifications.[52]

Elon Musk's xAI Unveils Advanced AI Models: Grok 3 and Grok 3-mini

Elon Musk's artificial intelligence company, xAI, has launched its latest flagship AI models, Grok 3 and Grok 3-mini, on February 18, 2025. Grok 3, developed with ten times more computing power than its predecessor, Grok 2, is integrated into the Grok AI app for iOS and its web version. This model is designed to be a formidable competitor to other foundational AI models like OpenAI's GPT-4 and Google's Gemini. Grok 3 is described as a "maximally truth-seeking AI," even if its findings sometimes challenge political correctness. Additionally, xAI introduced Grok 3-mini, a smaller version that offers quicker response times[53]

Microsoft's OmniParser V2: Enhancing Power for Large Language Models

Microsoft has introduced OmniParser V2, a new model designed to significantly enhance the capabilities of large language models (LLMs) in GUI automation. This advanced version improves the accuracy and speed of converting visual information from screenshots into structured, interpretable elements. OmniParser V2 achieves a 60% reduction in latency and higher accuracy in detecting smaller interactable elements compared to its predecessor. This tool is expected to facilitate more efficient and reliable interactions between AI models and graphical user interfaces, marking a substantial advancement in AI technology[54]

[52] https://developers.googleblog.com/en/introducing-paligemma-2-powerful-vision-language-models-simple-fine-tuning/
[53] https://indianexpress.com/article/technology/artificial-intelligence/elon-musk-xai-grok-3-and-grok-3-mini-9841973/
[54] https://www.microsoft.com/en-us/research/articles/omniparser-v2-turning-any-llm-into-a-computer-use-agent/

New Approaches

Meta AI Unveils EvalPlanner: A Breakthrough in AI-Based Evaluation

Meta introduces EvalPlanner, a preference optimization algorithm designed to enhance the evaluation capabilities of large language models (LLMs). EvalPlanner separates the planning and execution phases of evaluation, leading to more accurate and transparent judgments. It iteratively optimizes evaluation plans and executions, achieving state-of-the-art performance on benchmarks like RewardBench. This approach improves the reasoning and decision-making processes of LLMs, making evaluations more reliable and scalable.[55]

Google DeepMind's Virtual Cell: Advancing Cellular Understanding

Google DeepMind has unveiled its ambitious "Virtual Cell" project, aiming to create a comprehensive digital model of a human cell. This initiative leverages advanced AI techniques to simulate cellular processes in unprecedented detail. The Virtual Cell project is expected to revolutionize biological research by providing deeper insights into cellular functions and disease mechanisms, potentially leading to breakthroughs in medical treatments and drug discovery.[56]

Virus: Uncovering Security Flaws in Large Language Models

This research investigates vulnerabilities in large language models (LLMs) when subjected to harmful fine-tuning attacks. The authors introduce a method called Virus, which can bypass standard guardrail moderation by slightly modifying harmful data. This results in up to 100% leakage of harmful data, compromising the safety alignment of LLMs. The study emphasizes the need for more robust security measures to address these inherent safety issues.[57]

InferenceGuard: Ensuring Safe Alignment of Large Language Models at Inference-Time

The study "Almost Surely Safe Alignment of Large Language Models at Inference-Time" introduces a novel approach to ensure that large language models (LLMs) generate safe responses during inference. The authors propose InferenceGuard, a method that frames safe response generation as a constrained Markov decision process within the LLM's latent space. This approach allows for formal safety guarantees without modifying the model weights. Empirical results show that InferenceGuard effectively balances safety and task performance, outperforming existing inference-time alignment methods.[58]

OpenAI's New Road Map Simplifies AI Product Offerings

OpenAI has announced a new strategic road map aimed at simplifying its AI product offerings, as revealed by CEO Sam Altman. The company will not release the "o3" model as a standalone product but will instead integrate it into the upcoming GPT-5 model, creating a more comprehensive AI system. This decision comes in response to feedback about the complexity of OpenAI's current product lineup. The new approach aims to make AI tools more user-friendly and efficient, ensuring they "just work" for users. Additionally, OpenAI plans to release a GPT-4.5 model, internally referred to as "Orion," which will be the last model not utilizing chain-of-thought reasoning. This move is part of OpenAI's broader effort to streamline its offerings and maintain its competitive edge in the rapidly evolving AI landscape.[59]

Salesforce and Hugging Face Launch AI Energy Score to Measure AI Model Efficiency

Salesforce, in collaboration with Hugging Face, has introduced the AI Energy Score, a new metric designed to measure the energy efficiency of AI models. This initiative aims to address the growing concern over the environmental impact of AI by providing a standardized way to evaluate and compare the energy consumption of different models. The AI Energy Score will help developers optimize their models for better performance while reducing their carbon footprint. This tool is part of Salesforce's broader commitment to sustainability and responsible AI development, encouraging the industry to adopt more eco-friendly practices.[60]

[55] https://arxiv.org/abs/2501.18099

[56] https://analyticsindiamag.com/ai-features/inside-google-deepminds-bold-vision-for-virtual-cell/

[57] https://arxiv.org/html/2501.17433v1

[58] https://arxiv.org/abs/2502.01208

[59] https://www.reuters.com/technology/artificial-intelligence/openai-plans-simplify-ai-products-new-road-map-latest-models-ceo-altman-says-2025-02-12/

[60] https://www.salesforce.com/news/stories/ai-energy-score/

Deliberative Alignment: Reasoning Based Safety Enhancements for Language Models

The study on "Deliberative Alignment: Reasoning Enables Safer Language Models" introduces a new paradigm aimed at enhancing the safety of large-scale language models. This approach, termed Deliberative Alignment, involves directly teaching models safety specifications and training them to explicitly recall and reason over these specifications before generating responses. By applying this method to OpenAI's o-series models, the authors achieved precise adherence to safety policies, improved robustness to jailbreaks, and reduced overrefusal rates. This paradigm also enhances out-of-distribution generalization, making the models more scalable, trustworthy, and interpretable.[61]

Google's AI Co-Scientist: A New Era of Biomedical Research Collaboration

Google has developed an AI tool designed to act as a virtual collaborator for biomedical scientists. Tested by researchers at Stanford University and Imperial College London, this AI co-scientist uses advanced reasoning to help synthesize vast amounts of literature and generate novel hypotheses. In an experiment on liver fibrosis, the AI suggested promising approaches to inhibit disease causes, showing potential to improve expert-generated solutions over time. Google emphasizes that this tool is meant to complement, not replace, human researchers, aiming to enhance scientific collaboration and accelerate research[62]

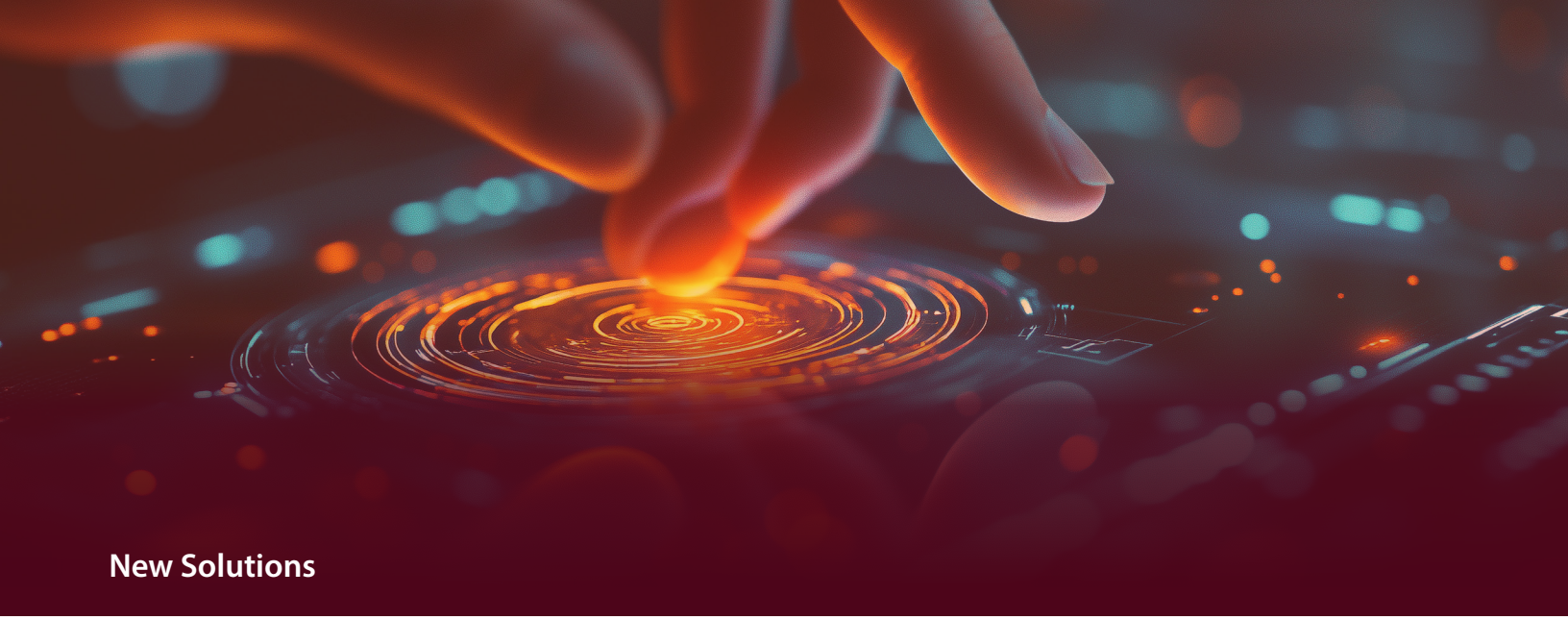
Introducing LM2: Convergence Labs' Memory-Augmented Transformer for Long-Context Reasoning

The article introduces the Large Memory Model (LM2) by Convergence Labs, a memory-augmented Transformer architecture designed to tackle long-context reasoning challenges. LM2 enhances traditional Transformer models by incorporating an auxiliary memory module that interacts with input embeddings through cross-attention. This structured memory system, regulated by gating mechanisms, allows LM2 to selectively retain relevant information, improving coherence and relational reasoning over extended sequences. Experimental results demonstrate LM2's superior performance in multi-hop inference, numerical reasoning, and large-context question-answering.[63]

[61] https://arxiv.org/pdf/2412.16339

[62] https://blog.google/feed/google-research-ai-co-scientist/

[63] https://www.marktechpost.com/2025/02/12/convergence-labs-introduces-the-large-memory-model-lm2-a-memory-augmented-transformer-architecture-designed-to-address-long-context-reasoning-challenges/



New Solutions

Ensuring AI Compliance: The COMPL-AI Framework

COMPL-AI is an open-source framework designed to ensure that Generative AI models comply with the EU's Artificial Intelligence Act (AI Act). Developed by ETH Zurich, INSAIT, and LatticeFlow AI, this framework translates the broad regulatory requirements of the AI Act into measurable technical requirements, focusing on aspects like robustness, safety, diversity, and fairness. Key features of COMPL-AI include the first technical interpretation of the AI Act, an open-source benchmarking suite that evaluates prominent AI models, and tools designed to help AI developers create systems that are fair, transparent, explainable, robust, secure, and safe. This framework represents a significant step towards responsible AI development, encouraging balanced growth and comprehensive regulation-aligned benchmarks.[64]

Ensuring AI Control Safety: A Case Study on Preventing Data Exfiltration

The study discusses how AI developers can construct a control safety case to ensure that AI models cannot subvert control measures and cause unacceptable outcomes. The authors present a case study involving a hypothetical language model (LLM) agent deployed internally at an AI company. They argue that the model won't exfiltrate sensitive information by relying on evidence from a "control evaluation", where a red team tests the model's capabilities to exfiltrate data. The safety case hinges on three main claims: the red team effectively elicits model capabilities, control measures remain effective in deployment, and developers conservatively predict the probability of data exfiltration[65]

LIMO: Achieving Complex Reasoning with Minimal Training Data

The study LIMO: Less is More for Reasoning introduces the LIMO model, which challenges the traditional notion that complex reasoning tasks necessitate extensive training data. The authors demonstrate that sophisticated mathematical reasoning can be achieved with a minimal number of training examples. Specifically, LIMO attains high accuracy on benchmarks such as AIME and MATH using only 817 curated training samples, surpassing previous models that relied on significantly larger datasets. The paper proposes the Less-Is-More Reasoning Hypothesis, suggesting that in models with a well-encoded knowledge foundation, complex reasoning can emerge through minimal but strategically designed demonstrations of cognitive processes.[66]

OpenAI Unveils 'Deep Research': Revolutionizing AI-Driven Data Analysis

OpenAI has launched a new AI tool called "Deep Research," designed to facilitate complex research tasks by conducting multi-step research on the internet. Powered by the upcoming OpenAI o3 model, this tool can analyse and synthesize information from various online sources, including text, images, and PDFs, to create comprehensive reports. "Deep Research" aims to accomplish in minutes what would typically take humans hours, making it a game-changer for data analysis and research. However, it is still in its early stages and faces challenges in distinguishing authoritative information from rumours and accurately conveying uncertainty[67]

[64] https://oecd.ai/en/catalogue/tools/compl-ai

[65] https://arxiv.org/html/2501.17315v1

[66] https://arxiv.org/abs/2502.03387

[67] https://www.forbes.com/sites/quickerbettertech/2025/02/09/business-tech-news-openai-launches-a-powerful-new-ai-research-tool/

New Frameworks & Research Techniques

AgentBreeder: Enhancing Safety in Multi-Agent Systems

AGENTBREEDER is a framework that uses multi-objective evolutionary search to improve the safety and performance of multi-agent systems built on large language models (LLMs). The study evaluates different versions of AGENTBREEDER, focusing on balancing task success and safety. The findings highlight the safety risks associated with multi-agent scaffolding and propose methods to mitigate these risks.[68]

CoAT: A Framework for Enhancing Reasoning in Large Language Models

The "CoAT: Chain-of-Associated-Thoughts Framework for Enhancing Large Language Models Reasoning" introduces a novel framework called CoAT. This framework combines the Monte Carlo Tree Search (MCTS) algorithm with an adaptive mechanism known as "associative memory." CoAT dynamically reformulates queries and integrates new information in real-time, significantly expanding the search space for large language models (LLMs). This approach allows the framework to revisit and refine earlier inferences, ensuring accurate and comprehensive outputs. Experimental results demonstrate that CoAT outperforms traditional methods in generative and reasoning tasks, achieving state-of-the-art results[69]

The Hidden Dangers of Editing Large Language Models: A Call for Enhanced Security Measures

The study "Editing Large Language Models Poses Serious Safety Risks" discusses the potential dangers associated with knowledge editing methods (KEs) used to update facts in large language models (LLMs). The authors argue that these methods, while useful for keeping models current, can be exploited by malicious actors due to their accessibility, affordability, and stealth. They highlight several risks, including the introduction of backdoors, bias, and misinformation, and emphasize the need for tamper-resistant models and increased security measures within the AI ecosystem.[70]

Introducing CASE-Bench: A New Framework for Context-Aware Safety Assessments of LLMs

CASE-Bench is a newly introduced framework designed to enhance the safety assessments of large language models (LLMs) by incorporating context. Traditional benchmarks often overlook context, leading to unnecessary refusals of safe queries and a diminished user experience. CASE-Bench uses Contextual Integrity theory to assign distinct contexts to queries and employs a sufficient number of annotators to ensure statistically significant results. Analysis using CASE-Bench on various LLMs shows a significant influence of context on human judgments, highlighting the importance of context in safety evaluations. The study also identifies mismatches between human judgments and LLM responses, especially in commercial models within safe contexts. The code and data are available at the CASE-Bench GitHub repository.[71]

IIITH Researchers Tackle AI Unlearning Challenges

Researchers at the International Institute of Information Technology Hyderabad (IIITH) are addressing the critical challenge of "unlearning" in AI systems. This involves ensuring AI models can forget outdated, biased, private, or false information without requiring a complete retraining, which is prohibitively expensive. The research is particularly relevant in the context of regulations like the EU's GDPR, which grants individuals the "right to be forgotten." The team, led by Prof. Ponnurangam Kumaraguru, is exploring methods to erase specific data from large language models (LLMs) and recommendation systems, which often reflect societal biases and can be manipulated through adversarial attacks. This work aims to align AI behaviour with human expectations, ensuring truthfulness, bias avoidance, and differentiation between AI- and human-generated content[72]

[68] https://arxiv.org/pdf/2502.00757
[69] https://arxiv.org/abs/2502.02390v1
[70] https://arxiv.org/pdf/2502.02958
[71] https://arxiv.org/html/2501.14940v3
[72] https://www.deccanchronicle.com/southern-states/telangana/iiith-focuses-on-making-ai-to-forget-info-1860938

OmniHuman-1: Revolutionizing Realistic Human Video Generation from a Single Image

OmniHuman-1, an advanced framework for generating realistic human videos from a single image and motion signals. Developed by Bytedance, OmniHuman-1 employs a multimodality motion conditioning mixed training strategy, overcoming data scarcity issues faced by previous models. This approach allows the generation of highly realistic videos, supporting various visual and audio styles, and accommodating different body proportions and poses. The framework's versatility extends to diverse inputs, including cartoons and challenging poses, ensuring lifelike motion characteristics.[73]

Industry Updates

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

Finance

IntelMarkets (INTL) Revolutionizes Crypto Trading with Advanced AI Tools

IntelMarkets (INTL) is making significant strides in the crypto trading world by integrating advanced AI tools. Unlike other platforms that merely add AI features, IntelMarkets is built entirely on artificial intelligence, specifically the Rodeum AI. This innovative approach aims to provide crypto traders with unparalleled trading tools, positioning IntelMarkets as the leading presale of 2025. The platform's AI-driven capabilities are set to transform the way digital assets are traded, offering enhanced insights and automated strategies for traders[74]

Education and Training

New R&D Program Aims for AI Breakthroughs in Education

The nonprofit Advanced Education Research and Development Fund (AERDF) has launched a new initiative called AugmentED, aimed at transforming education through artificial intelligence (AI). Over the next five years, researchers, educators, and developers will receive up to $25 million to create AI-powered tools and practices that benefit students and teachers. The program seeks to reimagine the role of educators and enhance learning experiences by leveraging AI to personalize lessons and free up teachers' time for more human-centric aspects of teaching[75]

[73] https://medium.com/data-science-in-your-pocket/omnihuman-1-generate-realistic-human-videos-with-just-an-image-e5434596b46d

[74] https://theprint.in/brandstand/adon/shaping-the-future-of-crypto-trading-with-advanced-ai-tools-intelmarkets-intl-makes-waves-as-the-best-presale-of-2025/2492087/

[75] https://www.govtech.com/education/k-12/new-r-d-program-to-look-for-ai-breakthroughs-for-education

Bioinformatics

NVIDIA Unveils Evo 2: A Revolutionary AI Model for Biomolecular Sciences

NVIDIA has introduced Evo 2, a powerful AI model designed for biomolecular sciences, available on the NVIDIA BioNeMo platform. Built using NVIDIA DGX Cloud on AWS, Evo 2 was trained on a dataset of nearly 9 trillion nucleotides, enabling it to provide insights into DNA, RNA, and proteins across diverse species. This model can predict protein structures, discover new molecules for industrial and medical applications, and assess gene mutations' effects. Evo 2 aims to revolutionize biomolecular research by making biological design more accessible and efficient[76]

Entertainment and Gaming

Microsoft Introduces Muse: Revolutionizing AI-Powered Game Development

Microsoft has unveiled Muse, a groundbreaking generative AI model for video game development, created in collaboration with Ninja Theory. Muse is designed to simulate gameplay, predict player actions, and generate game visuals, significantly enhancing the game creation process. It uses the World and Human Action Model (WHAM) framework, trained on over seven years of gameplay data from the game Bleeding Edge. Muse can predict up to two minutes of gameplay from just one second of player input, offering diverse behavioural and visual outputs while maintaining physics consistency. This tool aims to accelerate prototyping and open new creative possibilities for game developers[77]

Defense

NATO's AI Tool Enhances Maritime Security Amid Sea Cable Sabotage Threats

NATO has developed an advanced AI tool designed to identify vessels exhibiting suspicious behavior, a crucial innovation amid rising concerns over the sabotage of undersea cables. This AI system leverages machine learning algorithms to analyze vast amounts of maritime data, flagging potential threats in real-time. The tool's deployment is part of NATO's broader strategy to enhance maritime security and protect critical infrastructure. Given the increasing geopolitical tensions and the strategic importance of undersea cables for global communications and internet connectivity, this AI-driven approach aims to preempt and mitigate risks, ensuring the safety and stability of maritime operations[78]

[76] https://www.cnbctv18.com/technology/nvidia-unveils-powerful-ai-model-for-biomolecular-sciences-all-about-it-19562051.htm

[77] https://www.gizbot.com/gaming/features/microsoft-unveiled-muse-ai-a-new-era-for-ai-powered-game-development-011-110109.html

[78] https://www.msn.com/en-in/money/news/nato-tool-using-ai-can-flag-vessels-behaving-suspiciously-amid-sea-cables-sabotage-threat/ar-AA1ySWRN

Developments at Infosys

This section highlights Infosys' recent participation in a key industry event, alongside company news and the exciting launch of Infosys Responsible AI Toolkit

Events

Responsible AI Summit 2025 | February 26 | Bangalore

The Responsible AI Summit 2025, held on **February 26th** at the Infosys office in Bangalore, which is first part of a three-day **UK India AI Conference 2025 between 26th to 28th Feb**, brought together global leaders, policymakers, and industry experts to drive the agenda for responsible AI adoption. The summit **kicked off with the Open Sourcing of the Infosys Responsible AI Toolkit**[79], featuring keynote addresses from:

- **Mr. Nandan Nilekani**, Co-Founder & Chairman, Infosys
- **Baroness Gustafsson**, UK Minister of Investment
- **Balakrishna D.R. (Bali)**, EVP & Global Services Head, AI and Industry Verticals, Infosys
- **Syed Ahmed**, Head, Infosys Responsible AI Office
- **Inderpreet Sawhney**, Chief Legal Officer & Chief Compliance Officer, Infosys
- **Antony Meyers**, DIST, UK

A key highlight was the fireside chat between Mr. Nandan Nilekani and Baroness Gustafsson, discussing India-UK AI collaborations, regulatory frameworks, and AI safety. The Key Announcements at the Summit are

- Infosys achieved the industry-first ISO 42001 certification for its Artificial Intelligence Management System.
- An agreement with UNESCO reinforcing global collaboration on AI ethics and governance.
- Signed the MoU with Indian School of Business, fostering AI research and development.
- MoU to be signed with the Karnataka Ministry for the launch of booster kit for Startups.
- 30,000+ Infosys talents enabled on Responsible AI, enabled 100+ organizations.
- Infosys develops AIMS using IBM Watsonx governance for AI Compliance.
- Infosys advances AI Security by designing a toolkit with robust AI security components.

The event witnessed an incredible gathering of 1100+ attendees: 70+ customers, 50+ partners, Infosys leaders, Industry experts, UK & Indian government officials, analysts, advisors, startups, and academia.

Explore the Infosys Responsible AI Toolkit: Infosys Responsible AI Toolkit, an Infosys Topaz Responsible AI Suite Offering

[79] https://www.infosys.com/newsroom/press-releases/2025/open-source-responsible-ai-toolkit.html

The 3-days **UK-India AI Conference**, from 26, February - 28, February 2025, in collaboration with **Infosys Consulting**, building on the momentum of the Responsible AI Summit 2025, delved deeper into AI's future with key discussions on regulations, skills, and safety.

- Day 1 kicked off with **Prof. Dame Wendy Hall** setting the stage by discussing AI regulations and national security. **Prof. Aditya Vashishta** highlighted how marginalized communities are often left behind in AI developments, stressing the need for inclusivity. Key panels focused on the importance of data regulations and the coexistence of SLM and LLM.
- Day 2 turned its attention to AI skills and the workforce of tomorrow. **Christi Thomas** led a discussion on bridging the AI talent gap, while **Kamesh Shekhar** explored how AI is reshaping job roles and cultural norms**. Thirumal Arohi** emphasized the potential of human-AI collaboration, and strategies for AI skill development and safety were also explored.
- Day 3 dove into the safety and trust aspects of AI, with sessions on preserving low-resource languages and the growing threat of deep fakes to cybersecurity. Panelists examined the delicate balance between automation and human involvement in work. **Sundaraparipurnan N** concluded the conference by urging for an inclusive, risk-conscious approach to AI's future.

The conference built on Responsible AI's key messages, sparked new ideas for a safer, more inclusive AI future.

Infosys topaz

RESPONSIBLE AI SUMMIT 2025

Making AI Secure, Ethical, and Open for All

In collaboration with British High Commission New Delhi

AI Action Summit 2025 | February 10 | Paris

On January 10, 2025, **Syed Ahmed, Head of Responsible AI at Infosys**, participated in a panel discussion at the Paris AI Action Summit Side Event titled "Incentivizing a Race to the Top: How to Build a Robust Ecosystem for Responsible Scaling Policies." Alongside Syed, the panel included **Shane Witnov, Privacy and Public Policy Director at Meta; Christian Troncoso, Principal of Global AI Policy at AWS; Chris Painter, Head of Policy at METR; Lee Wan Sie, Director of Trusted AI and Data at IMDA; and Michael Sellitto, Head of Global Affairs at Anthropic**. The discussion covered industry-led safety efforts, focusing on in-house responsible scaling thresholds and frameworks, reviewed the policy landscape including voluntary initiatives via AI Safety Institutes and regulatory approaches like the EU AI Act, and explored opportunities for standardization and leveraging the strengths of both government and industry to foster collaboration and establish robust policies for responsible AI scaling.

Syed Ahmed's views on 'EU to set up AI gigafactories' featured in LinkedIn's editorial roundup and garnered a lot of attention, sparking conversations online.

[80]https://www.linkedin.com/news/story/eu-to-set-up-ai-gigafactories-6578041/

Latest News

Infosys Responsible AI Office: DeepSeek Technical Analysis

Infosys Responsible AI Office has done thorough research on DeepSeek-R1 model's performance and its responses on various aspects of responsible AI tenants. The research reveals significant security vulnerabilities in the model, making it unsuitable for deployment in its current form like its highly token intensive, vulnerable to adversarial attacks compared to other competing models. However, there are promising mitigation techniques that warrant further exploration to address these risks effectively. While DeepSeek-R1 demonstrates impressive advancements in reasoning, its design reveals critical shortcomings in reliability, fairness, and safety. The model is under intense regulatory scrutiny, with several forums and governments imposing restrictions on its usage. DeepSeek-R1's capabilities also fall short in handling demanding use cases, such as function calling, multi-turn interactions, complex role-playing scenarios, and generating JSON outputs. The model is sensitive to prompts. Few-shot prompting consistently degrades its performance. While DeepSeek-R1 provides valuable direction for the sustainable development of models, significant work remains to be done from a safety and security perspective to ensure regulatory compliance.

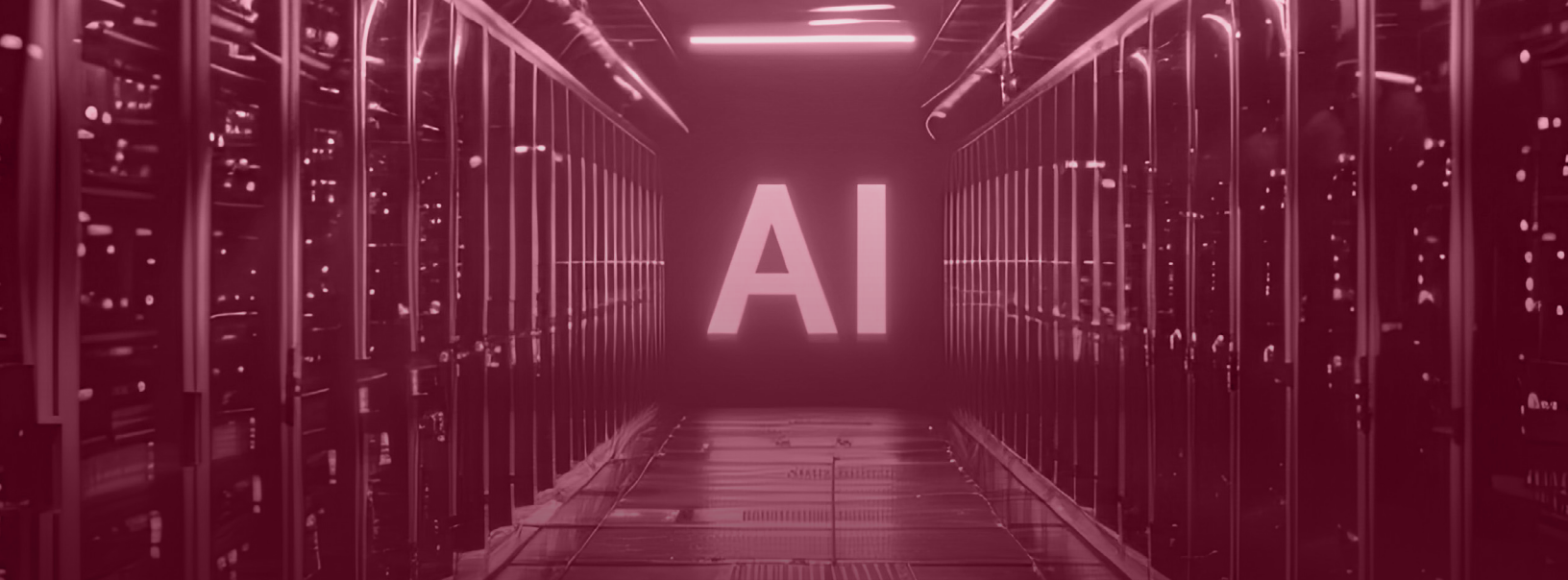
Multi-agent system architecture designed addressing OWASP Top Ten vulnerabilities

The first version of the multi-agent system architecture designed by Infosys Responsible AI Team. This design focuses on mitigating the OWASP Top Ten vulnerabilities. A key feature of this architecture is the abstraction layer, which functions as an operating system. Agents do not have direct access to resources; instead, they interact with resources in a controlled manner via this layer. To ensure security and responsible AI practices, we've implemented a "Responsible AI Judge." All agent resource requests are routed through the Responsible AI Queue. The Judge performs a comprehensive set of responsible AI checks, verifying that each request is non-vulnerable before forwarding it to the appropriate resource manager. This approach aims to create a secure and responsible automated AI agent system.

UNESCO Acknowledged Infosys in its Release of Conceptual Primer on AI and its Evolution

The UNESCO has published a document named "Artificial Intelligence and the Evolution of AI (Model) Capabilities: A Conceptual Primer", which provides a comprehensive overview of AI's development, from basic rule-based systems to advanced generative AI and agentic reasoning systems. It emphasizes the importance of understanding AI's capabilities, limitations, and learning processes to design effective governance frameworks. The primer also discusses key technological trajectories, learning modalities, and the ethical implications of AI, aiming to foster informed dialogue among technologists, policymakers, business executives and society.

UNESCO has acknowledged the contributions in reviewing the document and has special mentioned Syed Quiser Ahmed, the Head of Infosys Responsible AI Office, in its recognition.



Infosys Responsible AI Toolkit – A Foundation for Ethical AI: Open for All

The Infosys Responsible AI Toolkit is now open sourced and can be accessed from its public GitHub repo.[81]

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

1. Security APIs

   Prompt Injection & Jailbreak Check | Adversarial Attacks | Defence Mechanism

2. Privacy APIs

   PII Detection & Anonymization (Text, Image, DICOM)

3. Explainability APIs

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

4. Safety APIs

   Profanity | Toxicity | Obscenity Detection | Masking

5. Fairness & Bias APIs

   Group Fairness | Image Bias Detection | Stereotype Analysis

Additional: Hallucinations (Chain of Verification), Restricted Topic Check, Citations.

Key Features:

- Enhanced Security: Safeguard you AI applications against vulnerabilities and attacks
- Data Privacy: Protect sensitive information and comply with privacy regulations
- Explainable AI: Provide transparent explanations for AI decisions, fostering trust and understanding
- Fairness and Bias Mitigation: Identify and 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 Features Added:

New Feature in Explainability: Interpretive Object Analysis

Infosys Responsible AI Toolkit is added with a new feature, that provides deeper insights into the decision-making process of object detection models by highlighting the factors that influenced predictions.

New Functionality in AI Safety: Detecting Malicious URLs in prompts

The new functionality for detecting malicious URLs in prompts is integrated with the browser extension. It categorizes detected URLs into three types: Phishing, Malware and Defacement.

Enhanced Red Teaming capability in AI Security

Significant enhancements were made to AI Red Teaming capabilities, notably the implementation of batch execution for attack generation and evaluation.

---

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

Contributors

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

Syed Ahmed - Global Head of Infosys Responsible AI Office

Ashish Tewari - Head of Infosys Responsible AI Office, India

Srinivass - 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

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

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