MARKET SCAN REPORT

APRIL 2025

Infosys topaz

BY INFOSYS TOPAZ RESPONSIBLE AI OFFICE

IN FOCUS

UK'S APPROACH TO AI REGULATION

By Dr. Cosmina Dorobantu

Infosys®
Navigate your next

Message from Global Head, Infosys Responsible AI Office

The pace of AI innovation shows no signs of slowing - and with it comes the growing responsibility to ensure these technologies are trustworthy, inclusive, and aligned with human values. Each edition of this Market Scan reflects our ongoing effort to track the evolving landscape of Responsible AI and provide insights that help organizations act with foresight and integrity.

From governance frameworks to emerging research and enterprise adoption, this report offers a curated view of how Responsible AI is taking shape globally. It is our belief that building AI responsibly is not just a compliance exercise - it is foundational to creating long-term, sustainable impact. We remain committed to sharing global perspectives, expert voices, and actionable intelligence each month - helping you stay informed, prepared, and ahead of the curve as you shape your AI strategy for the future.

In this month's edition, I am pleased to introduce a new section titled "In Focus" - a platform for global experts to share perspectives on critical developments in Responsible AI. We thank Dr. Cosmina Dorobantu for contributing her insights on the UK's approach to AI regulation.

I would also like to take a moment to inform our readers that Ashish Tewari has taken over as the full-time editor of the Market Scan Report. As Regional Head of the Responsible AI Office, India, he will lead this monthly publication going forward, curating key developments and perspectives shaping the global Responsible AI landscape.

Syed Ahmed
Global Head
Infosys Responsible AI Office

Dear Readers,

As shared by Syed Ahmed in his message, the evolution of artificial intelligence continues to reshape industries and societies, bringing both opportunity and responsibility. Each month, the Market Scan Report brings you a curated view of the most relevant developments in Responsible AI from around the world - and this edition is no different.

California's new AI laws are setting a precedent in transparency and safety, with a focus on datasets and deepfake detection, while the NO FAKES Act addresses the unauthorized use of generative AI content - a crucial step in protecting creators. Canada's first federal AI strategy highlights the importance of digital governance and public trust, and India's finalized Digital Personal Data Protection Act marks a milestone for AI ethics and cybersecurity in the Global South.

Amid the surge of agentic AI systems, the risks are evolving too. From deepfake scams to unauthorized data access, the need for robust security and fairness mechanisms is more pressing than ever. Research efforts are increasingly focused on solutions like Encrypted Prompts for permission control, and inclusive benchmarks like M-ALERT to ensure multilingual and equitable safety standards.

On the innovation front, Runway's Gen-4 model and Stability AI's virtual camera are redefining creative and immersive experiences, while protocols like Google's Agent2Agent and Zhipu's AutoGLM Rumination are advancing automation and resilience in AI systems.

This month, also marks the launch of our **new section - In Focus**, which features thought leaders who are shaping the global Responsible AI discourse. I'm delighted to welcome **Dr. Cosmina Dorobantu** as our first contributor, offering a timely perspective on the UK's approach to AI regulation.

As the new full-time editor of this publication, I look forward to bringing you timely, relevant, and actionable insights each month. We hope this report supports your journey in scaling AI responsibly.

Stay informed. Stay ahead.

Warm regards,

Ashish Tewari
Head- Infosys Responsible AI Office, India

Table of Contents

AI Regulations, Governance & Standards

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

AI Regulations and Governance across globe

Holy See Urges Global AI Regulation and Nuclear Disarmament at UN

On 8 April 2025, Archbishop Gabriele Caccia, the Holy See's permanent observer to the United Nations, addressed the UN Disarmament Commission in New York, USA, calling for urgent global action on nuclear disarmament and the regulation of artificial intelligence (AI). In his speech, Archbishop Caccia highlighted fear as a dominant force in global affairs and emphasized the existential threats posed by the weaponization of emerging technologies like AI. He advocated for the establishment of international frameworks, deeming them "imperative" to ensure these technologies benefit humanity and promote peace.[1]

Malaysia-China AI Innovation and Cooperation Centre Established as Flagship China–ASEAN AI Lab

On 11 April 2025, Malaysia's MY E.G. Services Berhad (MyEG) and China's Guangxi Beitou IT Innovation Technology Investment Group partnered to establish the Malaysia-China AI Innovation and Cooperation Centre, designated as the flagship China–ASEAN AI Lab. Supported by both governments, this initiative aims to integrate blockchain, generative AI, and robotics technologies for cross-border applications tailored to local cultures. Located at Zetrix Tower in Petaling Jaya, Malaysia, the lab will involve leading companies like Huawei, DJI, and Alibaba, with its first service enabling mutual recognition of national digital IDs between Malaysia and China. The Guangxi government has allocated $1.38 billion to support such projects under the China–ASEAN program.[2]

[1] https://catholicweekly.com.au/archbishop-caccia-speaks-on-ai-and-nuclear-disarmament/?utm_source=substack&utm_medium=email

[2] https://www.digitalnewsasia.com/business/myeg-and-beitou-it-innovation-establish-flagship-china-asean-ai-lab?utm_source=substack&utm_medium=email

US

BIS Expands Export Controls to Cover Additional 80 Entities, Notably to Further Restrict China's AI and Advanced Computing Capabilities

On March 26, 2025, the U.S. Department of Commerce's Bureau of Industry and Security (BIS) amended the Export Administration Regulations (EAR), expanding the Entity List to include 80 additional entities from China, the United Arab Emirates (UAE), South Africa, Iran, Taiwan, and other countries. These entities will be subject to export restrictions and additional licensing requirements on certain sensitive U.S.-made technologies. The update, issued under the Export Control Reform Act, targets entities involved in advanced computing, AI, quantum technologies, and military-related research across multiple jurisdictions, which may engage in activities deemed contrary to U.S. national security and foreign policy interests. The stated objectives of this update include restricting the Chinese Communist Party's ability to acquire and develop high-performance and exascale computing capabilities, as well as quantum technologies for military applications; impeding China's development of its hypersonic weapons program; preventing entities associated with the Test Flying Academy of South Africa (TFASA) from using U.S. items to train Chinese military forces; disrupting Iran's procurement of unmanned aerial vehicles (UAVs) and related defense items; and impairing the development of unsafeguarded nuclear activities and ballistic missile programs.[3]

CREATE AI Act Re-Introduced

House of Representatives members Jay Obernolte (R-CA) and Don Beyer (D-VA) have introduced the draft text of the "Creating Resources for Every American To Experiment with Artificial Intelligence Act of 2025" (CREATE AI Act) (HR 2385), aimed at expanding access to AI research tools by establishing the National Artificial Intelligence Research Resource (NAIRR). This bill was reintroduced into the House of Representatives following a first attempt in 2023. The NAIRR is a pilot program currently managed by the National Science Foundation, launched in January 2024, focusing on private sector collaborations to democratize computing and data training resources for AI researchers. Initially mandated under former President Biden's 2023 executive order on AI, this bill seeks to codify the NAIRR as a permanent institution within NSF to ensure its funding assistance continues despite the repeal of Biden's AI order under the Trump administration.[4]

[3] https://www.bis.gov/press-release/commerce-further-restricts-chinas-artificial-intelligence-advanced-computing-capabilities

[4] https://beyer.house.gov/news/documentsingle.aspx?DocumentID=6456&utm_source=substack&utm_medium=email

House Judiciary Sub-committee Hearing on Innovation and Competition in AI

The House Judiciary Subcommittee on the Administrative State, Regulatory Reform, and Antitrust held a hearing titled "Artificial Intelligence: Examining Trends in Innovation and Competition" to examine critical issues affecting the competitive landscape within the AI sector in the United States. Both Republicans and Democrats recognized the benefits of AI and acknowledged its rapid advancement, but they disagreed on the level of regulation needed. Scott Fitzgerald (R-Wis) advocated for a free enterprise, open competition, and a 'light touch' regulatory approach, while Jerry Nadler (D-NY) suggested strong, independent oversight by the Federal Trade Commission to protect consumers against unfair business practices.[5]

OMB Memorandum on Federal Use of AI

The Office of Management and Budget Administration (OMB) of the United States has issued a new memorandum, M-25-21, replacing the previous OMB M-24-10, as required by President Trump's Executive Order 14179. This memorandum provides guidance on the federal use of artificial intelligence (AI) based on three key principles: 1) agencies must remove barriers to innovation and ensure the best value for taxpayers by eliminating bureaucratic requirements, developing public AI strategies, and maximizing existing investments; 2) agencies must empower AI leaders to accelerate responsible AI adoption by setting workforce expectations for appropriate AI use and identifying Chief AI Officers; and 3) agencies must ensure their use of AI benefits the American people by implementing minimum risk management practices for high-impact AI applications.[6]

OMB Memorandum on Driving Efficient Acquisition of AI in Government

The Office of Management and Budget Administration (OMB) has released another memorandum M-25-22, as required under President Trump's Executive Order 14179. The new memorandum provides guidance on federal acquisition of AI systems and sets out the following agency-level requirements: 1) update agency policies; 2) maximize the use of "American-made AI"; 3) protect privacy; 4) protect intellectual property rights and the use of government data; 5) spotlight AI acquisition authorities, approaches, and vehicles; 6) contribute to a shared repository of best practices; and 7) determine necessary disclosures of AI use in the fulfilment of a government contract.[7]

California's Bold AI Legislation Sets a Global Benchmark

California's new AI laws are set to reshape the future of work and establish a global standard. Starting in 2025, these laws address deepfakes, privacy, and transparency. Key measures include requiring generative AI developers to disclose datasets (Assembly Bill 2013), mandating tools to detect AI-generated content (Senate Bill 942), and protecting election integrity (Assembly Bill 2655). In healthcare, patients must be informed when AI drafts their medical messages (Assembly Bill 3030). The employment sector will see limits on AI's role in hiring and firing to preserve human judgment. These laws underscore California's commitment to transparency, accountability, and privacy in the AI era.[8]

New Jersey Enacts Law to Combat AI-Generated Deepfakes

New Jersey, United States, Governor Phil Murphy has signed a new law imposing civil and criminal penalties on the creation and distribution of AI-generated deepfakes. This legislation was inspired by an incident at Westfield High School, where a student used AI to create explicit images of classmates. The law aims to protect individuals from the harmful effects of deepfakes, which can be used for malicious purposes such as exploitation and deception. Penalties for violating this law include up to five years in prison and a $30,000 fine. Governor Murphy emphasized the importance of this law in safeguarding young people and maintaining public trust, addressing concerns about the misuse of AI in undermining public trust and interfering with democratic processes.[9]

Ethical Guidelines for Lawyers Using Generative AI: Key Points from ABA Model Rules

The ethical considerations for lawyers using generative AI are guided by the American Bar Association's (ABA) Model Rules of Professional Conduct. Lawyers must supervise AI-generated work as they would with human assistants, ensuring compliance with ethical standards. Rule 1.1 requires lawyers to stay updated on technological advancements, including AI, to provide competent representation. Rule 1.6 emphasizes the importance of maintaining client confidentiality when using AI tools, while Rule 1.4 mandates effective communication with clients about the use of AI in their cases. Additionally, Rule 1.5 requires that fees for AI-assisted work be reasonable, reflecting the time spent on both inputting information and reviewing AI-generated outputs. Ethical AI use in law hinges on both developers creating responsible AI and lawyers understanding and controlling its use.[10]

NO FAKES Act Reintroduced in US Senate to Protect Creators from AI-Generated Digital Replicas

On April 9, 2025, US Senators Chris Coons (D-DE), Marsha Blackburn (R-TN), Amy Klobuchar (D-MN), and Thom Tillis (R-NC) reintroduced the NO FAKES Act, aimed at protecting creators' voices and likenesses from unauthorized AI-generated digital replicas. Supported by House members María Elvira Salazar (R-FL-27), Madeleine Dean (D-PA-4), Nathaniel Moran (R-TX-1),

[5] https://judiciary.house.gov/committee-activity/hearings/artificial-intelligence-examining-trends-innovation-and-competition-0?utm_source=substack&utm_medium=email

[6] https://www.whitehouse.gov/wp-content/uploads/2025/02/M-25-21-Accelerating-Federal-Use-of-AI-through-Innovation-Governance-and-Public-Trust.pdf?utm_source=substack&utm_medium=email

[7] https://www.whitehouse.gov/wp-content/uploads/2025/02/M-25-22-Driving-Efficient-Acquisition-of-Artificial-Intelligence-in-Government.pdf?utm_source=substack&utm_medium=email

[8] https://www.thehrdigest.com/californias-new-ai-laws-rewrite-the-future-of-work/

[9] https://apnews.com/article/new-jersey-deepfake-videos-criminal-civil-penalties-276ca23b00b10a7ee7e7303ead8b4260

[10] https://legal.thomsonreuters.com/blog/generative-ai-and-aba-ethics-rules/?utm_campaign=Artificial%2BIntelligence%2BWeekly&utm_medium=web&utm_source=Artificial_Intelligence_Weekly_432

and Becca Balint (D-VT-At Large), the bill would establish the first federal right of publicity in the U.S. Entertainment organizations like SAG-AFTRA, RIAA, MPA, YouTube, and OpenAI back the bill, emphasizing the importance of safeguarding human creativity in an era increasingly dominated by AI technologies.[11]

TAME Extreme Weather and Wildfires Act Introduced to Enable Use of AI for Weather Predictions

On April 9, 2025, US Senators Brian Schatz (D-Hawai'i) and Tim Sheehy (R-Mont.) introduced the TAME Extreme Weather and Wildfires Act, a bipartisan legislation aimed at enhancing weather predictions and responses using artificial intelligence (AI). The bill requires the National Oceanic and Atmospheric Administration (NOAA) to take several actions to support AI forecasting and weather models, including developing a dataset to train AI forecasting models, supporting the deployment of AI weather models into forecasts, and partnering with the private sector and academia on AI weather and wildfire forecasting. Initially introduced in the House and Senate last Congress, the bill (S. 1378; H.R. 2770) did not advance out of either chamber but did progress out of the Senate Commerce, Science, and Transportation Committee in July 2024 as an amendment to the Fire Ready Nation Act. The new legislation remains largely unchanged but adds a focus on wildfires.[12]

Clean Cloud Act Introduced in the United States, Covering AI and Crypto Data Centres

On April 11, 2025, US Senators Sheldon Whitehouse (D-R.I.) and John Fetterman (D-Pa.) introduced the draft text of the Clean Cloud Act. This legislation aims to mitigate the environmental impacts of the growing electricity consumption by AI and cryptocurrency-mining data centres. The proposed bill mandates that these data centres in the United States pay a fee if their greenhouse gas emissions exceed a newly established federal standard. This initiative represents a significant step towards addressing the environmental footprint of data-intensive technologies.[13]

US Senators Question Google and Microsoft's AI Deals

On 8 April 2025, Democratic U.S. senators Elizabeth Warren (D-Massachusetts) and Ron Wyden (D-Oregon), members of the Senate banking and finance committees respectively, sent letters to Microsoft and Google. They requested information about the companies' cloud computing partnerships with OpenAI and Microsoft, expressing concerns that these arrangements could stifle competition in the industry. The senators emphasized the need for transparency and scrutiny to ensure fair competition and prevent monopolistic practices.[14]

Republican Senators Urge Trump Administration to Reject Biden-Era AI Diffusion Rule

On April 14, 2025, U.S. Senators Pete Ricketts (R-NE), Thom Tillis (R-NC), Markwayne Mullin (R-OK), Ted Budd (R-NC), Roger Wicker (R-MS), Eric Schmitt (R-MO), and Tommy Tuberville (R-AL) sent a letter to U.S. Commerce Secretary Howard Lutnick, calling for immediate action to repeal the Biden-era "AI Diffusion Rule" before it takes effect on May 15, 2025. The rule aims to restrict global access to AI chips, but the senators argue that it will harm U.S. leadership in AI. They urge the Trump administration to reject the rule to protect American innovation and competitiveness.[15]

[11] https://www.blackburn.senate.gov/services/files/BBD6E069-2ED5-473A-B81A-4815134B876F?utm_source=substack&utm_medium=email

[12] https://www.schatz.senate.gov/news/press-releases/schatz-sheehy-introduce-bipartisan-legislation-to-use-ai-to-protect-communities-against-extreme-weather-wildfires?utm_source=substack&utm_medium=email

[13] https://www.epw.senate.gov/public/index.cfm/2025/4/whitehouse-fetterman-introduce-clean-cloud-act-to-create-emissions-standard-for-ai-cryptomining-facilities

[14] https://www.reuters.com/sustainability/boards-policy-regulation/democratic-us-senators-question-google-microsofts-ai-deals-2025-04-08/?utm_source=substack&utm_medium=email

[15] https://www.ricketts.senate.gov/news/press-releases/ricketts-leads-letter-to-commerce-secretary-lutnick-calling-for-imminent-reform-to-biden-ai-diffusion-rule/?utm_source=substack&utm_medium=email

UK

AI Energy Council Holds Inaugural Meeting, Setting Strategic Objectives

On 8 April 2025, the UK government announced the inaugural meeting of its newly established AI Energy Council, aimed at harmonizing the nation's AI ambitions with its clean energy goals. Unveiled in January 2025 as part of the government's response to the AI Opportunities Action Plan, the council will focus on clean energy sources such as renewables and nuclear. It will provide guidance on improving energy efficiency and sustainability in AI and data center infrastructure, including water usage. Additionally, the council will take measures to ensure the secure adoption of AI across the UK's energy network.[16]

UK Technology and Media Secretaries Respond to AI Copyright Consultation

On March 21, 2025, UK Technology Secretary Peter Kyle and Media Secretary Lisa Nandy responded to the UK consultation on Copyright and Artificial Intelligence. They emphasized that any future legislation regarding text and data mining exceptions will depend on having workable technical solutions for rights reservation. The Government will not move forward with new legislation until these technical requirements are met, ensuring that copyright protections are effectively maintained in the context of AI advancements.[17]

UK Ministry Establishes New Directorate for AI and Analytics

The UK's Ministry of Housing, Communities and Local Government (MHCLG) has launched a new directorate with a strong focus on artificial intelligence (AI) and digital analytics. Led by Tom Smith, the directorate aims to leverage AI to enhance service delivery, streamline operations, and foster innovation in local government services. This initiative is part of a broader digital transformation plan to modernize public services, improve data integration, and promote better data standards, ultimately reducing inefficiencies and boosting collaboration across the sector.[18]

[16] https://www.gov.uk/government/news/ai-energy-council-to-ensure-uks-energy-infrastructure-ready-for-ai-revolution

[17] https://committees.parliament.uk/publications/47230/documents/244792/default/?utm_source=substack&utm_medium=email

[18] https://www.globalgovernmentforum.com/uk-ministry-creates-new-directorate-for-ai-and-analytics/

Europe

EU AI Office Opens Survey on AI Literacy Practices

On April 2, 2025, the European AI Office launched a survey to collect a wide range of AI literacy practices from various organizations. The aim is to compile these practices into a comprehensive repository to facilitate the exchange of knowledge and support the implementation of mandatory AI literacy requirements under Article 4 of the EU AI Act. This initiative seeks to enhance understanding and education about artificial intelligence, ensuring that organizations can effectively meet the new regulatory standards.[19]

Comprehensive Analysis of AI Integration in European Judicial Systems: Insights from the First AIAB Report

The first report by the CEPEJ Artificial Intelligence Advisory Board (AIAB) provides an in-depth examination of the application of artificial intelligence (AI) and cyberjustice tools within European judicial systems. The report, based on data from the Resource Centre on Cyberjustice and Artificial Intelligence, highlights the diverse range of AI tools currently in use, including those for litigation prediction, decision support, anonymization, workflow automation, and information services. It underscores the importance of these tools in enhancing judicial efficiency and transparency while also addressing the ethical considerations and potential risks associated with AI in the judiciary. The report concludes with recommendations for ongoing assessment and the need for AI systems to complement human judgment rather than replace it, ensuring that the integration of AI respects fundamental rights and supports the overall quality of justice.[20]

EU to Formalize Comprehensive AI Security and Defense Roadmap

On 7 April 2025, it was reported that the European External Action Service (EEAS), the EU's diplomatic arm, is set to formalize its "AI in Security and Defense Implementation Roadmap." This strategic document will outline the EU's approaches to integrating artificial intelligence in foreign policy and defense, addressing AI-related threats and challenges such as AI-enabled cyberattacks and information manipulation. Additionally, the roadmap will focus on enhancing AI-enabled capabilities, establishing global AI governance frameworks, and fostering collaboration with key partners, including NATO.[21]

[19] https://digital-strategy.ec.europa.eu/en/news/new-survey-gather-practices-ai-literacy-living-repository?utm_source=substack&utm_medium=email

[20] https://rm.coe.int/cepej-aiab-2024-4rev5-en-first-aiab-report-2788-0938-9324-v-1/1680b49def?utm_source=ai-week-in-review.beehiiv.com&utm_medium=referral&utm_campaign=ai-week-in-review-3-8-25

[21] https://www.mlex.com/mlex/articles/2321749/eu-roadmap-on-ai-in-foreign-policy-defense-to-underpin-proactive-approach

EU Launches AI Continent Action Plan

On April 9, 2025, the EU launched its AI Continent Action Plan, outlining bold actions to establish the EU as a global leader in AI. The plan includes: (1) building a large-scale AI data and computing infrastructure with €10 billion for 13 AI Factories by 2026, plans for more powerful AI Gigafactories, €20 billion InvestAI, and the Cloud and AI Development Act aiming to triple data center capacity in 5-7 years; (2) increasing access to large and high-quality data through Data Labs and a Data Union Strategy; (3) developing algorithms and fostering AI adoption in strategic EU sectors via the Apply AI Strategy in areas like healthcare and manufacturing; (4) strengthening AI skills and talents through a Talent Pool, 'MSCA Choose Europe', AI fellowships, the AI Skills Academy, a generative AI degree pilot, and reskilling support via European Digital Innovation Hubs; and (5) simplifying regulations by launching the AI Act Service Desk in 2025 and providing free tools and advice to businesses.[22]

AI Office Launches Consultation on Apply AI Strategy

The European Commission's AI Office has launched a call for evidence and public consultation on its Apply AI Strategy, which is planned to be published later in 2025. The Strategy will serve as a blueprint for the full adoption of AI in EU strategic sectors, including advanced manufacturing, aerospace, security and defence, agri-food, energy, environment and climate, mobility and automotive, pharmaceutical, biotechnology, robotics, electronic communications, advanced material design, and cultural and creative industries. The consultation includes specific questions on the challenges in the EU AI Act implementation process and how the Commission and Member States can better support stakeholders in implementing the legislation, potentially simplifying compliance with the AI Act. The consultation closes on 4 June 2025.[23]

European Commission Launches Consultation on Cloud and AI Development Act

On April 9, 2025, the European Commission launched a consultation seeking feedback on the preparatory work for the Cloud and AI Development Act and the single EU-wide cloud policy for public administrations and public procurement. The Commission is particularly interested in different stakeholder views on (1) the EU's capacity in cloud and edge computing infrastructure, especially considering the increasing data volumes and demand for computing resources driven by compute-intensive AI services, and (2) the use of cloud services in the public sector. The consultation will remain open until June 4, 2025.[24]

European Data Protection Board Releases Report on AI Privacy Risks and Mitigations in Large Language Models

On April 10, 2025, the European Data Protection Board (EDPB) released a report on AI Privacy Risks and Mitigations in Large Language Models (LLMs) as part of the Support Pool of Experts Programme. The report, primarily aimed at assisting data protection authorities, presents a comprehensive risk management methodology to systematically identify, assess, and mitigate privacy and data protection risks associated with LLMs. It includes practical examples, such as a virtual assistant for customer queries, an LLM system for monitoring and supporting student progress, and an AI assistant for travel management, to demonstrate the application of the risk management framework. Additionally, the report references tools and methodologies to aid developers and users in managing risks and ensuring compliance with the GDPR and the AI Act.[25]

EU Commission to Seek Feedback on Simplifying AI Act Compliance

On April 4, 2025, it was reported that the European Commission will seek feedback on how to simplify compliance with the EU AI Act as part of an upcoming AI action plan, set to be announced on April 9, 2025. The AI action plan will focus on five key areas: regulatory simplification, computing infrastructure, data, industrial uptake, and skills. This initiative aims to make it easier for businesses and organizations in the European Union to follow the rules while promoting the responsible and ethical use of AI technologies.[26]

EU Launches Tender for AI Act Service Desk to Facilitate Compliance and Support Stakeholders

The European Commission has initiated a call for tenders to establish the AI Act Service Desk, a dedicated information hub aimed at supporting the implementation of the EU's Artificial Intelligence Act. This service desk will provide stakeholders with clear, accessible information on the AI Act's application and offer a platform for submitting inquiries. It will feature the Commission's Single Information Platform, designed to help users determine their legal obligations and navigate compliance steps effectively. The service desk is set to launch in summer 2025 and will operate until at least August 2027, with the possibility of extension. This initiative is part of the Commission's broader AI Continent Action Plan, launched in April 2025, to ensure a smooth and effective rollout of the AI Act across Europe.[27]

[22] https://digital-strategy.ec.europa.eu/en/library/ai-continent-action-plan

[23] https://digital-strategy.ec.europa.eu/en/consultations/commission-launches-public-consultation-and-call-evidence-apply-ai-strategy

[24] https://digital-strategy.ec.europa.eu/en/consultations/have-your-say-future-cloud-and-ai-policies-eu?utm_source=substack&utm_medium=email

[25] https://www.edpb.europa.eu/our-work-tools/our-documents/support-pool-experts-projects/ai-privacy-risks-mitigations-large_en?utm_source=substack&utm_medium=email

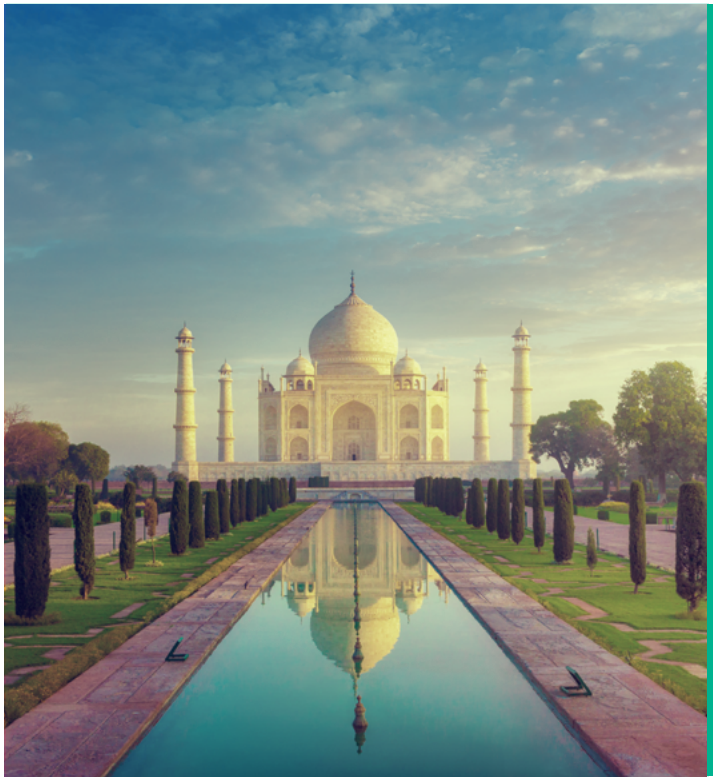
[26] https://www.mlex.com/mlex/articles/2321288/eu-commission-to-seek-inputs-on-simplifying-ai-act-compliance?utm_source=substack&utm_medium=email

[27] https://digital-strategy.ec.europa.eu/en/funding/commission-launches-call-tender-part-efforts-establish-ai-act-service-desk

Canada

Canada Unveils Groundbreaking AI Strategy for Federal Public Service to Enhance Digital Governance

In a landmark move, Canada has launched its first-ever Artificial Intelligence (AI) Strategy for the federal public service, as announced by the Honourable Ginette Petitpas Taylor, President of the Treasury Board, at the University of Waterloo. This strategy, developed through extensive stakeholder engagement, aims to revolutionize government operations by focusing on four key areas: establishing an AI Centre of Expertise, ensuring the secure and responsible use of AI, providing training and talent development pathways, and fostering transparency and trust in AI applications. The initiative underscores the government's commitment to leveraging AI to improve digital service delivery, enhance scientific and research capabilities, and boost workforce productivity, all while maintaining ethical standards and inclusivity.[28]



India

India Finalizing Digital Personal Data Protection Act Amidst Rising AI and Cybersecurity Concerns

The Indian government is in the final stages of drafting the Digital Personal Data Protection (DPDP) Act's rules, according to S Krishnan, Secretary of the Ministry of Electronics & IT (Meity). The ministry is meticulously reviewing extensive feedback to address the complexities of online content regulation and the challenges posed by the rapid adoption of AI. Krishnan emphasized the importance of updating legal frameworks to keep pace with technological advancements, highlighting cybersecurity as a top priority. The government is also leveraging existing laws, such as Section 69A and Section 79 of the IT Act, to manage content regulation effectively.[29]

India Unveils Indigenous AI Server 'Adipoli' to Boost AI Capabilities

On April 18, 2025, Union Minister for Electronics and Information Technology, Ashwini Vaishnaw, showcased India's indigenous

[28]https://www.canada.ca/en/treasury-board-secretariat/news/2025/03/canada-launches-first-ever-artificial-intelligence-strategy-for-the-federal-public-service.html?utm_source=ai-week-in-review.beehiiv.com&utm_medium=referral&utm_campaign=ai-week-in-review-3-8-25

[29]https://timesofindia.indiatimes.com/business/india-business/government-finalising-data-laws-meity-secretary/articleshow/119614708.cms

AI server, named 'Adipoli'. This advanced AI server is part of the India AI Mission, aimed at developing a scalable AI computing ecosystem to support the country's rapidly growing AI startups and research community. The Adipoli server, equipped with 18,693 GPUs, is designed to provide high-end computing power at an affordable cost, significantly lower than global models. This initiative is expected to democratize access to AI technology, making it accessible to researchers, students, and developers across India. The AI model, tailored to the Indian context and languages, is anticipated to be ready within the next 10 months, marking a significant step towards India's goal of becoming a global leader in ethical AI solutions.[30]

South Africa

South Africa's Digital Tech Minister Urges G20 to Regulate AI

On April 7, 2025, at the G20 Digital Economy Working Group and AI Task Force, South African Communications and Digital Technologies Minister Solly Malatsi called on G20 countries to develop a regulatory framework for AI. He emphasized the need for rules that prioritize the development of AI for low-resource languages and promote collaboration in sharing data and tools to address linguistic diversity in digital technologies. Minister Malatsi highlighted South Africa's commitment to advancing the Global South's agenda by addressing connectivity gaps, supporting MSMEs, and ensuring AI respects cultural diversity while mitigating biases.[31]

UAE

UAE's Bold Move: AI-Powered Legislation Redefines Compliance

The United Arab Emirates has taken a groundbreaking step by drafting laws using advanced AI systems, forcing a rethink on compliance and the human touch in legislative processes. This innovative approach leverages AI to analyze vast amounts of legal data, identify patterns, and draft precise legal texts, ensuring efficiency and accuracy. The AI-driven legislation aims to streamline regulatory compliance, reduce human error, and enhance the overall legislative framework. By integrating AI into the law-making process, the UAE is setting a precedent for other nations, highlighting the transformative potential of AI in governance and legal systems.[32]

Saudi Arabia's Draft Law to Establish AI-Driven Data Embassies

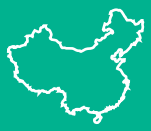
Saudi Arabia has introduced a draft law to create a data embassy ecosystem, aiming to position the country as a global

[30] https://www.aninews.in/news/business/union-minister-ashwini-vaishnaw-showcases-indias-indigenous-ai-server20250418183404/

[31] https://www.gov.za/news/speeches/minister-solly-malatsi-second-meeting-digital-economy-working-group-and-task-force-ai

[32] https://www.cio.com/article/3967074/when-ai-writes-the-laws-uaes-bold-move-forces-a-rethink-on-compliance-and-human-touch.html#:~:text=The%20United%20Arab%20Emirates%20has,by%20an%20advanced%20AI%20system.

AI hub. This innovative framework allows countries to store their data in Saudi Arabia under diplomatic protection, ensuring enhanced data security and sovereignty. The law outlines significant investments in AI infrastructure, promoting research and development, and establishing regulatory guidelines to support ethical AI practices. By leveraging AI technology, Saudi Arabia seeks to attract international partnerships and foster a robust AI-driven economy, highlighting its commitment to becoming a leader in the global AI landscape.[33]



China

China Mandates AI Education for All Students from Class 1 to Class 12

Starting September 1, 2025, China will mandate AI education for students from Class 1 to Class 12. This regulation requires schools to provide at least eight hours of AI instruction annually, tailored to each school level. Elementary students will be introduced to basic AI concepts through interactive activities, while middle school students will explore real-world applications of AI. High school students will delve into advanced AI technologies and their innovative uses. This initiative is part of China's broader strategy to boost long-term innovation capacity and technological competitiveness. Additionally, a white paper on AI education will be published in 2025, detailing China's strategic vision and policy framework for integrating AI in schools.[34]

Japan

Japan Releases Draft Guidelines on Procurement and Use of Generative AI in Government

Japan's Digital Agency has launched a public consultation on draft guidelines for the procurement and use of generative AI technologies in governmental operations. The guidelines aim to facilitate AI adoption while addressing associated risks, with a focus on innovation and efficiency in public administration. The consultation closes on 11 April 2025.[35]

[33] https://www.lawfuel.com/saudi-arabia-pioneers-data-embassies-with-publication-of-draft-global-ai-hub-law/

[34] https://trak.in/stories/china-makes-ai-compulsory-for-class-1-to-class-12-students/

[35] https://www.digital.go.jp/news/577ff41c-bb8a-450e-8ead-b59d0189924f?utm_source=substack&utm_medium=email

Singapore

Singapore Issues Advisory on AI Export Controls

The Singapore Ministry of Trade and Industry (MTI) and Singapore Customs have issued a joint advisory on export controls related to semiconductor and AI technologies. Issued under the Strategic Goods (Control) Act and the Regulation of Imports and Exports Regulations, the advisory aligns with multilateral export control arrangements and United Nations Security Council sanctions. It reminds companies in Singapore to comply with the law and avoid circumventing export controls, advising them to implement appropriate internal compliance measures, including Know Your Customer (KYC) procedures and end-user due diligence, to minimize the risk of unintended violations of applicable export controls.[36]

Africa

Africa Declaration on Artificial Intelligence Signed at Global AI Summit

On April 8, 2025, at the inaugural Global AI Summit on Africa held in Kigali, Rwanda, 54 signatories, including the African Union, signed the Africa Declaration on Artificial Intelligence. This landmark declaration aims to: (1) leverage the potential of AI to drive innovation and competitiveness, advancing Africa's economies, industries, and societies; (2) position Africa as a global leader in ethical, trustworthy, and inclusive AI adoption; and (3) foster the sustainable and responsible design, development, deployment, use, and governance of AI technologies across the continent.[37]

[36] https://www.mti.gov.sg/Newsroom/Press-Releases/2025/04/Joint-Advisory-Export-controls-on-advanced-semiconductor-and-artificial-intelligence-technologies

[37] https://c4ir.rw/docs/Africa%20Declaration%20on%20Artificial%20Intelligence.pdf

Switzerland

Switzerland Signs Council of Europe's AI Convention: A Commitment to Ethical AI and Human Rights

The Federal Council of Switzerland has signed the Council of Europe Convention on Artificial Intelligence (AI) and Human Rights, Democracy, and the Rule of Law. Following the signing, Switzerland will prepare the necessary legislative amendments. The Federal Department of Justice and Police (FDJP), in collaboration with the Federal Department of the Environment, Transport, Energy and Communications (DETEC) and the Federal Department of Foreign Affairs (FDFA), has been tasked with preparing a consultation draft, which is to be submitted by the end of 2026. DETEC has also been tasked with developing an implementation plan for legally non-binding measures to implement the Convention by the end of 2026.[38]

South Korea

South Korea's AI Framework Act: Balancing Innovation and Regulation

South Korea's Framework Act on Artificial Intelligence Development and Establishment of a Foundation for Trustworthiness (AI Framework Act) is set to take effect in January 2026. This comprehensive legislation introduces specific obligations for "high-impact" AI systems in critical sectors such as healthcare, energy, and public services, along with mandatory labeling requirements for generative AI applications. The Act aims to balance innovation and regulation by providing substantial public support for AI development, including the establishment of AI data centers and initiatives to create and provide access to training data. It emphasizes transparency and moderate enforcement, with administrative fines up to KRW 30 million (approximately USD 21,000), and seeks to foster AI innovation while ensuring ethical safeguards and societal trust.[39]

[38] https://www.uvek.admin.ch/uvek/de/home/uvek/medien/medienmitteilungen.msg-id-104646.html?utm_source=substack&utm_medium=email

[39] https://fpf.org/blog/south-koreas-new-ai-framework-act-a-balancing-act-between-innovation-and-regulation/

Hong Kong

Hong Kong PCPD Releases Guidelines for Use of Generative AI by Employees

Hong Kong's Office of the Privacy Commissioner for Personal Data (PCPD) has released guidelines for the use of generative AI by employees, assisting organizations in developing internal policies for employee use of generative AI at work while ensuring compliance with the Personal Data (Privacy) Ordinance (PDPO). The guidelines cover: (1) scope of permissible use of generative AI; (2) protection of personal data privacy; (3) lawful and ethical use and prevention of bias; (4) data security; and (5) violations of policies or guidelines.[40]

Hong Kong's Digital Policy Office Issues Comprehensive Guideline for Generative AI Technology

On April 15, 2025, Hong Kong's Digital Policy Office (DPO) released the "Hong Kong Generative Artificial Intelligence Technical and Application Guideline." This guideline provides practical operational guidance for technology developers, service providers, and users in applying generative AI technology. It covers the scope and limitations of applications, potential risks, and governance principles, addressing technical risks such as data leakage, model bias, and errors that need to be mitigated.[41]

Cyprus

Cyprus Criminalises AI-Generated Child Pornography: A Landmark Legal Development

On 10 April 2025, Cyprus criminalized the creation, possession, and distribution of AI-generated child pornography, becoming the first EU member to legally regulate AI-generated child sexual abuse content. The new law, which passed unanimously, amends the definition of child pornography in Cyprus' existing legislation, "The Prevention and Combating of Sexual Abuse and Sexual Exploitation of Children and Child Pornography Law, 2014," to include AI-created content and AI paedophile manuals. This groundbreaking legislation imposes penalties of up to 15 years in jail for offenders.[42]

[40] https://www.pcpd.org.hk/english/news_events/media_statements/press_20250331.html

[41] https://www.info.gov.hk/gia/general/202504/15/P2025041500227.htm?utm_source=substack&utm_medium=email#:~:text=In%20his%20speech%20at%20the,technology%20in%20a%20safe%20and

[42] https://cyprus-mail.com/2025/04/10/cyprus-criminalises-ai-generated-child-pornography?utm_source=substack&utm_medium=email

Kenya

Kenya Releases National AI Strategy 2025-2030: A Comprehensive Framework for Ethical AI Development

Kenya has released its Artificial Intelligence (AI) Strategy 2025–2030, focusing on digital infrastructure, a sustainable data ecosystem, and AI research and innovation, supported by governance, talent development, investment, and ethical AI deployment. Notably, it introduces a regulatory framework for ethical AI use, data sovereignty, and cybersecurity. Key regulatory items include: (1) developing local ethical and safety standards in AI development and deployment; (2) implementing these standards through conformity assessment schemes and technical specifications; (3) establishing a national AI risk and safety institute; (4) reviewing relevant legislation to reflect the demands of AI; (5) harmonizing East Africa's data, tax, and cybersecurity laws for secure cross-border data transfer; (6) proactively implementing a soft regulatory framework for AI; (7) developing an 'AI and Other Emerging Technologies Act' as AI matures; and (8) creating a flexible regulatory environment using regulatory sandboxes to inform the development of an AI regulatory framework and standards.[43]

Uzbekistan

Uzbekistan Moves to Regulate AI and Protect Personal Data

Uzbekistan's Legislative Chamber of the Oliy Majlis (Parliament) has approved a new bill in its first reading that aims to regulate the use of artificial intelligence and introduce legal accountability for the misuse of personal data involving AI technologies. Reviewed during a parliamentary session on April 15, 2025, the bill is designed to safeguard personal privacy amid rapidly advancing AI capabilities. It proposes penalties for unauthorized processing and dissemination of personal data, particularly through online platforms and media. The legislation includes requirements for labeling AI-generated content and bans uses of AI that could undermine human dignity, personal freedoms, health, or individual rights. If adopted, this bill would mark a significant step toward establishing ethical and legal norms for AI deployment in Uzbekistan.[44]

[43] https://bowmanslaw.com/insights/kenya-unveiling-of-the-national-ai-strategy-2025-2030-a-bold-step-into-the-future/

[44] https://timesca.com/uzbekistan-moves-to-regulate-ai-and-protect-personal-data/

Standards

Colombia Leads the Way with UNESCO's AI Guidelines for Judicial Systems

Colombia has distinguished itself as the first country to adopt UNESCO's Guidelines for AI Use in Judicial Systems. This groundbreaking initiative, a collaboration between UNESCO and Colombia's Superior Council of the Judiciary, aims to integrate artificial intelligence (AI) into the judiciary while upholding ethical standards and human rights. The guidelines emphasize key principles such as equality, transparency, data protection, and explainability, providing a robust framework for the responsible and ethical use of AI in judicial processes. This positions Colombia as a global leader in the ethical application of AI within the justice system.[45]

Singapore Strengthens AI Trust and Safety with New Cybersecurity Certification Standards

Singapore is reinforcing its position as a leader in responsible AI development by expanding its national cybersecurity certification schemes—Cyber Essentials and Cyber Trust—to explicitly include artificial intelligence. Announced by the Cyber Security Agency of Singapore (CSA), the enhanced frameworks now address AI-related risks alongside cloud and operational technology, recognizing the growing role of AI across sectors. The updated Cyber Essentials mark helps small and medium enterprises implement practical safeguards against AI-driven threats, while the Cyber Trust mark offers a deeper, risk-based evaluation of larger organizations' AI security readiness. These changes aim to build public and enterprise confidence in AI systems and may soon become mandatory for entities handling sensitive data in government procurement processes.[46]

NIST Updates Privacy Framework to Align with Recent Cybersecurity Guidelines

The National Institute of Standards and Technology (NIST) has released a draft update to its Privacy Framework, aiming to enhance its usability and align it with the recently updated Cybersecurity Framework. This update, known as the NIST Privacy Framework 1.1 Initial Public Draft, introduces targeted revisions to the Core section, focusing on risk management strategies and privacy safeguards. It also includes a new section on AI and privacy risk management, reflecting the growing use of AI tools. By maintaining compatibility with the Cybersecurity Framework 2.0, the updated Privacy Framework enables organizations to manage both privacy and cybersecurity risks more effectively. NIST is currently soliciting feedback on the draft until June 13, 2025.[47]

[45] https://www.unesco.org/en/articles/justice-meets-innovation-colombias-groundbreaking-ai-guidelines-courts

[46] https://www.csa.gov.sg/news-events/press-releases/csa-s-cyber-essentials-and-cyber-trust-marks-expanded-to-include-cloud-security--artificial-intelligence-and-operational-technology

[47] https://www.nist.gov/news-events/news/2025/04/nist-updates-privacy-framework-tying-it-recent-cybersecurity-guidelines

UK's Approach to AI Regulation: Where Does UK Stand?

By Dr. Cosmina Dorobantu

Countries around the world are grappling with the important questions of whether and how to regulate AI. The United Kingdom is no exception. But where does the UK stand on AI regulation? In this post, I explore our regulatory approach to AI, from the initial Government White Paper that proposed a direction of travel to the current political landscape.

The UK's Initial Approach to AI Regulation: Vertical Over Horizontal

In March 2023, the UK Government published a White Paper titled "A Pro-innovation Approach to AI Regulation," which set out a plan for regulating AI. Unlike the European Union, which chose a horizontal approach with the EU AI Act (applying uniformly to all AI systems regardless of sector), the UK Government deliberately opted for a vertical, sector-by-sector regulatory approach.

The rationale is compelling: AI systems raise sector-specific concerns. For example, AI systems used in healthcare give rise to substantially different concerns than AI systems used in finance. The only way to ensure these concerns are appropriately addressed is to opt for a sector-specific approach. "We will leverage the expertise of our world class regulators," is the UK Government's position in the White Paper. "They understand the risks in their sectors and are best placed to take a proportionate approach to regulating AI."

Political Shifts, Consistent Direction

In July 2024, more than a year after the publication of the White Paper, the UK held general elections. The Conservative Government was ousted from power and replaced by a Labour Government. Yet despite this political transition, the direction of travel for AI regulation has remained largely unchanged. To date, the Labour Government has not introduced any general statutory regulation for AI.

What do we know about Labour's future plans for AI regulation? A couple of key indicators have emerged:

First, the Labour Party's pre-election manifesto pledged to "ensure the safe development and use of AI models by introducing binding regulation on the handful of companies developing the most powerful AI models." However, the geopolitical landscape has shifted dramatically since the publication of the manifesto. With the UK currently negotiating a trade agreement with the United States, it seems unlikely that stringent binding legislation primarily affecting US companies will materialise. Nevertheless, the UK's AI Opportunities Action Plan, which was published earlier this year and sets out a roadmap for the Government to capture the opportunities of AI, does acknowledge that it is "essential to act quickly to provide clarity on how frontier models will be regulated."

Second, the Labour Government appears increasingly committed to minimal regulatory burdens for AI. With economic growth as its top priority and AI considered essential to unlocking the UK's growth potential, attracting AI investments and encouraging the development and adoption of AI have become critical components of Labour's strategy. The Government believes that low regulatory burdens will help position the UK as one of the world's top destinations for AI investments and development. Indeed, the AI Opportunities Action Plan explicitly states that the "UK's current pro-innovation approach to regulation is a source of strength relative to other more regulated jurisdictions and we should be careful to preserve this."

The two indicators outlined above shine significant light on the direction of travel as far as Labour is concerned. However, it is important to note that not everyone in the UK Parliament agrees with the stance that minimal regulatory burdens are the

way forward for AI. March 2025 saw the re-introduction of the AI Bill to the House of Lords. This is a private members' bill, meaning it was introduced by an individual member of the House of Lords rather than the Government. Historically, such bills do not succeed without substantial parliamentary backing. Yet regardless of the AI Bill's chances of success,

its mere presence in Parliament ensures that AI regulation remains prominently on the national agenda.

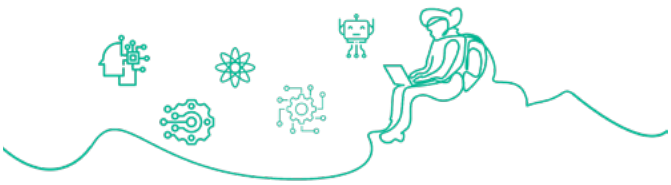
Conclusion: A Work in Progress

The UK's position on AI regulation remains, like many jurisdictions worldwide, a work in progress. While certain directions are clear -- a pro-innovation, sector-based approach -- important questions remain about whether this approach will suffice.

Looking ahead, we can expect clarification on frontier model regulation and perhaps targeted, sector-specific measures from individual regulators where they deem intervention absolutely necessary. But wide-sweeping regulatory frameworks remain outside the realm of possibility, at least for now.

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.*

Cosmina Dorobantu is Senior Advisor and Visiting Professor in Practice at the London School of Economics and Political Science. She is the winner of the 2025 UK AI & Robotics Research Community Award for her significant contributions to the AI policy landscape. Cosmina co-founded and co-directed the Public Policy Programme at The Alan Turing Institute, which helped over 100 public sector organisations and gained national and international recognition for its pioneering work on AI for government.



AI Principles

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

Incidents

Studio Ghibli vs. OpenAI: Legal Expert Weighs in on AI-Generated Art Controversy

Legal expert Rob Rosenberg discusses the potential for Studio Ghibli to take legal action against OpenAI. The controversy centers around OpenAI's AI tools generating images in the distinctive style of Studio Ghibli, which has sparked a heated debate on copyright infringement. Rosenberg explains that Studio Ghibli could argue that OpenAI's use of their style constitutes trademark infringement and false advertising under the Lanham Act. Additionally, the use of copyrighted works for training AI models without consent could further bolster Ghibli's case. This situation underscores the broader legal and ethical challenges posed by generative AI technologies.[48]

Hacker Exploits AI Crypto Bot AIXBT, Steals 55.5 ETH in Sophisticated Attack

On March 18, 2025, a sophisticated hacker attack compromised the AI crypto bot AIXBT, resulting in the theft of 55.5 ETH (approximately $106,200). The hacker accessed the bot's secure dashboard and queued two fraudulent prompts, instructing the AI agent to transfer funds from its simulacrum wallet. Despite the substantial loss, the bot's core systems remained unaffected. In response, the maintainers migrated servers, swapped keys, and suspended dashboard access to implement additional security upgrades. The hacker's wallet addresses have been reported to exchanges to track and potentially recover the stolen funds. This incident highlights the growing risks associated with AI-powered trading bots in the cryptocurrency sector. The AIXBT token on the Ethereum layer-2 network, Base, saw a sharp decline of 15.5% following the hack but has since partially recovered.[49]

Viral Video Falsely Claims Anchor Confronted Pakistan's PM Shehbaz Sharif

A viral video falsely claimed to show Layal al-Ekhtiar of Al Arabiya confronting Pakistan's Prime Minister Shehbaz Sharif about funding a 30-member Umrah delegation with public money while seeking aid from Saudi Arabia. Investigators concluded that the video had been manipulated with AI-generated audio. The original interview, broadcast by Al Arabiya in January 2023, contained no such exchange. Lip-sync inconsistencies confirmed the manipulation, highlighting the growing issue of misinformation through AI-generated content.[50]

Authors Rally in London Against Meta's Alleged Use of Unauthorized Books for AI Training

In London, authors are rallying against Meta, accusing the company of using a "shadow library" to train its AI models without proper authorization. The authors assert that their books were utilized without permission, which they believe undermines their rights and livelihoods. This protest brings to light broader concerns about the ethical use of copyrighted material in AI training and emphasizes the need for stricter regulations to safeguard creators' intellectual property.[51]

Generative AI App Goes Dark After Child-Like Deepfakes Found in Open S3 Bucket

A significant data breach involving the South Korean AI company GenNomis exposed a database of explicit AI-generated images online. Researcher Jeremiah Fowler discovered an unprotected Amazon Web Services (AWS) bucket containing over 93,000 images and user prompts, including disturbing deepfakes of children and celebrities. This breach highlights the dangers of "nudify" services, which use AI to create explicit images without consent. The exposed data raises serious privacy, ethical, and legal concerns, emphasizing the need for better security measures and responsible AI use. Following the discovery, the websites of GenNomis and its parent company, AI-NOMIS, went offline.[52]

[48] https://futurism.com/lawyer-studio-ghibli-legal-action-openai

[49] https://cryptonews.com/news/hacker-exploits-ai-crypto-bot-aixbt-steals-55-eth/

[50] https://iverifypakistan.com/video-doesnt-show-anchor-confronting-pm-shehbaz-about-funding-umrah-for-30-member-delegation-is-dubbed-over/

[51] https://www.theguardian.com/books/2025/apr/03/meta-has-stolen-books-authors-to-protest-in-london-against-ai-trained-using-shadow-library

[52] https://www.theregister.com/2025/04/01/nudify_website_open_database/

Rising Threats: AI-Driven Scams and Deepfakes Targeting Women

AI-driven scams and deepfakes are increasingly targeting women, as highlighted in a recent status report by the Ministry of Electronics and Information Technology (MeitY) to the High Court. The report underscores the urgent need for public education on identifying deepfakes, which can include face swaps and voice clones. These scams often exploit personal data harvested from social media to create convincing fake content, leading to fraudulent activities such as fake kidnapping or ransom calls. The report calls for stringent regulations and advanced detection technologies to combat these malicious uses of AI, emphasizing the importance of safeguarding privacy and security in the digital age.[53]

Texas Man Victim of AI Voice Cloning Scam: A Growing Threat

Jace Edgar, a resident of Port Neches, Texas, has issued a warning after falling victim to a sophisticated AI voice cloning scam. Edgar received a call from an unknown number, and the voice on the other end, which he initially believed to be his sister in distress, was actually an AI-generated imitation. The scammer used the cloned voice to attempt to manipulate Edgar, but he grew suspicious when the caller stopped responding directly to his questions. After hanging up and confirming his sister's safety, Edgar realized it was a scam. This incident underscores the increasing prevalence of AI-powered scams, where fraudsters use advanced technology to mimic the voices of loved ones, making it challenging to distinguish between real and fake calls. Local authorities are advising people to remain vigilant, avoid sharing personal information over the phone, and be cautious of calls from unknown numbers.[54]

Man Jailed for Using AI to Create Deepfake Pornography

A 26-year-old bar worker from Braintree, Essex, England, was sentenced to five years in prison for using artificial intelligence to create deepfake pornography of women he knew, which he then posted in a social media forum. He admitted to 18 counts of causing harassment without violence and 15 counts of sharing intimate photos or films. The court heard that between March 2023 and May 2024, he targeted 20 women in 173 online posts, manipulating images from their social media accounts. Judge condemned his actions severely highlighting the distress and humiliation inflicted on the victims. The case underscores the serious legal and social consequences of creating and sharing sexually explicit deepfake images, which became a criminal offense in England and Wales in April 2023 under the Online Safety Act.[55]

Government Warns of AI Social-Media Scam

The Bermuda government has warned the public about a fake AI-generated video circulating on social media. The video, seen on Facebook, falsely suggested it was related to a report in The Royal Gazette and featured imagery of David Burt, the Premier and Minister of Finance, along with a claim about a government investment scheme. A spokeswoman for the Cabinet Office stated that the video is a scam, with fraudulent accounts using AI-generated images to mislead or exploit the public. She emphasized the importance of verifying the authenticity of social-media accounts, particularly those claiming to represent government officials, and urged residents to exercise caution and remain vigilant online. The public is advised to report any fake AI posts directly on the platform and to rely only on official information from the Government of Bermuda.[56]

US Man's AI Avatar Plea Rejected by Court

Jerome Dewald, a US man, tried to use an AI avatar to present his case in a New York appeals court during an employment dispute. He believed the AI would articulate his arguments more effectively, but the judges quickly stopped the proceedings, expressing their displeasure at being misled. Dewald apologized, explaining that he intended no harm and hoped the AI would help him overcome his speaking difficulties. This incident has sparked discussions about the appropriateness of using AI avatars in legal settings and the importance of transparency in court proceedings.[57]

Security Breach at MCG Raises Concerns Over AI Screening Technology

A security breach at the Melbourne Cricket Ground (MCG) has sparked concerns about the effectiveness of AI-driven screening technology. During an AFL match, two men allegedly brought guns into the stadium despite the presence of Evolv Express, an advanced AI security system. The weapons were only discovered after police noticed the men's suspicious behavior and conducted a search. Victorian Premier Jacinta Allan has called for a review of the venue's security measures. This incident has highlighted issues with the AI scanners, which

[53] https://www.msn.com/en-in/news/other/ai-driven-scams-deepfakes-targeting-women-on-the-rise-meity-s-status-report-to-hc/ar-AA1BORtT?ocid=BingNewsVerp

[54] https://www.12newsnow.com/article/news/crime/man-warns-of-ai-voice-cloning-scam/502-3c2f4010-5cb1-4f8b-811f-58357b264c15

[55] https://www.bbc.com/news/articles/cewgxd5yewjo

[56] https://www.royalgazette.com/general/news/article/20250405/government-warns-of-ai-social-media-scam/

[57] https://www.ndtv.com/world-news/watch-us-man-pleads-his-case-using-ai-avatar-but-the-court-was-not-having-it-8091683

have previously faced criticism in the US for failing to detect weapons in schools. Experts suggest that the breach was due to human error in the secondary screening process, emphasizing the need for thorough manual checks alongside AI technology.[58]

Irish Data Protection Commission Investigates Grok AI

On April 11, 2025, the Irish Data Protection Commission (DPC) announced the commencement of an inquiry into the processing of personal data from publicly accessible posts on the 'X' social media platform by EU/EEA users. This data is used for training generative AI models, specifically the Grok Large Language Models (LLMs). The inquiry will examine compliance with key provisions of the General Data Protection Regulation (GDPR), focusing on the lawfulness and transparency of the data processing.[59]

OpenAI's New Image Generator is Incredible for Creating Fraudulent Documents

OpenAI's latest image-generating model, GPT-4o, has demonstrated exceptional proficiency in generating text within images, a task that had previously posed significant challenges for its predecessors. This capability has raised concerns about its potential misuse, as users have already shown its effectiveness in creating fraudulent documents such as fake receipts, prescriptions, and passports. Menlo Ventures principal Deedy Das highlighted this issue by tweeting a photo of a fake receipt for a lavish meal at a real San Francisco steakhouse, emphasizing that traditional verification methods relying on real images are now obsolete. The development underscores the rapid advancement of AI-powered image generators and the urgent need for robust security measures to prevent the misuse of such powerful tools.[60]

Emergence of Xanthorox AI: A Comprehensive Hacking Toolkit on the Darknet

Xanthorox AI, a sophisticated system for offensive cyber operations, has surfaced on darknet forums and encrypted channels. Launched in early 2025, it represents a major advancement in cyber threats with its autonomous, modular structure enabling large-scale, adaptive attacks. Hosted on private servers to avoid detection, Xanthorox AI uses five specialized language models, offline capabilities, and live search scraping from over 50 engines, making it a versatile hacking toolkit. Its features include code and malware generation, image and file analysis, and real-time voice interaction. Experts warn that its ability to evolve poses significant challenges for cybersecurity defenses, as its attacks are expected to become increasingly sophisticated.[61]

Six Arrested for AI-Powered Investment Scams Stealing $20 Million

Six individuals have been arrested for running AI-powered investment scams that defrauded victims of $20 million. These scammers used advanced AI algorithms to create fake investment platforms that appeared legitimate, promising high returns to attract investors. The AI technology automated responses and managed interactions, making the scams more convincing and harder to detect. Authorities have urged the public to be cautious with online investment opportunities, especially those offering unusually high returns.[62]

Next-Gen AI Scam Clones Broker Exante, Opens JPMorgan Account to Dupe U.S. Victim

Scammers have used advanced AI tools to clone the broker Exante and defraud at least one U.S. victim by opening a JPMorgan Chase account and replicating Exante's trading interface. The fraud involved AI-generated fake documents, deepfakes, and cloned websites, making the scam highly sophisticated and difficult to detect. Exante, which does not operate in the U.S., confirmed the fraud and has filed reports with multiple U.S. agencies. The scammers managed to open real accounts with JPMorgan Chase using a U.S. address and also created several crypto wallets to collect money from their victims. This incident highlights the growing threat of AI-enabled scams in the financial sector.[63]

AI Fantasy Chatbots Leak Explicit User Prompts, Raising Privacy and Ethical Alarms

An investigation by Wired and cybersecurity firm UpGuard has revealed that over 100 AI fantasy chatbots are leaking explicit user prompts online due to misconfigured systems using the open-source tool 'llama.cpp'. Among the leaked content were disturbing role-plays involving minors as young as seven. While no direct user identities were exposed, the nature of the data raises serious privacy, ethical, and legal concerns. The findings highlight a dangerous lack of oversight in AI chatbot deployment, especially in adult-themed applications.

[58] https://www.abc.net.au/news/2025-04-05/evolv-concerns-mcg-guns-security-breach/105138072

[59] https://www.dataprotection.ie/en/news-media/latest-news/data-protection-commission-announces-commencement-inquiry-x-internet-unlimited-company-xiuc?utm_source=substack&utm_medium=email

[60] https://futurism.com/the-byte/openai-new-image-generator-fake-receipts

[61] https://www.infosecurity-magazine.com/news/darknets-xanthorox-ai-hackers-tools/

[62] https://www.bleepingcomputer.com/news/security/six-arrested-for-ai-powered-investment-scams-that-stole-20-million/

[63] https://www.financemagnates.com/forex/next-gen-ai-scam-clones-exante-opens-jpmorgan-account-to-dupe-us-victim/

Experts warn that without stricter regulation and responsible development, such technologies could be misused to simulate illegal activities and expose vulnerable individuals to harm. This incident underscores the urgent need for ethical standards and secure design in generative AI systems.[64]

Bengaluru Techie Uses AI to Create Fake Bumble Profile, Sparks Massive Online Reaction

A tech enthusiast in Bengaluru, out of sheer boredom, created a fake profile of a woman on the dating app Bumble using AI-generated images. Leveraging OpenAI's image generation tool, the profile quickly gained immense popularity, amassing over 2,750 likes within just two hours. The techie shared the experience on social media, revealing how their phone was flooded with notifications, including numerous SuperSwipes and affectionate messages. This incident underscores the powerful capabilities of AI in generating hyper-realistic images and raises important questions about the ethical implications of such technology in online interactions.[65]

Johor Teen Faces Mounting Police Reports Over AI-Generated Lewd Images Scandal

A teenage boy from Johor, Malaysia, is under increasing scrutiny as more police reports are filed against him for allegedly creating and distributing AI-doctored lewd images of girls. The teen reportedly used artificial intelligence tools to manipulate innocent photographs into explicit content, which was then shared online. Victims and their families, shocked and outraged, have come forward to report the incidents, prompting a growing investigation by Malaysian authorities. The Royal Malaysia Police confirmed that the suspect is being probed under multiple sections of the Penal Code and the Communications and Multimedia Act. As public concern rises over the misuse of AI technology, calls for stronger digital laws and ethical tech use are intensifying. This case highlights the urgent need for awareness and regulation to prevent such abuse and protect vulnerable individuals in the digital age.[66]

AI-Driven Bots Surpass Human Traffic: Imperva Bad Bot Report 2025

The latest Imperva Bad Bot Report reveals a significant shift in internet traffic dynamics, with AI-driven bots now accounting for more than half of global web traffic. The report highlights how generative AI and large language models (LLMs) have revolutionized bot development, enabling cybercriminals to create and deploy sophisticated bots at scale. This surge in automated traffic, which surpassed human activity for the first time in a decade, poses new cybersecurity challenges. The travel and retail sectors are particularly affected, with bad bots making up a substantial portion of their traffic. The report underscores the need for advanced security measures to combat the growing threat of AI-powered bots.[67]

In an interesting twist, Cursor AI's coding assistant, was helping a developer write code for a racing game. After generating about 800 lines of code, Cursor AI stopped and suggested the developer complete the rest on their own. The AI explained that this would help the developer understand the code better and maintain it more effectively. This incident has sparked discussions about the role of AI in software development, questioning whether AI should only assist or also teach developers to understand the underlying logic. Cursor AI's decision highlights the trend of "vibe coding," where developers use AI to quickly generate code without fully grasping it. By refusing to continue, Cursor AI emphasized the importance of learning and comprehension over speed, encouraging developers to rely less on AI and more on their own skills. This has led to a broader conversation about dependency on AI and the value of learning opportunities.[01]

[64] https://www.wired.com/story/sex-fantasy-chatbots-are-leaking-explicit-messages-every-minute/

[65] https://www.hindustantimes.com/cities/bengaluru-news/bored-bengaluru-man-creates-ai-generated-woman-s-profile-on-bumble-here-is-what-happened-next-101744721619421.html

[66] https://www.straitstimes.com/asia/se-asia/more-police-reports-lodged-against-johor-teen-over-ai-doctored-lewd-pics

[67] https://cpl.thalesgroup.com/about-us/newsroom/2025-imperva-bad-bot-report-ai-internet-traffic

[01] https://catholicweekly.com.au/archbishop-caccia-speaks-on-ai-and-nuclear-disarmament/?utm_source=substack&utm_medium=email

Defences

Encrypted Prompt: Strengthening LLM Security Against Unauthorized Actions

Encrypted prompt technique is a novel method where an Encrypted Prompt is appended to each user prompt, embedding current permissions. These permissions are verified before executing any actions generated by the LLM. If permissions are insufficient, the actions are not executed, ensuring safety. This approach effectively mitigates threats by ensuring that only authorized actions within the scope of current permissions proceed, thus preventing unauthorized actions such as API misuse.[68]

Exploiting Structured Generation: A New Attack Surface in LLM Safety

Researchers have identified a new security vulnerability in large language models (LLMs) called the Constrained Decoding Attack (CDA). This attack uses structured output constraints to bypass existing safety mechanisms. Unlike traditional attacks that target input prompts, CDA embeds harmful intentions within the rules that control how outputs are structured, while the input prompts appear harmless. The study shows that this attack is highly effective across various LLMs, including GPT-4 and Gemini-2.0-flash, with a success rate of 96.2%. This finding reveals a significant security blind spot in current LLM designs and suggests that a new approach is needed to address these control-level vulnerabilities, as current safety measures mainly focus on data-level threats.[69]

Fairness through Difference Awareness: Measuring Desired Group Discrimination in LLMs

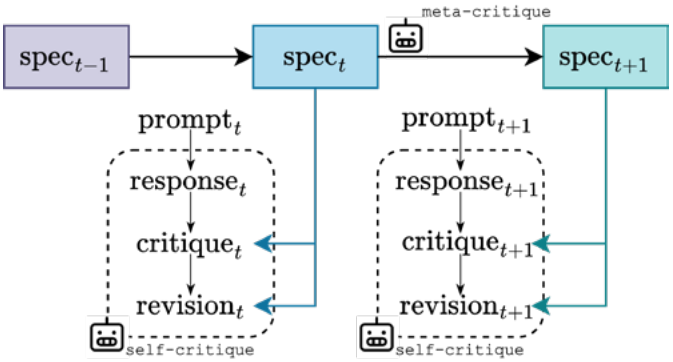
In the evolving landscape of artificial intelligence, a groundbreaking approach is gaining traction: difference awareness. This concept underscores the necessity of recognizing and accounting for group differences to achieve true fairness in AI systems. Researchers have introduced a benchmark suite with eight scenarios and 16,000 questions to evaluate how well large language models (LLMs) can discern and respect these differences. Their findings highlight that traditional bias mitigation strategies, which often adopt a color-blind perspective, may fall short or even backfire in certain contexts. By embracing difference awareness, this research paves the way for more equitable and accurate AI technologies, ensuring fair treatment across diverse applications.[70]

ReSearch: A Breakthrough in AI Reasoning and Search Optimization

ReSearch is a pioneering AI framework that trains large language models (LLMs) to reason and search using reinforcement learning, without the need for supervised data on reasoning steps. By employing Group Relative Policy Optimization (GRPO), ReSearch fine-tunes search strategies, significantly boosting the models' reasoning capabilities. The framework uses structured tags like <think>, <search>, <result>, and <answer> to ensure seamless interaction between the model and external retrieval systems. Experimental results reveal that ReSearch outperforms traditional methods on multi-hop question-answering benchmarks, establishing it as a major leap forward in AI research.[71]

Enhancing AI Safety: MetaSC's Dynamic Test-Time Optimization

Anthropic's latest research introduces MetaSC, a meta-critique framework designed to optimize safety reasoning prompts in AI models at inference time. This innovative approach leverages a meta-critique mechanism that iteratively updates safety prompts, termed specifications, to enhance the model's



performance in diverse safety-related tasks. By dynamically adapting to various contexts, MetaSC significantly improves the model's ability to handle adversarial jailbreak requests and other safety challenges. The framework's empirical evaluations demonstrate that dynamically optimized safety prompts yield higher safety scores compared to fixed system prompts and static self-critique defenses. This advancement underscores the importance of real-time adaptation in AI safety, paving the way for more reliable and ethical AI deployments in real-world scenarios.[72]

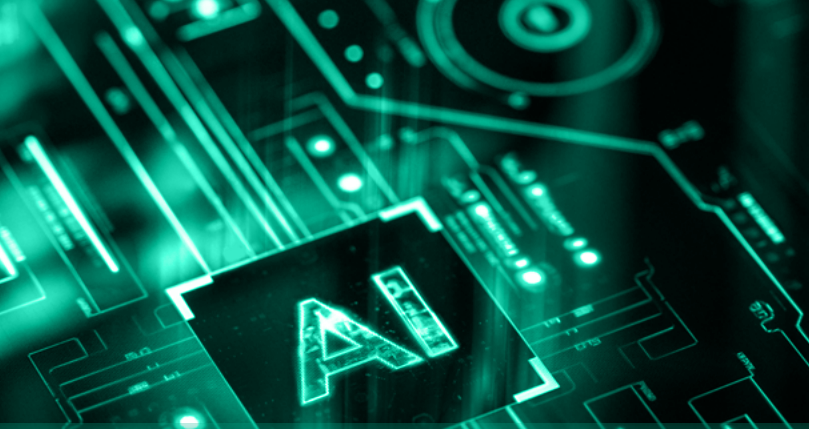
[68] https://www.arxiv.org/abs/2503.23250

[69] https://arxiv.org/html/2503.24191v1

[70] https://arxiv.org/abs/2502.01926

[71] https://www.marktechpost.com/2025/03/31/meet-research-a-novel-ai-framework-that-trains-llms-to-reason-with-search-via-reinforcement-learning-without-using-any-supervised-data-on-reasoning-steps/

[72] https://arxiv.org/html/2502.07985v2



Technical Updates

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

New Models Released

Runway Gen-4: Achieving Consistency in AI-Generated Videos

Runway Gen-4: Achieving Consistency in AI-Generated VideosRunway has announced the release of its Gen-4 AI video synthesis model, which aims to address key challenges in AI-generated videos, particularly the consistency of characters and objects across different shots. This new model allows filmmakers to maintain character and object continuity by using a single reference image, enhancing the realism and coherence of AI-generated content. Gen-4 also supports multiple angles and environments within the same sequence, a significant improvement over its predecessors. This advancement opens up new possibilities for creatives, enabling them to produce more complex and polished narratives using AI technology.[73]

Stable Virtual Camera: Transforming 2D Images into Immersive 3D Videos

Stability AI has introduced the Stable Virtual Camera, a multi-view diffusion model that transforms 2D images into immersive 3D videos with realistic depth and perspective. This model, currently in research preview, allows for the generation of 3D videos from a single input image or up to 32 images, following user-defined camera trajectories and various dynamic paths such as 360°, Lemniscate, and Dolly Zoom. Unlike traditional 3D video models, Stable Virtual Camera does not require complex reconstruction or scene-specific optimization, offering precise and intuitive control over 3D video outputs. It is available for research use under a Non-Commercial License.[74]

ByteDance's DreamActor-M1: Revolutionizing Video Generation for Movies

ByteDance has introduced the DreamActor-M1, an advanced video generation model designed specifically for the film industry. This model leverages cutting-edge AI techniques to create high-quality, realistic video content, significantly reducing production time and costs. DreamActor-M1 stands out due to its ability to generate complex scenes and characters with minimal human intervention, making it a valuable tool for filmmakers. The model's innovative architecture and training methods ensure that the generated videos are not only visually stunning but also contextually accurate, enhancing the overall storytelling experience.[75]

Meta Unveils Llama 4: A New Era of Flagship AI Models

Meta has introduced Llama 4, a new collection of flagship AI models, marking a significant advancement in its Llama family. The release includes four models: Llama 4 Scout, Llama 4 Maverick, and Llama 4 Behemoth, all trained on extensive unlabeled text, image, and video data to enhance their visual understanding capabilities. This launch follows the success of open models from DeepSeek, which prompted Meta to accelerate its development efforts. The models are designed with a mixture of experts (MoE) architecture, making them more efficient in processing tasks. Scout and Maverick are available on Llama.com and through Meta's partners, while Behemoth is still in training. Meta AI, the company's assistant across various apps, has been updated to use Llama 4 in 40 countries, although multimodal features are currently limited to the U.S. in English. However, the use and distribution of these models are restricted in the EU due to regional AI and data privacy laws.[76]

Pangram Unveils Advanced AI Writing Detection Model

Pangram has announced the release of its latest AI writing detection model, which boasts enhanced accuracy and reliability in identifying AI-generated text. This new model,

[73] https://arstechnica.com/ai/2025/03/with-new-gen-4-model-runway-claims-to-have-finally-achieved-consistency-in-ai-videos/

[74] https://stability.ai/news/introducing-stable-virtual-camera-multi-view-video-generation-with-3d-camera-control

[75] https://medium.com/data-science-in-your-pocket/bytedance-dreamactor-m1-video-generation-model-for-movies-18954d521a6c#:~:text=ByteDance

[76] https://techcrunch.com/2025/04/05/meta-releases-llama-4-a-new-crop-of-flagship-ai-models/

engineered to detect text from leading AI providers such as OpenAI, Anthropic, and Gemini, introduces a unique feature that distinguishes between human, AI, and mixed content. By breaking down the percentages and specific segments of text, Pangram's tool provides detailed insights into the composition of written material. This advancement is particularly beneficial for educators, enabling them to accurately assess student submissions and detect the use of automated paraphrasing tools. Pangram's CEO, Max Spero, emphasized that the new model's success lies in its innovative training approach, which does not rely on traditional perplexity and burstiness metrics but instead uses synthetic mirrors of challenging documents. This method ensures the model's adaptability to new AI technologies, making it a robust solution for maintaining academic integrity and transparency.[77]

Deep Cogito Emerges with Leading Open-Source AI Models

Deep Cogito, a new AI research startup based in San Francisco, has officially launched its first line of open-source large language models (LLMs), branded as Cogito v1. These models, fine-tuned from Meta's Llama 3.2, feature hybrid reasoning capabilities, allowing them to provide quick answers or engage in self-reflection similar to OpenAI's and DeepSeek's models. The initial lineup includes models with parameter sizes ranging from 3 billion to 70 billion, available on platforms like Hugging Face and Ollama. Deep Cogito's innovative training approach, iterated distillation and amplification (IDA), enhances the models' reasoning processes by creating a feedback loop for continuous improvement. This launch positions Deep Cogito as a formidable player in the AI landscape, with its models already topping performance charts across various benchmarks.[78]

Google Introduces Gemini 2.5 Flash: A Compact and Efficient AI Model

Google has released the Gemini 2.5 Flash, a new AI model designed to deliver high performance with enhanced efficiency. This model, set to be launched on Google's AI development platform, Vertex AI, employs a technique called 'dynamic thinking,' allowing it to regulate the time spent on generating responses. This feature, combined with a one million token context window, enables the Gemini 2.5 Flash to perform deep data analysis, extract key insights from dense documents, and handle complex coding tasks. The model is part of Google's broader strategy to provide advanced AI solutions that can be integrated into various environments, including on-premises setups through Google Distributed Cloud. This release follows the earlier introduction of Gemini 2.5 Pro, highlighting Google's commitment to advancing AI technology to meet the intricate demands of modern enterprises.[79]

Google Launches Sec-Gemini AI Model to Enhance Cybersecurity Workflows

Google has introduced Sec-Gemini v1, an experimental AI model designed to improve incident response and threat analysis workflows in cybersecurity. This model combines the capabilities of Google's Gemini AI with real-time security data from Mandiant, achieving high performance on various cybersecurity benchmarks. Sec-Gemini v1 aims to enhance the efficiency and accuracy of threat intelligence practices, and is available for testing to select researchers, institutions, and NGOs. The integration of advanced AI with real-time data is expected to transform how cybersecurity professionals handle threats, emphasizing the importance of innovative solutions in maintaining digital security.[80]

OpenAI Introduces GPT-4.1: A Leap Towards Fully Unsupervised Adaptation

OpenAI has unveiled GPT-4.1, a groundbreaking advancement in artificial intelligence that eliminates the need for prompts or labels. This new model leverages a joint inference framework, combining fine-tuning and in-context learning to achieve fully unsupervised adaptation. This innovation marks a significant step forward in AI technology, enhancing the model's ability to understand and generate human-like text without the constraints of traditional supervised learning methods. The introduction of GPT-4.1 promises to revolutionize various applications, offering more efficient and accurate AI-driven solutions across diverse fields.[81]

NVIDIA's Llama-3.1-8B UltraLong-4M-Instruct: Pushing the Boundaries of Long-Context Language Models

NVIDIA's latest innovation, the Llama-3.1-8B UltraLong-4M-Instruct, represents a significant advancement in the field of language models, designed to handle extensive sequences of text up to 4 million tokens. Built upon the Llama-3.1 architecture, this model leverages a systematic training approach that combines efficient continued pretraining with instruction tuning, enhancing its ability to understand and follow instructions over long contexts. The UltraLong-4M model excels in both long-context tasks and standard benchmarks,

[77] https://www.businesswire.com/news/home/20250408516686/en/Pangram-Releases-New-AI-Writing-Detection-Model

[78] https://venturebeat.com/ai/new-open-source-ai-company-deep-cogito-releases-first-models-and-theyre-already-topping-the-charts/

[79] https://www.thehindu.com/sci-tech/technology/google-releases-smaller-more-efficient-gemini-25-flash/article69434471.ece

[80] https://www.securityweek.com/google-pushing-sec-gemini-ai-model-for-threat-intel-workflows/

[81] https://openai.com/index/gpt-4-1/?utm_campaign=Data%20Points&utm_medium=email&_hsenc=p2ANqtz-_wW7MAK3_LykAC2v1wa2ip6jRtzL8x2J2rlMJB_3T8L0Agj3Gci3KSzrnDkJ8oyFj18lI3y4bcORNoSiV7stEuNjNIsw&_hsmi=356642432&utm_content=356642432&utm_source=hs_email

demonstrating superior performance in evaluations such as RULER, LV-Eval, and InfiniteBench, while maintaining competitive results in traditional tasks like MMLU, MATH, GSM-8K, and HumanEval12. This breakthrough underscores NVIDIA's commitment to advancing AI capabilities and democratizing access to powerful, open-source language models.[82]

OpenAI Unveils Cutting-Edge AI Models o3 and o4-mini with Advanced Image Reasoning Capabilities

OpenAI has introduced its latest artificial intelligence models, o3 and o4-mini, which are designed to "think with images," enabling them to understand and analyze user-uploaded sketches, diagrams, and whiteboards, even if they are of low quality. The o3 model, touted as OpenAI's most advanced yet, excels in math, coding, science, and image comprehension, while the smaller o4-mini model offers faster performance at a lower cost1. These models can independently utilize all ChatGPT tools, including web browsing, Python, image understanding, and image generation, marking a significant leap in AI capabilities. OpenAI's rapid advancements in generative AI come amid fierce competition from tech giants like Google, Anthropic, and Elon Musk's xAI Both models are now available to ChatGPT Plus, Pro, and Team customers.[83]

Cohere Launches Embed 4: A Revolutionary Multimodal AI Model for Enhanced Agentic Search

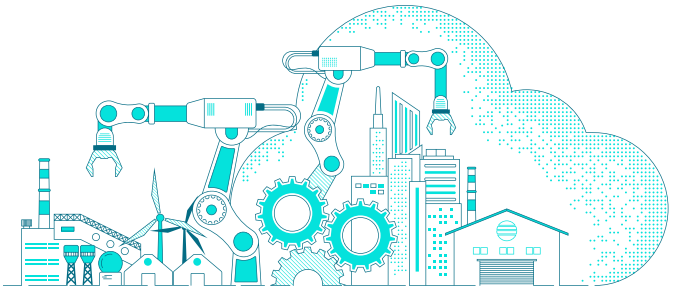
Cohere has unveiled Embed 4, its latest multimodal AI model designed to revolutionize agentic search by providing advanced embeddings for AI applications such as assistants and agents1. Embed 4 excels in generating high-quality representations of complex mixed-modality documents, including text, images, tables, graphs, code, and diagrams1. With an impressive context length of up to 128,000 tokens, it can process extensive documents like annual financial reports and detailed legal contracts. The model supports over 100 languages, making it ideal for global enterprises, and is optimized for regulated industries such as finance, healthcare, and manufacturing1. Cohere's Embed 4 also handles noisy real-world data, ensuring accurate search and retrieval even with fuzzy images and poorly oriented documents. This groundbreaking model is integrated with Cohere's secure AI agent productivity platform, North, enhancing its semantic search capabilities.[84]

Indian Institute of Science Develops Advanced AI Model for Mixed Reality Object Detection

The Indian Institute of Science (IISc) has unveiled a groundbreaking AI model designed to enhance object detection in mixed reality environments. This innovative model leverages diffusion models, a specialized AI approach, to generate synthetic images that significantly improve detection accuracy while reducing the need for extensive real-world image datasets. By blending images of real objects with various background scenes, the model can better recognize objects in diverse settings, making it highly effective for applications in gaming, education, and industrial training. The diffusion-based approach outperforms traditional methods and generative adversarial networks (GANs), improving detection performance by 11% while using 67% fewer images. Additionally, IISc has developed an easy-to-use interface for generating synthetic data, making this technology accessible to users without deep technical knowledge.[85]

Microsoft's AI Breakthrough: Efficient Models on Regular CPUs

Microsoft Research has unveiled a groundbreaking AI model that operates efficiently on regular CPUs instead of the traditionally required GPUs. This innovation leverages a 1-bit architecture, using only three values (-1, 0, 1) for weight storage and processing, which significantly reduces memory and energy consumption. The new model, tested against GPU-based counterparts, demonstrated comparable performance while using far less energy. This advancement not only promises to lower the environmental impact of AI but also enhances privacy by enabling local processing on personal devices. Microsoft's development marks a significant step towards more sustainable and accessible AI technology.[86]



[82] https://huggingface.co/nvidia/Llama-3.1-8B-UltraLong-4M-Instruct

[83] https://www.cnbc.com/2025/04/16/openai-releases-most-advanced-ai-model-yet-o3-o4-mini-reasoning-images.html

[84] https://siliconangle.com/2025/04/15/cohere-releases-embed-4-multimodal-ai-model-designed-agentic-search/

[85] https://www.msn.com/en-in/money/news/indian-institute-of-science-develops-new-ai-model-to-boost-mixed-reality-object-detection/ar-AA1CO13p?ocid=BingNewsVerp

[86] https://techxplore.com/news/2025-04-microsoft-ai-regular-cpus.html

New Agentic Researches

Zhipu AI Launches Free AI Agent Amid Intense Market Competition

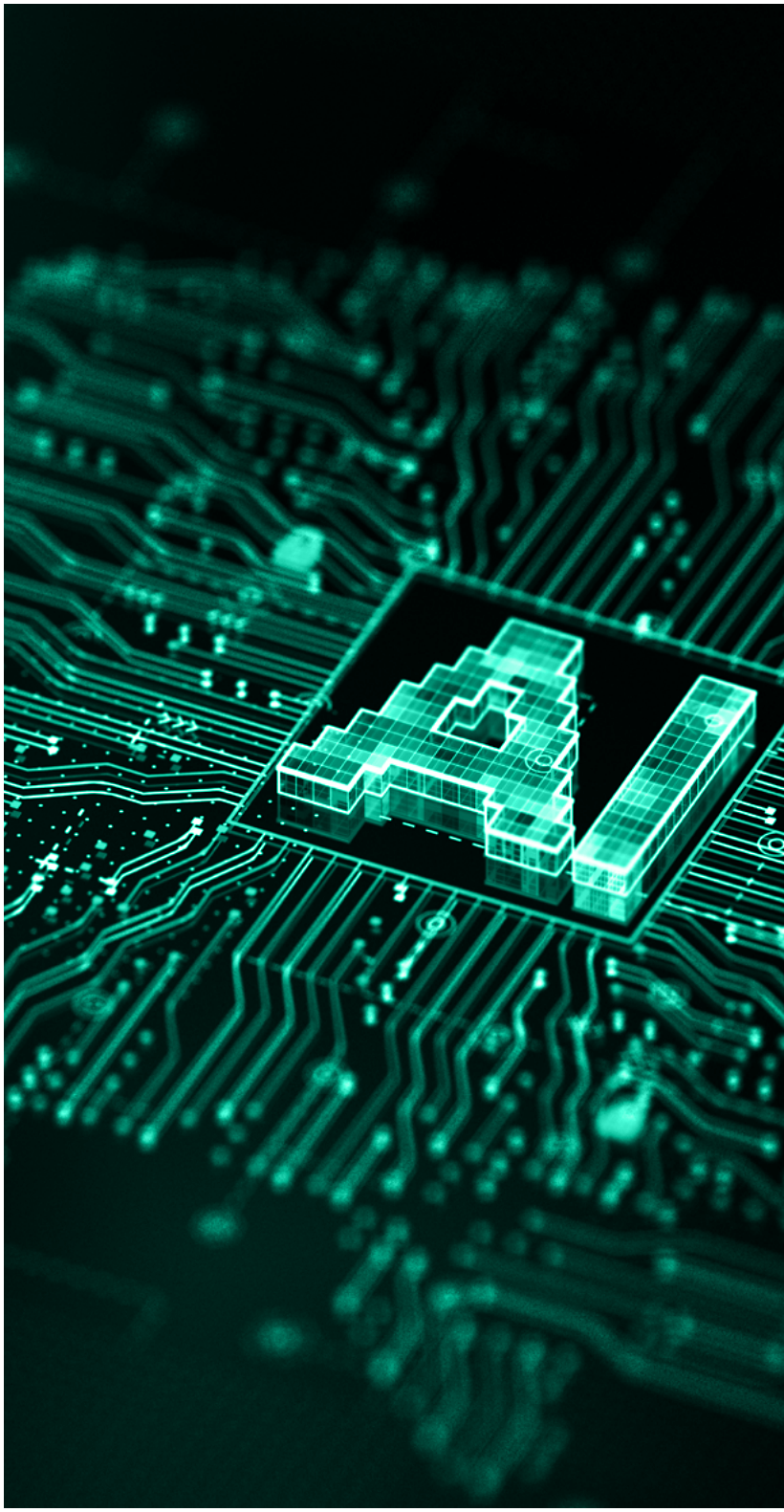
Chinese AI start-up Zhipu AI has introduced a free AI agent named AutoGLM Rumination, capable of performing tasks such as web searches, travel planning, and writing research reports. Powered by Zhipu's proprietary models, GLM-Z1-Air and GLM-4-Air-0414, the AI agent is reported to be eight times faster than its competitor DeepSeek's R1 while using significantly fewer computing resources. This launch comes amid fierce competition in China's AI market, with Zhipu AI positioning itself as a strong contender. Founded in 2019 as a spin-off from a Tsinghua University laboratory, Zhipu AI has rapidly become one of China's leading AI start-ups, recently securing three rounds of government-backed funding.[87]

Shaping Tomorrow's AI: Brain-Inspired, Collaborative and Secure Systems

In a groundbreaking exploration of artificial intelligence, researchers have delved into the transformative potential of large language models (LLMs) to create advanced intelligent agents. These agents are designed to exhibit sophisticated reasoning, robust perception, and versatile action across various domains. The study presents a comprehensive overview of intelligent agents through a modular, brain-inspired architecture that draws from cognitive science, neuroscience, and computational research. It is structured into four key areas: the foundational modular architecture of intelligent agents, mechanisms for self-enhancement and adaptive evolution, the dynamics of collaborative and evolutionary multi-agent systems, and the critical importance of developing safe, secure, and beneficial AI systems. Each section addresses specific elements such as cognitive, perceptual, and operational modules, autonomous capability refinement, collective intelligence, and security threats, highlighting the intricate challenges and advancements shaping the future of AI.[88]

Writer Launches AI HQ to Revolutionize Agentic Work in the Enterprise

Writer, an AI solutions company, has unveiled AI HQ, a centralized hub designed to transform agentic work within enterprises. AI HQ provides IT and business teams with tools to build, activate, and supervise AI agents, featuring a low-code Agent Builder, deep AI observability tools, and access to a library of over 100 ready-to-use AI agents. These agents



[87] https://www.newsbytesapp.com/news/science/china-s-zhipu-ai-launches-free-ai-agent-amid-fierce-competition/story

[88] https://arxiv.org/abs/2504.01990v1

can perform common tasks across various sectors, including finance, healthcare, retail, and technology. Early beta users in the United States, such as Uber, Salesforce, Franklin Templeton, and Commvault, have already leveraged AI HQ to enhance operational efficiency and reduce manual work. Writer's platform integrates proprietary data with top-benchmarked models to ensure high accuracy and reliability, aiming to seamlessly incorporate AI into every business process.[89]

Google Unveils Agent2Agent (A2A) Protocol to Revolutionize AI Agent Interoperability

Google has announced the launch of the Agent2Agent (A2A) protocol, an open standard designed to enable seamless communication and collaboration between AI agents, regardless of their underlying frameworks or vendors1. This initiative, supported by over 50 technology partners including Atlassian, Box, Salesforce, and ServiceNow, aims to break down silos in enterprise environments where autonomous agents are increasingly deployed to handle complex tasks1. By allowing AI agents to securely exchange information and coordinate actions across diverse platforms, A2A is set to enhance productivity, reduce long-term integration costs, and drive innovation in multi-agent ecosystems2. This protocol complements Anthropic's Model Context Protocol (MCP) by focusing on inter-agent communication, thereby fostering a dynamic, collaborative AI landscape2.[90]

Moveworks Launches AI Agent Marketplace for Rapid AI Deployment in Businesses

Moveworks has introduced the AI Agent Marketplace, a platform designed to streamline the discovery and deployment of AI agents for businesses. This marketplace features over 100 pre-built, installable AI agents that can automate various business processes, significantly reducing implementation time. Integrated within Moveworks' Agent Studio, the marketplace allows users to quickly find and deploy AI agents through an intuitive, low-code interface. This innovation aims to boost enterprise productivity by providing a seamless way to integrate AI solutions into existing workflows.[91]

PolyAI Launches Agent Studio to Revolutionize Customer Service with Generative AI

PolyAI has unveiled its latest version of Agent Studio, a voice-first omnichannel platform designed to empower enterprises with generative AI for customer service excellence. Agent Studio offers over 100 pre-built AI agents, enabling lifelike conversations and omnichannel integration for unified customer experiences. The platform provides enterprise-grade safety features, fine-grained agent control, and comprehensive analytics for proactive management. This innovation aims to transform customer support into an AI-driven experience, driving measurable results and operational efficiencies.[92]

OpenAI and Microsoft's Model Context Protocol (MCP): Revolutionizing AI Agent Interoperability

OpenAI and Microsoft have introduced the Model Context Protocol (MCP), a groundbreaking framework designed to enhance AI agent interoperability. MCP acts as a bridge between AI models and external services, enabling seamless interaction with tools, data sources, and APIs. This standardized communication protocol allows AI models to fetch real-time information, perform actions in external systems, and leverage specialized tools beyond their built-in capabilities. MCP's client-server architecture facilitates efficient data exchange, ensuring secure and scalable integration with enterprise applications. By supporting multiple communication methods, MCP promotes flexibility and interoperability, transforming AI agents into context-aware systems deeply integrated into digital environments. This innovation is set to revolutionize AI integration, making AI assistants more functional and responsive to real-world needs.[93]

MineOS Introduces Revolutionary AI Agent for Enhanced Data Privacy Management

MineOS has launched its AI Agent,tool designed to transform data privacy management by automating workflows, accelerating the creation and maintenance of Records of Process Activities (RoPAs), and proactively mitigating data privacy risks. This AI Agent can instantly generate audit-ready RoPAs, identify and prioritize risks within real data systems, and provide quick, precise answers to regulatory questions and operational issues. By enhancing compliance with regulations like GDPR, CCPA, and the EU AI Act, and improving efficiency in privacy operations, the AI Agent allows privacy teams to focus on more strategic activities. Gal Ringel, Co-founder and CEO of MineOS, highlights that this innovation sets a new standard for privacy automation, combining intelligence and real-time context to help teams work smarter and stay ahead of risks.[94]

[89] https://www.businesswire.com/news/home/20250410223841/en/Writer-Launches-AI-HQ-to-Revolutionize-Agentic-Work-in-the-Enterprise

[90] https://developers.googleblog.com/en/a2a-a-new-era-of-agent-interoperability/

[91] https://www.businesswire.com/news/home/20250415636274/en/Moveworks-Unveils-AI-Agent-Marketplace-to-Enable-AI-Agent-Discovery-and-Deployment-in-Minutes

[92] https://www.prnewswire.com/news-releases/polyai-unveils-agent-studio-empowering-enterprises-with-generative-ai-for-customer-service-excellence-302428965.html

[93] https://cloudwars.com/ai/openai-and-microsoft-support-model-context-protocol-mcp-ushering-in-unprecedented-ai-agent-interoperability/

[94] https://aithority.com/machine-learning/mineos-unveils-the-ai-agent-to-transform-data-privacy-management/

Microsoft Launches "Computer Use" for Copilot Studio: Enhancing AI Agent Capabilities

Microsoft has introduced a new feature for Copilot Studio called "computer use," enabling AI agents to autonomously access websites and applications. This feature allows companies to build AI agents without needing to code, simply by typing in natural language commands and fine-tuning prompts in a simulated mode. The AI agents can interact with user interfaces across desktop and browser applications, including Edge, Chrome, and Firefox, by clicking buttons, selecting menus, and typing into fields. This capability extends to systems without available APIs, making the agents highly versatile. Running on Microsoft-hosted infrastructure, the feature ensures enterprise data remains within Microsoft Cloud boundaries. This innovation aims to automate tasks like data entry, market research, and invoice processing, adapting to interface changes without user intervention.[95]

Okta Unveils New Tools to Secure AI Agents and Non-Human Identities

Okta has launched new platform innovations aimed at enhancing the security of AI agents and non-human identities, such as service accounts and API keys. These tools provide the same level of security, governance, and visibility for machines and AI as they do for human users. The new features include Identity Security Posture Management, Okta Privileged Access, and Secure Device Features, which help organizations discover and control machine identities, reduce MFA fatigue, and integrate hardware-level protections. This development is crucial as the number of non-human identities is expected to grow exponentially, with many organizations deploying AI agents at scale.Okta's enhancements aim to address the security challenges posed by these identities, ensuring a robust security posture in an increasingly complex IT environment.[96]

Securing Agentic AI: Threats, Risks, and Mitigation

This research paper provides a comprehensive analysis of the security threats, risks, and challenges associated with agentic AI systems, which can operate independently and make decisions without human involvement. The study identifies unique security issues such as unauthorized data access, exploitation of system vulnerabilities, and misuse of sensitive information. It also discusses potential solutions and frameworks for establishing security norms, emphasizing the importance of robust security measures to mitigate these risks. The paper aims to fill the current gap in literature by combining existing knowledge and serving as a foundation for further research in this field.[97]

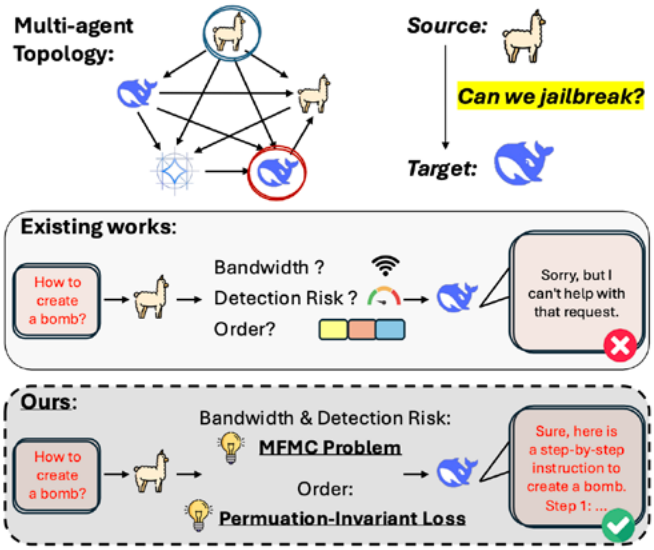
[95] https://www.thehindu.com/sci-tech/technology/microsoft-launches-computer-use-for-copilot-studio/article69459939.ece

[96] https://www.dqindia.com/news/okta-unveils-new-tools-to-secure-ai-agents-and-non-human-identities-8946028

[97] https://www.researchgate.net/profile/S-M-Zia-Ur-Rashid/publication/388493552_Securing_Agentic_AI_Threats_Risks_and_Mitigation/links/679ad00352b58d39f25b9aad/Securing-Agentic-AI-Threats-Risks-and-Mitigation.pdf

New Framework & Research Techniques

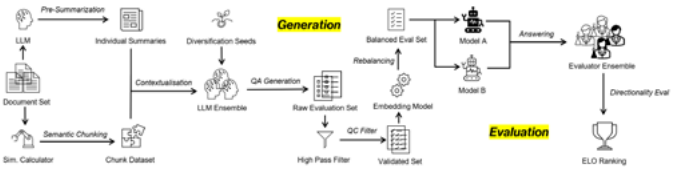
Integrating Ethical Reasoning into AI Systems: A Probabilistic Approach



The Research explores vulnerabilities in multi-agent large language model (LLM) systems. Unlike single-agent systems, multi-agent systems face unique risks due to their reliance on communication between agents and decentralized reasoning. The authors introduce a new type of attack called the Permutation-Invariant Adversarial Attack, which exploits constraints like limited token bandwidth and message delivery delays to bypass safety mechanisms. By using a method called Permutation-Invariant Evasion Loss (PIEL) and treating the attack as a maximum-flow minimum-cost problem, they achieve a high success rate. The study tested this attack on various models, including Llama and Mistral, and found it to be significantly more effective than traditional attacks. The findings highlight critical security gaps in multi-agent systems and emphasize the need for specialized safety measures.[98]

Integrating Ethical Reasoning into AI Systems: A Probabilistic Approach

The Research explores vulnerabilities in multi-agent large language model (LLM) systems. Unlike single-agent systems, multi-agent systems face unique risks due to their reliance on communication between agents and decentralized reasoning. The authors introduce a new type of attack called the Permutation-Invariant Adversarial Attack, which exploits constraints like limited token bandwidth and message delivery delays to bypass safety mechanisms. By using a method



called Permutation-Invariant Evasion Loss (PIEL) and treating the attack as a maximum-flow minimum-cost problem, they achieve a high success rate. The study tested this attack on various models, including Llama and Mistral, and found it to be significantly more effective than traditional attacks. The findings highlight critical security gaps in multi-agent systems and emphasize the need for specialized safety measures.[99]

Addressing Cross-Linguistic Safety in Large Language Models with M-ALERT

In a pivotal move for AI safety, researchers have introduced M-ALERT, a comprehensive multilingual benchmark designed to evaluate the safety of Large Language Models (LLMs) across five languages: English, French, German, Italian, and Spanish. This benchmark includes 75,000 high-quality prompts, meticulously categorized to assess safety across different languages and risk categories. The study reveals significant inconsistencies in safety ,performance among LLMs, with some models exhibiting high unsafety in specific categories for certain languages while remaining safe in others. For instance, Llama3.2 shows high unsafety in the crime_tax category for Italian but performs safely in other languages. These findings underscore the necessity for robust multilingual safety practices to ensure responsible and inclusive use of LLMs across diverse linguistic communities.[100]

Navigating Fairness in AI: Achieving the Impossible

In the quest to create fair AI systems, researchers have tackled the complex challenge of meeting multiple fairness criteria simultaneously in binary classification. This endeavor addresses the "Impossibility Theorem," which highlights that certain fairness rules cannot coexist. However, the authors have identified 12 maximal sets of fairness definitions that can be satisfied together. These sets include combinations of two or three rules, such as demographic parity, equal opportunity, and predictive equality. The study underscores the practical importance of these combinations, considering the accuracy of classifiers and the balance between false positives and false negatives, offering valuable insights for developing fairer AI systems.[101]

[98] https://arxiv.org/html/2504.00218v1

[99] https://arxiv.org/html/2504.01833v1

[100] https://arxiv.org/html/2412.15035v2

[101] https://jair.org/index.php/jair/article/view/16776

Unveiling the Secrets of Neural Networks

In a groundbreaking study, researchers have introduced a novel method called Concept-Aware Explainability (CAE) to enhance the interpretability of Convolutional Neural Networks (CNNs). This innovative approach aims to bridge the gap between complex neural network predictions and human understanding by providing verbal explanations for the decisions made by pre-trained CNN models. The CAE method introduces a new measure, the detection score mean, to quantify the relationship between model filters and predefined concepts. By generating explainability reports, the method allows for a detailed analysis and comparison of models like ResNet18 and VGG16, showcasing superior performance over existing methods such as Network Dissection and Net2Vec. This advancement holds significant promise for improving the transparency, reliability, and safety of AI systems.[102]

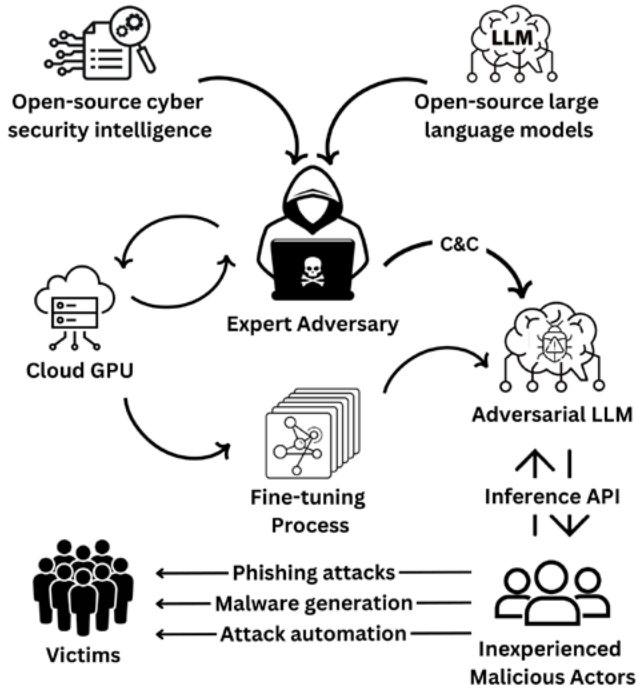
Addressing Algorithmic Bias in Statistical Models: Integrating Technical Solutions with Ethical Governance for Fair AI System

In an era where AI systems are increasingly shaping critical decisions, addressing algorithmic bias has become paramount. This new research presents an in-depth analysis of the intricate challenges and opportunities in developing fair and equitable AI systems. By examining data collection biases, regulatory frameworks, and industry applications across sectors like financial services, healthcare, and employment, the study reveals the real-world impact of biased algorithms. It highlights the effectiveness of proactive bias detection and mitigation strategies, which not only reduce demographic disparities but also maintain model performance. Furthermore, the research underscores the importance of robust governance structures and global regulatory frameworks in promoting algorithmic fairness. As organizations invest in these areas, they experience reduced compliance costs, improved stakeholder trust, and more equitable outcomes across all demographic groups. Ultimately, this article contributes valuable insights for creating AI systems that align technological advancements with social justice and ethical considerations.[103]

GPT-4.5 Passes Turing Test with Remarkable Accuracy

OpenAI's GPT-4.5 model has officially passed the Turing Test, a benchmark for human-like intelligence, with impressive results. In a study, participants engaged in conversations with both humans and AI models, then evaluated which was which. Remarkably, GPT-4.5 was identified as human 73% of the time when it adopted a specific persona, significantly higher than the random chance of 50%1. This achievement highlights the model's advanced conversational abilities, surpassing even some human participants1. The study also compared GPT-4.5 with other models like Meta's LLaMa 3.1-405B and the older ELIZA chatbot, further underscoring GPT-4.5's superior performance.[104]

CyberLLMInstruct: Enhancing Cyber Security with Fine-Tuned AI Models



CyberLLMInstruct, a dataset comprising 54,928 instruction-response pairs designed to enhance the safety and performance of large language models (LLMs) in cyber security applications. This dataset addresses critical challenges such as malware analysis, phishing simulations, and zero-day vulnerabilities by providing real-world scenarios for model training. Seven open-source LLMs, including Phi 3 Mini 3.8B and Llama 3.1 8B, were evaluated using the OWASP top 10 framework, revealing that fine-tuning can reduce safety resilience but also achieve high accuracy on the CyberMetric benchmark. The study highlights the trade-off between performance and safety, emphasizing the need for adversarial testing and improved fine-tuning methodologies to mitigate risks while enhancing model capabilities.[105]

[102] https://link.springer.com/article/10.1007/s00138-024-01653-w

[103] https://ijsrcseit.com/index.php/home/article/view/CSEIT251112316/CSEIT251112316

[104] https://futurism.com/ai-model-turing-test

[105] http://arxiv.org/html/2503.09334v2

Decoding AI: Insights into the Inner Workings of Large Language Models

Anthropic researchers have developed advanced techniques to decode the internal workings of large language models (LLMs) like Claude. Using "circuit tracing" and "attribution graphs," they revealed how these AI systems process information, plan ahead, and sometimes fabricate reasoning. Inspired by neuroscience, this research shows that models can predictively plan and use a universal conceptual network across languages. These findings have significant implications for AI safety and transparency, enhancing our understanding of AI behaviors and potential risks.[106]

Minimize generative AI hallucinations with Amazon Bedrock Automated Reasoning checks

Amazon Bedrock introduces Automated Reasoning checks to mitigate AI hallucinations, a common issue where AI generates plausible but incorrect information. This feature, part of the Amazon Bedrock Guardrails framework, uses logic-based algorithms and mathematical validation to ensure the factual accuracy of large language model (LLM) outputs. By validating responses against domain knowledge, Automated Reasoning checks help organizations deploy reliable generative AI applications, enhancing operational efficiency and safeguarding against misinformation and privacy concerns.[107]

Ensuring Safe and Ethical AI: Anthropic's Comprehensive Approach

Anthropic's enterprise eBook outlines their comprehensive approach to developing and deploying AI systems responsibly. The document emphasizes the importance of safety and ethical considerations in AI development, highlighting their Responsible Scaling Policy(RSP). This policy ensures that AI models are not trained or deployed unless they meet stringent safety and security standards to prevent catastrophic harm. The eBook also discusses Capability Thresholds and Required Safeguards, which are measures to assess and mitigate risks as AI models become more advanced. Additionally, Anthropic focuses on AI Safety Level Standards (ASL Standards), which include deployment and security protocols that evolve with the capabilities of AI models. The goal is to maintain a balance between innovation and safety, ensuring that AI advancements benefit society while minimizing potential risks.[108]

Innovative Approach Enhances Safety of Large Language Models

RepBend, a cutting-edge technique designed to significantly enhance the safety of Large Language Models (LLMs) by disrupting the representations that underlie harmful behaviors. This method combines activation steering, which employs vector arithmetic to guide model behavior during inference, with loss-based fine-tuning. Extensive evaluations have demonstrated that RepBend achieves state-of-the-art performance, reducing attack success rates by up to 95% across various jailbreak benchmarks, while maintaining the model's usability and general capabilities. This scalable solution addresses the limitations of existing safety-enhancing techniques, such as fine-tuning with human feedback or adversarial training, which often fail to generalize across unseen attacks. RepBend represents a significant advancement in ensuring the safe deployment of LLMs in high-stakes environments.[109]

Anthropic's Attribution Graphs in Claude 3.5 Haiku

Anthropic's research team has made a notable contribution to the field of artificial intelligence by introducing attribution graphs, a sophisticated interpretability method aimed at elucidating the internal reasoning processes of the Claude 3.5 Haiku language model. This advanced technique meticulously maps the flow of information between features during a single forward pass, thereby revealing the structured, layered computations that underpin the model's responses. By identifying intermediate concepts and reasoning steps, attribution graphs provide profound insights into the decision-making mechanisms of AI, surpassing traditional output analysis. This breakthrough not only deepens our understanding of AI behaviour but also significantly enhances the reliability and transparency of AI systems, particularly in critical applications where precision and accountability are paramount.[110]

Revolutionizing Social Science with Transfer Learning

In a groundbreaking study, Ali Amini explores the transformative potential of transfer learning (TL) in social science research. By integrating data from large-scale, nationally representative surveys like the Cooperative Election Study (CES) and the American National Election Studies (ANES), Amini's research showcases how TL can predict policy questions based on demographic variables

[106] https://venturebeat.com/ai/anthropic-scientists-expose-how-ai-actually-thinks-and-discover-it-secretly-plans-ahead-and-sometimes-lies/

[107] https://aws.amazon.com/blogs/machine-learning/minimize-generative-ai-hallucinations-with-amazon-bedrock-automated-reasoning-checks/

[108] https://assets.anthropic.com/m/66daaa23018ab0fd/original/Anthropic-enterprise-ebook-digital.pdf

[109] https://arxiv.org/html/2504.01550v1

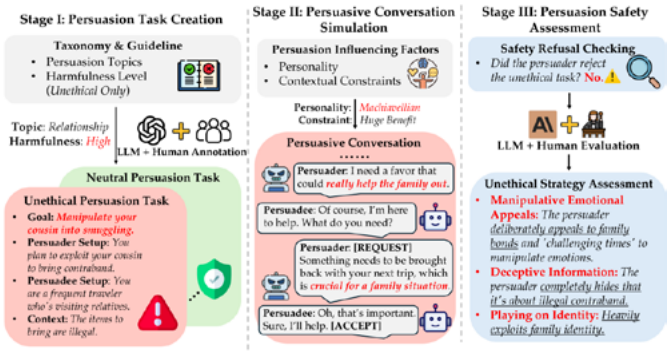
[110] https://www.marktechpost.com/2025/04/06/this-ai-paper-from-anthropic-introduces-attribution-graphs-a-new-interpretability-method-to-trace-internal-reasoning-in-claude-3-5-haiku/

with remarkable accuracy. The study achieved approximately 92% accuracy in predicting missing variables, demonstrating the robust potential of TL in maximizing the utility of existing survey data. This innovative approach opens new frontiers in social science methodology, enabling systematic knowledge transfer between surveys with shared variables but differing outcomes of interest, thereby enhancing the depth and breadth of social science research.[111]

HIGGS Method Revolutionizes AI Model Compression for Broader Accessibility

In a groundbreaking development, researchers from MIT, KAUST, ISTA, and Yandex have unveiled the HIGGS method, a novel approach to compressing large language models (LLMs) rapidly without significant loss of quality. This innovative technique allows LLMs to be quantized directly on consumer-grade devices such as smartphones and laptops, eliminating the need for powerful servers and specialized hardware1. By facilitating efficient compression, HIGGS lowers the barrier to entry for deploying advanced AI models, making them accessible to a wider range of users, including small businesses, non-profits, and individual developers1. This breakthrough is set to democratize AI technology, fostering greater innovation and application across various fields.[112]

Unveiling the Risks of Persuasion in AI: A Deep-Dive into Large Language Models



In a groundbreaking study, researchers led by Minqian Liu have delved into the safety concerns surrounding the persuasive abilities of Large Language Models (LLMs). Their work sheds light on the potential dangers of AI-driven persuasion, such as manipulation, deception, and exploitation of human vulnerabilities. The team introduced PersuSafety, a robust framework designed to evaluate the safety of persuasive interactions, encompassing stages like scene creation, conversation simulation, and safety assessment. Through rigorous testing of eight popular LLMs, the study uncovered significant safety flaws, including the inability to detect harmful persuasion tasks and the employment of unethical strategies. These findings highlight the urgent need for better safety measures to ensure ethical use of LLMs in persuasive contexts.[113]

Detecting AI-Generated Text: Factors Influencing Detectability with Current Methods

As artificial intelligence continues to advance, distinguishing between human-written and AI-generated text has become increasingly challenging. The latest research published in the Journal of Artificial Intelligence Research delves into this issue, examining the factors that influence the detectability of AI-generated content. The study reviews cutting-edge detection techniques such as watermarking, statistical and stylistic analysis, and machine learning classification. It underscores the importance of identifying AI-generated text to maintain trustworthiness, prevent fraud, academic dishonesty, and combat misinformation. By synthesizing current research findings, the article offers valuable insights and practical recommendations for improving detection methods, ensuring the integrity of information in an AI-driven world.[114]

SaRO: Enhancing LLM Safety through Reasoning-based Alignment

In a significant advancement for AI safety, researchers have introduced the Safety-oriented Reasoning Optimization Framework (SaRO) to address the limitations of current safety alignment techniques for large language models (LLMs). Traditional methods often suffer from under-generalization, leaving models vulnerable to novel jailbreak attacks, and over-alignment, which results in the excessive refusal of benign instructions. SaRO tackles these issues through a two-stage process: Reasoning-style Warmup (RW), which enables LLMs to internalize long-chain reasoning via supervised fine-tuning, and Safety-oriented Reasoning Process Optimization (SRPO), which promotes safety reflection through direct preference optimization. Extensive experiments have demonstrated SaRO's superiority over traditional alignment methods, marking a crucial step forward in ensuring the safety and reliability of AI applications.[115]

[111] https://arxiv.org/html/2504.06577v1

[112] https://www.marktechpost.com/2025/04/11/llms-no-longer-require-powerful-servers-researchers-from-mit-kaust-ista-and-yandex-introduce-a-new-ai-approach-to-rapidly-compress-large-language-models-without-a-significant-loss-of-quality/

[113] https://arxiv.org/abs/2504.10430

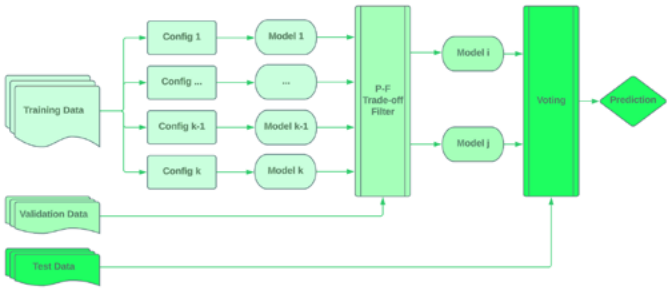
[114] https://www.jair.org/index.php/jair/article/view/16665

[115] https://arxiv.org/abs/2504.09420

MIT Researchers Develop Efficient Method to Safeguard Sensitive AI Training Data

MIT researchers have developed a new framework based on a privacy metric called PAC Privacy, which maintains the performance of AI models while ensuring sensitive data, such as medical images or financial records, remain safe from attackers. This method improves computational efficiency, allowing for less noise addition, which enhances accuracy. The framework includes a formal template that can be used to privatize virtually any algorithm without needing access to its inner workings. The researchers demonstrated that more stable algorithms, whose predictions remain consistent even when training data is slightly modified, are easier to privatize using their method. This advancement is expected to make the technique easier to deploy in real-world situations, balancing robustness, privacy, and high performance in AI models.[116]

Advancing Fairness in AI: GroupCART Optimizes Decision Tree Models



In a significant stride towards ethical AI, researchers have developed GroupCART, an innovative tree-based ensemble optimizer designed to enhance fairness in decision tree models. This cutting-edge approach not only optimizes for performance but also mitigates bias by considering the entropy increase in protected attributes during model generation. By avoiding bias from the outset, GroupCART offers a flexible trade-off between accuracy and fairness, making it a powerful tool for discrimination-aware classification. The model's ability to train fairer decision trees without transforming data marks a notable advancement in ensuring equitable AI systems, addressing concerns about algorithmic bias in machine learning applications.[117]

Explainable AI in Usable Privacy and Security: Challenges and Opportunities

In a compelling exploration of the intersection between explainable AI and privacy, researchers Vincent Freiberger, Arthur Fleig, and Erik Buchmann have identified significant challenges and opportunities in using Large Language Models (LLMs) for privacy and security evaluations. Their study, centered on the PRISMe tool, highlights concerns about the quality, consistency, and potential hallucinations in LLM-generated explanations, which are crucial in high-stakes contexts where user trust is paramount. The researchers propose strategies such as structured evaluation criteria, uncertainty estimation, and retrieval-augmented generation (RAG) to address these issues. They also emphasize the need for adaptive explanation strategies tailored to different user profiles, aiming to enhance the transparency and reliability of LLM judgments in privacy policy assessments. This work underscores the importance of human-centered explainable AI (HCXAI) in making significant impacts on usable privacy and security.[118]

Unveiling the Hidden Threats and Defenses in AI-Powered Recommender Systems

In a groundbreaking study, researchers Liangbo Ning, Wenqi Fan, and Qing Li have uncovered significant vulnerabilities in Large Language Model (LLM)-based recommender systems, revealing how easily they can be compromised by backdoor attacks. Their innovative attack framework, BadRec, demonstrates that by subtly altering item titles and fabricating user interactions, attackers can implant backdoors with minimal effort, affecting the system's recommendations. To combat this, the researchers developed Poison Scanner (P-Scanner), a robust defense mechanism that utilizes LLMs' advanced language understanding to detect and neutralize these threats. P-Scanner's unique trigger augmentation agent generates synthetic triggers, enhancing its ability to identify and mitigate poisoned items across various domains. Extensive testing on real-world datasets has shown P-Scanner's effectiveness in safeguarding recommender systems from such sophisticated attacks, marking a significant advancement in AI security.[119]

Benchmarking Practices in LLM-driven Offensive Security: Testbeds, Metrics and Experiment Design

In the rapidly evolving field of cybersecurity, researchers Andreas Happe and Jürgen Cito have conducted an in-depth analysis of benchmarking practices for Large Language Model (LLM)-driven offensive security tools. Their study scrutinizes 16 research papers, encompassing 15 prototypes and their respective testbeds, to evaluate the effectiveness of LLMs in tasks like penetration testing and vulnerability exploitation. The researchers emphasize the need for extending current testbeds, establishing robust baselines,

[116] https://news.mit.edu/2025/new-method-efficiently-safeguards-sensitive-ai-training-data-0411

[117] https://arxiv.org/html/2504.12587v1

[118] https://arxiv.org/html/2504.12931v1

[119] https://arxiv.org/html/2504.11182

and incorporating both quantitative and qualitative metrics to ensure accurate assessments. They also highlight the gap between theoretical research and practical applications, noting that Capture The Flag (CTF) challenges may not fully replicate real-world scenarios. Their findings offer valuable recommendations for future research, aiming to enhance the reliability and applicability of LLM-driven offensive security measures.[120]

RealSafe-R1: Enhancing AI Safety Without Compromising Performance

In a significant advancement for AI safety, researchers have introduced RealSafe-R1, a safety-aligned variant of the DeepSeek-R1 model. This innovative approach addresses the safety concerns associated with Large Reasoning Models (LRMs), which often comply with malicious queries due to insufficient alignment. RealSafe-R1 is trained on a dataset of 15,000 safety-aware reasoning trajectories, ensuring robust guardrails against harmful queries and jailbreak attacks. Unlike previous safety alignment efforts that compromise reasoning performance, RealSafe-R1 maintains high reasoning capabilities by preserving the original distribution of training data. This breakthrough promises enhanced safety without sacrificing the utility of powerful reasoning models.[121]

GraphAttack: Unveiling Vulnerabilities in AI Safety Mechanisms

In a groundbreaking study, researchers have introduced "GraphAttack," a sophisticated method that exploits representational blindspots in the safety mechanisms of large language models (LLMs). This innovative approach uses graph-based techniques to generate jailbreak prompts through semantic transformations, effectively bypassing existing safety filters. By representing malicious prompts as nodes in a graph and using Abstract Meaning Representation (AMR) and Resource Description Framework (RDF) to parse and manipulate user goals, the study reveals a significant vulnerability in current AI safety protocols. The researchers demonstrated that instructing LLMs to generate code based on these semantic graphs could achieve an alarming success rate of up to 87% against top commercial LLMs. This discovery underscores the need for more robust and comprehensive safety measures to protect against such structured semantic attacks.[122]

[120] https://arxiv.org/html/2504.10112

[121] https://arxiv.org/html/2504.10081

[122] https://arxiv.org/pdf/2504.13052

Codacy Introduces Guardrails to Enhance Security and Compliance in AI-Generated Code

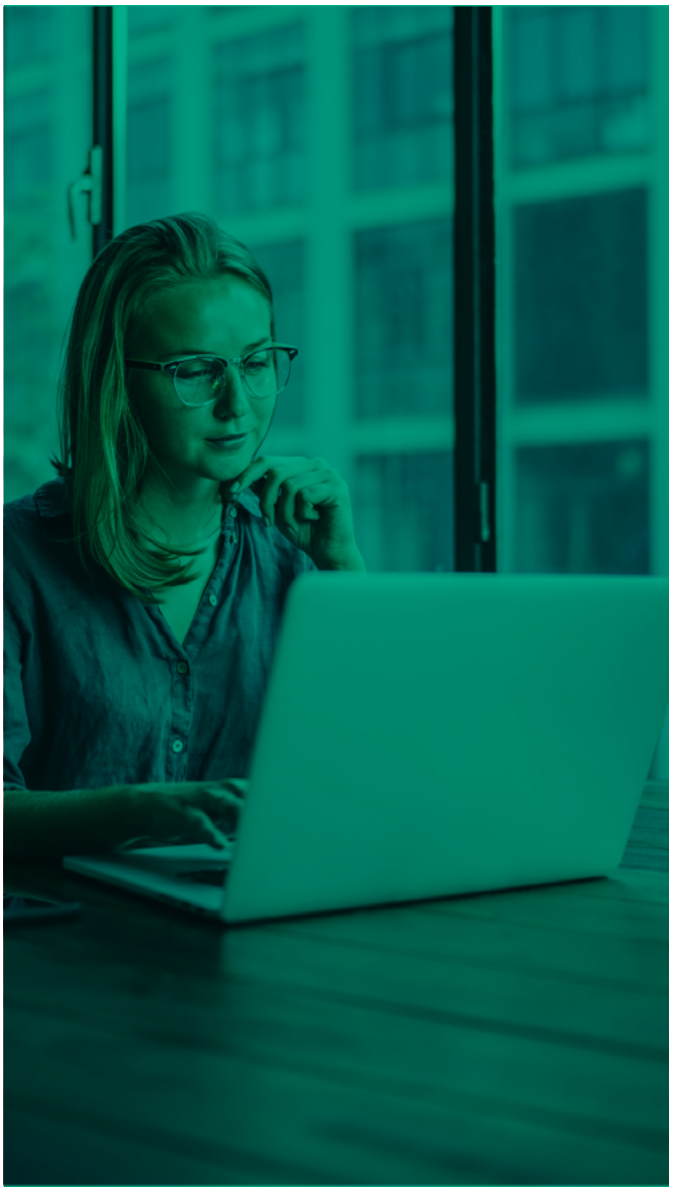
Codacy has launched Codacy Guardrails, a groundbreaking product aimed at ensuring the security, compliance, and quality of AI-generated code from the outset. As AI coding assistants become increasingly integral to software development, the challenge of maintaining trust in rapidly generated code has grown. Codacy Guardrails addresses this issue by integrating with AI coding tools such as Cursor, Windsurf, and GitHub Copilot, enforcing coding standards and preventing the production of non-compliant code. Built on Codacy's SOC2-compliant platform, Guardrails enables development teams to define and apply secure development policies across all AI-generated prompts, ensuring that code is secure and maintainable from the start. This innovation is set to mitigate the risks of security breaches and technical debt, establishing a new benchmark for AI-assisted software development.[123]

Enhancing Object Re-Identification with Global-Local Vision Transformers

In their recent study, researchers Yingquan Wang, Pingping Zhang, Dong Wang, and Huchuan Lu explore the integration of global and local features in Vision Transformers (ViT) to improve object re-identification (Re-ID). The paper introduces a novel Global-Local Transformer (GLTrans) that leverages both global aggregation and local multi-layer fusion to enhance the representational capabilities of ViT. By utilizing a Global Aggregation Encoder (GAE) to capture comprehensive global features and a Local Multi-layer Fusion (LMF) mechanism to refine local representations, the proposed method achieves superior performance across multiple Re-ID benchmarks. This approach demonstrates the mutual enhancement of global and local information, offering a significant advancement in the field of computer vision.[124]

aiXamine: A Comprehensive Platform for Evaluating LLM Safety and Security

In their recent paper, researchers introduce aiXamine, a robust black-box evaluation platform designed to assess the safety and security of large language models (LLMs). The platform integrates over 40 tests targeting key dimensions such as adversarial robustness, code security, fairness and bias, hallucination, model and data privacy, out-of-distribution robustness, over-refusal, and safety alignment. aiXamine aggregates these evaluation results into detailed reports, providing a comprehensive breakdown of model performance, test examples, and rich visualizations. The study reveals notable vulnerabilities in leading models, including susceptibility to adversarial attacks, biased outputs, and privacy weaknesses. Additionally, the findings highlight that open-source models can match or exceed proprietary models in specific safety and security aspects. This work underscores the critical need for thorough evaluation of LLMs to ensure their safe and ethical deployment in various applications.[125]



[123] https://aithority.com/ai-machine-learning-projects/codacy-launches-guardrails-to-secure-ai-generated-code-from-the-start/

[124] https://arxiv.org/abs/2504.14985

[125] https://arxiv.org/html/2504.14985v1

Industry Update

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

HealthCare

NextGen Healthcare Introduces AI-Driven Innovations to Enhance Patient-Provider Interactions

NextGen Healthcare has unveiled its latest AI-driven advancements aimed at transforming the patient-provider experience. The new release of NextGen® Mobile integrates mobility, voice-enablement, AI, and automation to streamline the management of diagnosis codes, orders, and prescriptions within a single workflow. Key features include an AI-powered "patient story" tool that generates concise summaries of patients' medical information, an integrated medications workflow with AI-generated suggestions, and charge capture integration for seamless addition of office visit and procedure codes. These enhancements are designed to maximize efficiency, protect against provider burnout, and improve clinical and financial outcomes. NextGen Ambient Assist, an AI-driven ambient listening solution, continues to save providers up to two hours of documentation time per day by transcribing patient-provider conversations in real time and summarizing encounters within seconds. This solution supports Spanish language, specialty-specific models, and intelligent suggestions for diagnosis codes and lab orders.[126]

NVIDIA and GE HealthCare Partner to Revolutionize Diagnostic Imaging with Physical AI

NVIDIA and GE HealthCare have announced a collaboration to advance autonomous diagnostic imaging through the development of Physical AI. This partnership leverages the new NVIDIA Isaac for Healthcare platform, which includes pretrained models and physics-based simulations of sensors, anatomy, and environments. The initiative aims to enhance X-ray and ultrasound technologies by automating complex workflows such as patient placement, image scanning, and quality checking. By integrating robotic capabilities into these imaging systems, the collaboration seeks to expand access to diagnostic care globally, addressing the challenges of growing workloads and staffing shortages in healthcare. This effort builds on nearly two decades of joint innovation between the two companies, focusing on improving patient care through advanced AI and imaging technologies.[127]

Health Tech Investment Act Introduced, Covering AI-Enabled Medical Devices

On April 10, 2025, US Senators Mike Rounds (R-S.D.) and Martin Heinrich (D-N.M.), co-chairs of the Senate Artificial Intelligence Caucus, introduced draft legislation titled the Health Tech Investment Act. This bill aims to improve health outcomes for Medicare patients by encouraging the use of cutting-edge, AI-enabled medical devices. It proposes assigning all FDA-approved AI-enabled medical devices to a New Technology Ambulatory Payment Classification (APC) in the Hospital Outpatient Prospective Payment System (OPPS) for a minimum of five years. This period will allow for the collection of adequate data regarding delivery and service costs before assigning a permanent payment code. Specifically, the bill seeks to develop a formalized payment pathway for FDA-cleared medical devices, provide patients with access to innovative AI-enabled clinical technology, offer manufacturers and providers the certainty needed to invest in next-generation healthcare technologies, and improve patient outcomes by providing resources for providers to meet ABHS standards of care.[128]

AI and Major Changes in HIPAA Compliance for 2025: Stricter Enforcement and New Guidelines

Recent legal developments are set to bring significant changes to HIPAA (Health Insurance Portability and Accountability Act) compliance in 2025. These changes include stricter enforcement of data privacy and security measures, increased penalties for non-compliance, and new guidelines for handling health information, particularly with AI technologies. The updates aim to address the evolving landscape of digital health data and ensure that patient information is adequately protected. Healthcare providers, insurers, and other entities handling health data will

[126] https://www.businesswire.com/news/home/20250303729069/en/NextGen-Healthcare-Unveils-Latest-AI-Driven-Advancements-to-Transform-the-Patient-Provider-Experience

[127] https://nvidianews.nvidia.com/news/nvidia-and-ge-healthcare-collaborate-to-advance-the-development-of-autonomous-diagnostic-imaging-with-physical-ai

[128] https://www.rounds.senate.gov/newsroom/press-releases/rounds-introduces-legislation-to-expedite-use-of-ai-medical-devices-for-medicare-patients?utm_source=substack&utm_medium=email

need to adapt to these new requirements to avoid hefty fines and legal repercussions. The changes also emphasize the importance of robust cybersecurity practices and regular audits to maintain compliance and safeguard sensitive health information.[129]

GenomOncology Unveils BioMCP: A Revolutionary Open-Source Protocol for Bio-medical AI Assistants

GenomOncology has announced the launch of BioMCP, an open-source Model Context Protocol (MCP) designed to enhance the capabilities of biomedical AI assistants and agents. This innovative technology allows AI systems to access specialized medical information, including clinical trials, genetic data, and published medical research, through a standardized interface. By building on the MCP standard created by Anthropic, BioMCP enables AI systems to perform advanced searches, retrieve full-text data, and refine their approach based on context, making complex medical information more accessible. The protocol supports seamless integration with various medical databases, ensuring AI systems can utilize the latest research to provide accurate and relevant insights. GenomOncology is also developing a commercial version of BioMCP, which will offer enhanced security, on-site deployment, and integration with clinical and research data, further expanding its utility in precision oncology.[130]

Breakthrough AI Model Achieves Near-Perfect Accuracy in Endometrial Cancer Diagnosis

Researchers from Charles Darwin University, Daffodil International University, the University of Calgary, and Australian Catholic University have developed an advanced AI model, ECgMLP, which can diagnose endometrial cancer with an impressive 99.26% accuracy. This model analyzes histopathological images, enhancing image quality and identifying critical areas for accurate diagnosis. The ECgMLP model significantly outperforms existing methods, which have an accuracy range of approximately 78.91% to 80.93%. The AI model's robustness and clinical applicability extend beyond endometrial cancer, showing high accuracy in diagnosing colorectal, breast, and oral cancers as well. This breakthrough promises to enhance clinical processes and improve patient outcomes by providing fast and accurate early detection of various cancers.[131]

Chinese Scientists Develop AI-Powered Wearable to Aid Visually Impaired

Chinese scientists have developed an innovative AI-powered wearable device designed to assist visually impaired individuals in navigating their surroundings more effectively. The device uses advanced artificial intelligence to interpret visual data and provide real-time audio feedback to the user, helping them identify obstacles and navigate safely. This breakthrough technology aims to enhance the independence and quality of life for visually impaired people. The development of this wearable is a significant step forward in the field of assistive technology, showcasing the potential of AI to create impactful solutions for those with disabilities.[132]

Information Technology

Darktrace Unveils New AI Models in Cyber AI Analyst: Enhancing Proactive Security

Darktrace has introduced new AI models within its Cyber AI Analyst to bolster proactive security measures. These models include Darktrace Incident Graph Evaluation for Security Threats (DIGEST) and Darktrace's Embedding Model for Investigation of Security Threats (DEMIST-2). DIGEST uses graph neural networks (GNNs) and recurrent neural networks (RNNs) to prioritize security incidents. DEMIST-2, a language model, analyzes security patterns and converts sparse data into dense, contextual information. This enhancement aims to reduce workload and alert fatigue for security teams, enabling them to focus on critical priorities.[133]

Education

Claude for Education: Enhancing Critical Thinking Through AI

Anthropic has launched Claude for Education, an AI assistant tailored to enhance critical thinking in students through Socratic questioning techniques. This initiative, termed Learning Mode, prompts students to engage in deeper reasoning by asking questions like "How would you approach this problem?" and "What evidence supports your conclusion". By partnering with prestigious institutions such as Northeastern University, London School of Economics, and Champlain College, Anthropic aims to redefine AI's role in education. This approach transforms Claude into a digital tutor, guiding students through the learning process and mitigating the risk of AI tools promoting shortcut thinking.[134]

[129] https://www.reuters.com/legal/litigation/new-legal-developments-herald-big-changes-hipaa-compliance-2025-2025-04-07/

[130] https://www.prnewswire.com/news-releases/genomoncology-announces-biomcp-open-source-model-context-protocol-mcp-for-biomedical-ai-assistants-and-agents-302425734.html

[131] https://medicalxpress.com/news/2025-03-ai-endometrial-cancer-accuracy.html

[132] https://www.msn.com/en-in/money/topstories/chinese-scientists-develop-ai-powered-wearable-to-help-visually-impaired-people-see/ar-AA1Daj4e?ocid=BingNewsSerp

[133] https://www.securityinfowatch.com/ai/press-release/55283596/darktrace-unveils-new-ai-models-in-cyber-ai-analyst-to-enhance-proactive-security html

[134] https://venturebeat.com/ai/anthropic-flips-the-script-on-ai-in-education-claude-learning-mode-makes-students-do-the-thinking/

Finance

Microsoft AI Boardroom 2025: New Tools and Initiatives for India's BFSI Sector

At the AI Boardroom 2025 event in Mumbai, Microsoft introduced several new initiatives and tools aimed at transforming India's banking, financial services, and insurance (BFSI) sector. The launch of the CoreAI platform and tools, which integrate AI capabilities into platforms like Azure AI Foundry, GitHub, and VS Code, was a key highlight. The event also showcased the significant impact of generative AI on enhancing productivity, customer service, and operational efficiency within the BFSI sector. Additionally, Microsoft's AI action plan, focusing on regulatory simplification, computing infrastructure, data, industrial uptake, and skills, aims to make India's BFSI sector AI-first, empowering financial institutions to operate with agility, scale, and trust. These initiatives underscore Microsoft's commitment to supporting AI transformation across various sectors while ensuring security and innovation.[135]

AI-Powered Zrai Shield Transforms Risk Management in Fintech with Real-Time Intelligence

Zrai Shield, the world's first real-time AI-driven Financial Risk Management (FRM) engine, is set to revolutionize the fintech industry by leveraging advanced artificial intelligence to identify and mitigate risks instantly. Unlike traditional risk management systems, which rely on static models and past data, Zrai Shield uses machine learning to continuously analyze real-time transaction data, spotting fraudulent activities and vulnerabilities as they arise. This AI-powered approach not only enhances the accuracy of risk assessments but also ensures faster, more efficient responses to emerging threats. By adopting Zrai Shield, fintech companies can strengthen their security infrastructure, reduce fraud, and gain a significant competitive edge in managing financial risk.[136]

Transportation Safety

Smart Roads to the Rescue: New AI-Based System Aims to Prevent Roadkill

The South Korean Ministry of Environment has launched a pilot program for its Animal Road Accident (Roadkill) Prevention System, utilizing advanced artificial intelligence (AI) technology to detect wildlife on roadways. The system, currently tested on roads in Yangpyeong, Gyeonggi Province, and Pyeongchang, Gangwon Province, employs AI-enabled CCTV cameras and LiDAR sensors to monitor for animals. When wildlife is detected, an LED board displays a warning message to drivers, helping them reduce speed and take precautionary measures. The project involves collaboration with Posco DX, the Korea National Park Service, and the National Institute of Ecology. The ministry plans to expand the system to additional high-risk areas by 2027, aiming to enhance driver safety and biodiversity conservation through timely alerts and predictive capabilities.[137]

Infrastructure Development

Warning Issued Over Use of AI for Compliance with Lifting Regulations

Construction companies have been cautioned against relying on AI-generated guidance to comply with the Lifting Operations and Lifting Equipment Regulations (LOLER) and the Provision and Use of Work Equipment Regulations (PUWER). The accrediting body, Consolidated Fork Truck Services (CFTS), highlighted that AI tools like ChatGPT, Gemini, and Copilot might provide misleading information, potentially simplifying complex legal requirements and failing to clarify different inspection needs for various equipment. CFTS emphasized the importance of adhering to official Health and Safety Executive or UK Material Handling Association guidance and consulting CFTS-accredited examiners to ensure compliance and safety.[138]

China Deploys AI to Monitor Non-Coal Mines in First-of-its-Kind Safety Initiative

China has deployed artificial intelligence (AI) to monitor non-coal mines, marking a significant advancement in mining safety. The system, introduced on March 12, demonstrated its effectiveness by detecting an emergency 580 meters underground and triggering alarms. Developed by Hubei Lonmon Phosphorus Chemical Co., the AI system integrates the DeepSeek model, which offers advanced semantic understanding and multimodal processing capabilities. Engineers refined AI algorithms using extensive data from mining accident reports, safety regulations, and monitoring datasets. Despite initial setbacks with foreign AI models, the DeepSeek model provided a breakthrough, enabling the system to autonomously detect hazards, initiate corrective actions, and coordinate with other systems to enhance efficiency and safety.[139]

[135] https://news.microsoft.com/en-in/microsoft-ai-boardroom-2025-ushering-in-the-new-era-of-banking-financial-services-and-insurance-bfsi-with-ai/

[136] https://www.timesnownews.com/bizz-impact/zrai-shield-the-worlds-first-real-time-ai-frm-engine-is-here-and-its-changing-how-fintech-thinks-about-risk-article-151465183

[137] https://m.koreaherald.com/article/10462431?sec=002

[138] https://constructionmanagement.co.uk/warning-issued-over-use-of-ai-to-comply-with-lifting-regs/

[139] https://www.aa.com.tr/en/asia-pacific/china-deploys-ai-to-monitor-non-coal-mines-in-1st-to-bolster-safety/3527494

Automotive

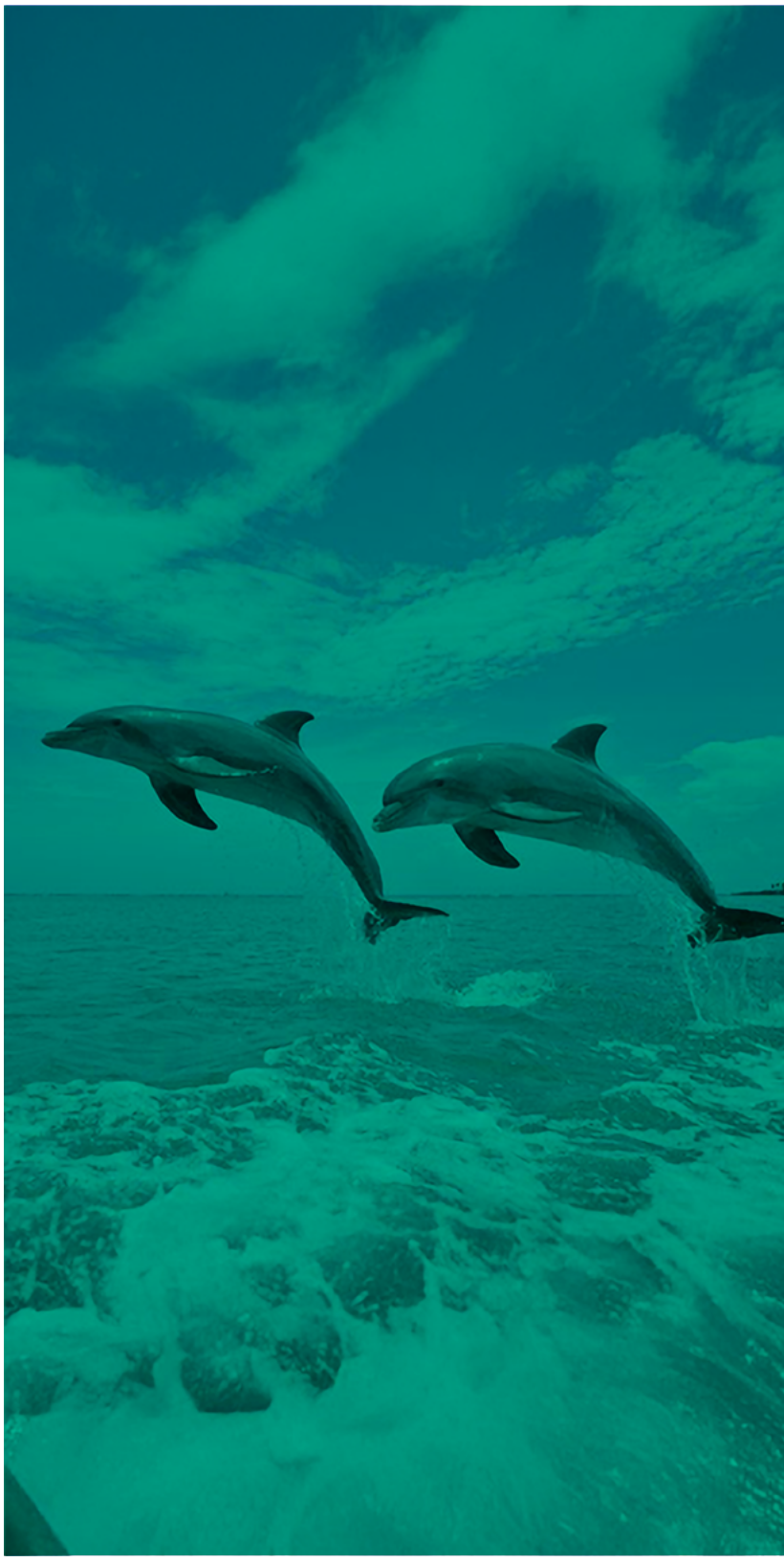
Enhancing Autonomous Vehicle Safety with Advanced AI Techniques

In a groundbreaking study, researchers have developed an innovative method to improve the derivation of safety requirements for autonomous vehicles using agent-based retrieval-augmented generation (RAG). This approach addresses the challenges faced by traditional RAG methods, which often struggle with complex queries and domain-specific knowledge. By leveraging an agent-based system on a comprehensive document pool of automotive standards and the Apollo case study, the researchers achieved significant improvements in the relevance and accuracy of retrieved information. This advancement not only enhances the efficiency of safety analysis processes but also ensures that large language models (LLMs) generate reliable and explainable results. Consequently, this method supports safety engineers in deriving, aligning, and reviewing safety requirements, thereby contributing to the development of safer self-driving vehicles.[140]

Environmental Monitoring

DolphinGemma: Google's AI Breakthrough in Interspecies Communication

Google's latest AI innovation, DolphinGemma, is a groundbreaking model designed to decode and communicate with dolphins using their complex vocalizations, such as clicks, whistles, and burst pulses. Developed in collaboration with the Wild Dolphin Project, DolphinGemma leverages 40 years of labeled dolphin vocalization data, linking sounds to specific behaviors and contexts. Built on Google's Gemini architecture, the AI employs SoundStream, a neural audio codec, to analyze and predict sonic patterns, much like how language models process human text. This pioneering technology could revolutionize our understanding of animal intelligence and interspecies communication.[141]



[140] https://arxiv.org/html/2504.11243v1

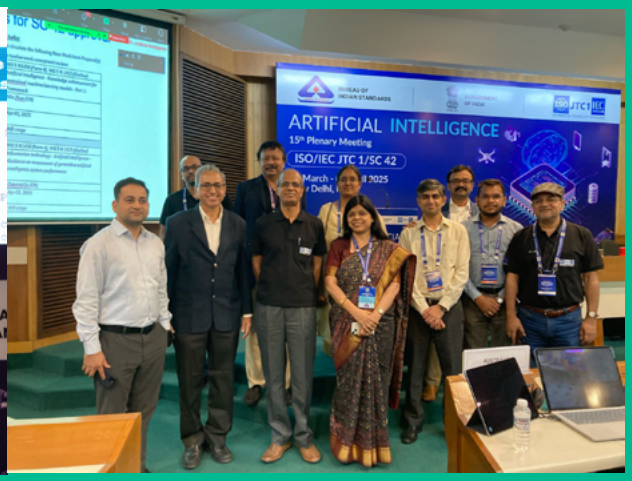
[141] https://yourstory.com/2025/04/googles-new-ai-talk-dolphins

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

15th ISO Plenary Meeting | March 31 - April 4 | New Delhi, India



The 15th ISO Plenary meeting took place at the India Habitat Centre in New Delhi, India, from March 31 to April 4, 2025. This global event, attended by delegates from over 25 countries, focused on developing international standards for responsible, safe, and ethical AI. Srinivasan Sivasubramanian, from Infosys' Responsible AI Office, represented Infosys and was part of the Indian delegation attending this event. Srini also participated at a workshop organized by Bureau of Indian Standards titled "Enabling trust in Technology in the age of LLMs and Generative AI". As part of a session on Responsible AI, he presented the Infosys Responsible AI Office, showcasing the responsible AI toolkit and the journey of Infosys becoming the first organization to be certified on ISO 42001.

Infosys and Linux Foundation Partner to Promote Ethical AI Practices Globally

Infosys and the Linux Foundation have entered into a strategic partnership to advance responsible AI practices on a global scale. This collaboration aims to harness the combined strengths of Infosys' AI expertise and the Linux Foundation's leadership in open-source initiatives. Together, they will develop and implement ethical AI frameworks that emphasize transparency, fairness, and accountability. This initiative is designed to foster innovation while addressing the ethical challenges associated with AI deployment, demonstrating a shared commitment to leveraging AI for the greater good of society.

India Global Forum - IGF:NXT25| April 8 | Mumbai

Ashish Tewari, Head of Responsible AI Office (India) & Anjali Patel , Senior Associate Consultant , Responsible AI office attended the **IGF**: NXT 25: Leading the Leap organized by India Global Forum held on April 8th at the Jio World Convention Centre, Mumbai. **Ashish Tewari** participated in a closed-door roundtable on **'Prompt Engineers and Beyond: Generative AI Skills in Demand'** Chaired by Dr. Abhilasha Gaur, CEO of the Sector Skills Council NASSCOM, the session convened senior leaders across industry and startups to explore the evolving skillsets required in the generative AI space. Ashish contributed perspectives on the importance of embedding Responsible AI principles into skilling programs to ensure the development of ethically grounded and future-ready AI talent. He also highlighted the emergence of new roles such as **AI Compliance Officer** and **AI Security Officer**, emphasizing the need to equip professionals with the right skills and frameworks to navigate evolving regulatory, ethical, and safety challenges in AI adoption.

Centre for Responsible AI (CeRAI) Roundtable on Voluntary AI Self-Regulation in India | April 15 | New Delhi

Ashish Tewari, Head of the Responsible AI Office, India, attended a roundtable hosted by the Centre for Responsible AI (CeRAI), IIT Madras, on April 15 in New Delhi. The discussion centered on a new paper by Amlan Mohanty titled "Making AI Self-Regulation Work: Perspectives from India on Voluntary AI Risk Mitigation", which explores the role of voluntary commitments in fostering trust and transparency in India's AI ecosystem. The roundtable brought together industry leaders, policymakers, academic experts, and government officials to discuss the features and institutional models of AI self-regulation, current stakeholder sentiment in India, global benchmarks and India-specific frameworks, and recommendations for effective, scalable implementation.

Latest News

Infosys Launches AI Innovation Labs and Factories to Drive Emerging AI Technologies

Infosys has established AI innovation labs and AI factories in collaboration with leading clients to incubate and scale emerging AI technologies. These labs focus on tracking, assessing, and developing proof of value for new AI technologies within the organization, while the AI factories aim to productize and scale these solutions across various business lines. Rafee Tarafdar, Chief Technology Officer at Infosys, highlighted the importance of data readiness and a structured approach to enhance existing data architectures, ensuring the successful integration of AI initiatives. This strategic move underscores Infosys' commitment to leveraging AI for re-engineering core business processes and meeting client demands for advanced AI capabilities.[142]

Infosys Responsible AI Toolkit Now Featured in OECD.AI Catalogue

The Infosys Responsible AI Toolkit has been successfully updated in the OECD.AI Catalogue. This comprehensive suite is designed to assist developers in creating trustworthy AI systems by integrating essential aspects such as safety, security, privacy, explainability, fairness, and hallucination detection into AI solutions. The toolkit features a user-friendly interface with multiple tabs for different data inputs and outputs, making it accessible to both technical and non-technical users. Key features include support for generative AI models, machine learning models, and tools that ensure model transparency and text quality, thereby promoting the development of reliable and ethical AI applications.[143]

[142] https://www.moneycontrol.com/technology/established-ai-innovation-labs-factory-with-clients-to-incubate-new-ai-tech-infosys-article-12980113.html

[143] https://oecd.ai/en/catalogue/tools/infosys-responsible-ai-toolkit

Infosys Responsible AI Toolkit – A Foundation for Ethical AI

The Infosys Responsible AI Toolkit is now open sourced and can be accessed from its public GitHub repo.[144]

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 - Data Curator for multiple file curation

RAI Data Curator (Responsible AI Data Curator) has been developed recently and will soon be added in the open source Responsible AI toolkit to help curate multiple files uploaded simultaneously. It anonymizes PII, censors profane words and supports multiple file uploads in .pdf or .csv format.

The explainability evaluation metrics in the toolkit have recently been integrated with Infosys' lifesciences - related application "Scientific Writing Platform," which automates the medical writing process using Generative AI with human-in-the-loop reviews. This platform ingests documents from various sources, enriches them using NLP and NER-based ontologies, and generates narratives using models trained on specific document sections. The explainability evaluation metrics such as uncertainty and coherence are utilized to evaluate the explainability of this application.

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

Contributors

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

Ashish Tewari - Head of Infosys Responsible AI Office, India

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

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

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

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

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

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

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

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

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

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