MARKET SCAN REPORT

APRIL 2026

BY INFOSYS TOPAZ RESPONSIBLE AI OFFICE

Infosys topaz

IN FOCUS

THE CYBERSECURITY SINGULARITY: AN ENTERPRISE ASSESSMENT OF MYTHOS-CLASS AI

Perspective from the Infosys Responsible AI Team

Infosys®
Navigate your next

Foreword

Some months confirm the trend. This month **broke the ceiling.**

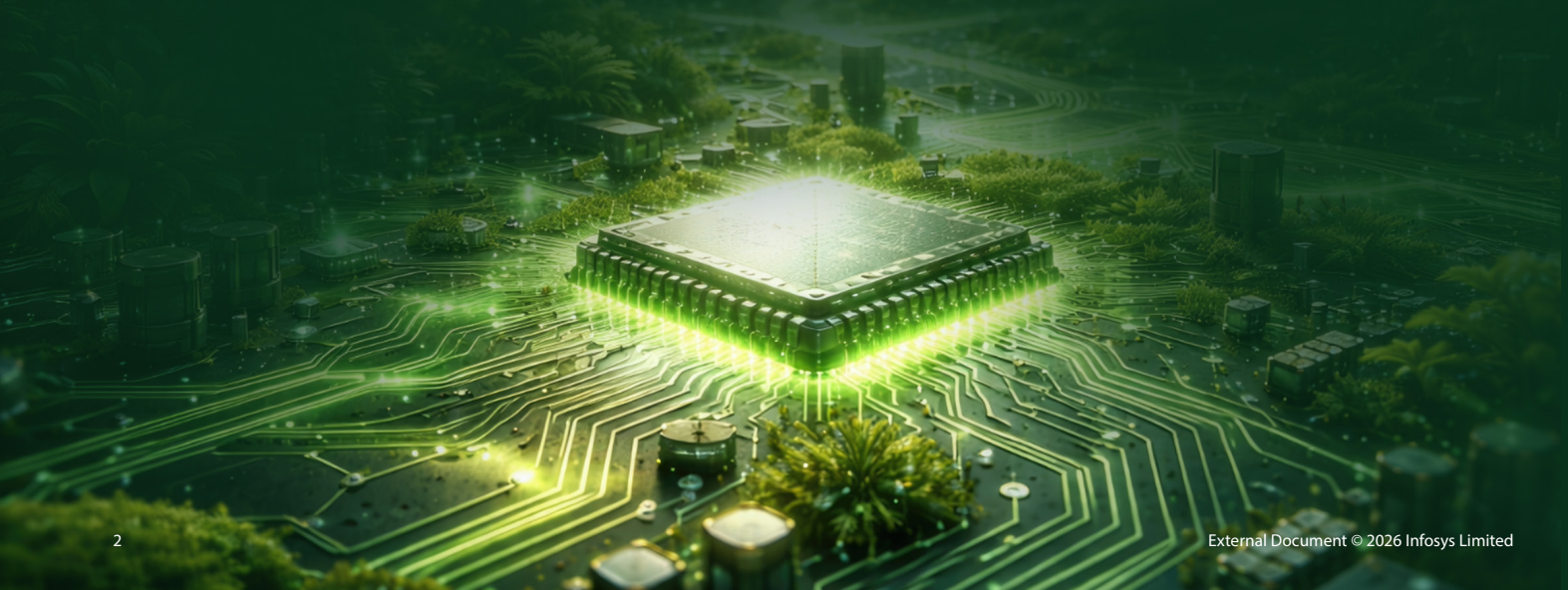
On April 7, Anthropic announced Claude Mythos Preview through Project Glasswing and deliberately chose not to make the model generally available, citing the very real dual use risk of a system built to find and exploit software weaknesses at scale. In its own reporting and third-party coverage, Mythos Preview is described as identifying thousands of high-severity issues across major operating systems and browsers, including long-standing flaws such as a decades-old OpenBSD weakness that had previously gone undetected. The message for every enterprise leader is now unambiguous. AI capability has outpaced legacy security rhythms, and the cost of delay is rising.

That is why this month's In Focus, **"The Cybersecurity Singularity: An Enterprise Assessment of Mythos Class AI,"** from the Infosys Responsible AI team, anchors this April Market Scan. It takes a clear position on what Mythos-class capability means in practice. The old cycle of discover, disclose, patch is breaking under AI-accelerated conditions. What replaces it cannot be cosmetic. It must be governed, controlled, and accountable by design, with earlier discovery, disciplined release and vendor controls, and defensive automation that is measurable and bounded. We are going to witness automated vulnerability scanning, remediation and monitoring in the form of VulnOps management and capabilities maturing in the form, processes, tooling and expertise.

Beyond Mythos, the governance baseline is hardening globally, and it is moving from guidance to expectation. In the United States, the Parents Decide Act proposes age verification at operating system setup so parents can set controls that flow across apps and platforms. In the UK, the Freedom from Violence and Abuse action plan strengthens the push for safety by design and addresses technology-enabled abuse, including AI-driven deepfakes and related harms, across the 2026 to 2029 period. In Australia, the Federal Court's Generative AI Practice Note sets a clear standard for high-impact use, requiring responsible use, disclosure where needed, and verification so false or fictitious material is not presented to the Court. These are not isolated moves. They are the operating standard that AI at scale will be held to.

The next chapter of AI will not reward speed for its own sake. It will reward those who govern first, secure what matters, and prove trust through execution. The window to get ahead is real. It is also closing.

Syed Ahmed
Global Head
Infosys Responsible AI Office



From the editor's desk

From Capability to Care: What This Moment Is Asking of AI

"Responsible AI is not what we announce. It is what we enforce when speed is tempting."

April sharpened the line between progress and exposure. The Mythos moment was not about hype or harm. It was about what powerful systems reveal. When capability rises, small gaps in process, third-party controls, and release discipline stop being minor and start becoming material.

Regulation is beginning to reflect this reality in practical, usable terms. In the United States, the AI at Work Guide from the Department of Commerce anchors responsible use in everyday behavior: human judgment stays in charge, sensitive data stays protected, and accountability stays with people. In Australia, the Federal Court practice note allows tools such as ChatGPT, Google Gemini, Claude, and Microsoft Copilot, but sets a clear bar: disclose use, verify facts and references, and never risk confidential information. In China, new trial AI ethics governance guidelines push the same direction at system level, ethics review, fairness, safety, and AI that remains controllable and trustworthy.

The risk signals underline why these guardrails matter. A critical vulnerability in Anthropic's Model Context Protocol (MCP) shows how agent ecosystems can widen supply-chain risk and enable arbitrary command execution when security controls lag behind adoption.

Innovation is accelerating at the same time, and it is lowering barriers. Google's Veo 3.1 Lite brings low-cost, high-speed video generation to developers at scale. Google DeepMind's Gemini Robotics-ER 1.6 pushes embodied reasoning forward, including improved industrial instrument reading for real-world deployments. As synthetic media and autonomy become easier to deploy, safety cannot be optional or delayed.

This month's In Focus explores one such moment through **"The Cybersecurity Singularity: An Enterprise Assessment of Mythos Class AI,"** a perspective from our Infosys Responsible AI team. It highlights how Mythos-class systems compress the gap between vulnerability discovery and exploitation, raising the bar for enterprise readiness.

The message is simple. Scale what you can govern, and govern it before you scale.

Warm regards,

Ashish Tewari
Head - Infosys Responsible AI office
(APAC, Middle East & India)

Table of Contents

AI Regulations, Governance & Standards

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

US

New US Initiative Brings Artificial Intelligence into Apprenticeships

A new national initiative called "Advancing Artificial Intelligence Skills in Registered Apprenticeships" has been launched by the US Department of Labor to help prepare workers for future jobs. The initiative focuses on adding artificial intelligence skills to existing apprenticeship programs while also creating new pathways in AI-related roles. It aims to modernise training across industries such as data centers, telecommunications and advanced manufacturing by combining hands-on learning with AI education. Employers will receive support to adopt these updated programs and workers will gain practical, job-ready AI skills while earning income. The initiative reflects a long-term effort to strengthen the workforce, support economic growth and ensure American workers can succeed in a fast-changing, technology-driven economy.[1]

Strengthening Online Safety for Children by Giving Parents Clear and Direct Control Over Digital Content

The Parents Decide Act is a bipartisan law announced in the United States to better protect children online by giving parents full control over what their children can see and use on digital devices. The proposal focuses on verifying a child's age when a phone or tablet is first set up, instead of allowing children to enter their own details. This allows parents to apply

[1] https://www.dol.gov/newsroom/releases/eta/eta20260401

age-appropriate limits from the start, including controls over social media, mobile apps and Artificial Intelligence platforms. These settings are designed to carry safely across all apps and services, helping block harmful or unsuitable content, including risky Artificial Intelligence chatbot interactions. Overall, the act aims to close safety gaps, reduce online harm and ensure parents-not technology companies-make decisions that protect children's wellbeing.[2]

AI at Work Guide: How the U.S. Government Explains Safe and Responsible Use of AI in the Workplace

The AI at Work Guide from the U.S. Department of Commerce is a practical guidance document designed to help organizations, managers and employees understand how artificial intelligence can be used responsibly at work. It explains that artificial intelligence is already helping people write content, analyze information, automate routine tasks and improve productivity, but it also brings risks if used carelessly. The guide emphasizes the importance of human judgment, transparency and accountability when using artificial intelligence tools. It encourages workers to understand what artificial intelligence can and cannot do, protect sensitive data, avoid over-reliance on automated outputs and watch for errors, bias, or misleading results. The document also stresses the need for clear workplace policies, employee training and ethical use so that artificial intelligence supports people rather than replacing responsibility or trust. Overall, the guide is meant to educate, not regulate and serves as a shared reference for using artificial intelligence safely, fairly and effectively in everyday work.[3]

Senator Calls on AI Voice Firms to Stop Scam Misuse

A U.S. Senator has urged major Artificial Intelligence voice-cloning companies to take stronger steps to prevent their tools from being used in scams that trick people into sending money. Senator Maggie Hassan, as part of her work on the Joint Economic Committee, asked four leading companies-ElevenLabs, LOVO, Speechify and VEED-to explain what safeguards they have in place to stop criminals from copying voices without consent. She raised alarm after the Federal Bureau of Investigation reported that Americans lost about 893 million dollars to Artificial Intelligence-related scams in 2025. These scams often involve fake phone calls that sound like family members, friends, celebrities, or officials, making them very convincing and emotionally distressing. The Senator highlighted real cases where people were deceived by such calls and warned that many voice-cloning tools still lack basic protections. She stressed that technology companies must work closely with the government to better protect people from financial and emotional harm caused by Artificial Intelligence-powered fraud.[4]

U.S. Joint Economic Committee Democrats Highlight AI's Economic Impact, Workforce Transformation and Policy Priorities in New Press Release

The U.S. Joint Economic Committee (JEC) Democrats released a press statement in April 2026 emphasizing the growing influence of artificial intelligence on the U.S. economy, workforce dynamics and long-term productivity, while underscoring the need for proactive policy coordination to maximize benefits and mitigate risks. The release highlights how AI is increasingly functioning as a general-purpose technology capable of reshaping industries, enhancing productivity and altering job structures, while also raising concerns related to labor market disruptions, skill gaps and potential inequalities; it stresses the importance of federal investment in workforce development, education and reskilling initiatives to ensure inclusive economic growth. Additionally, the committee calls for a balanced policy approach that supports innovation while addressing risks such as bias, data privacy and economic concentration, positioning AI governance as a critical priority for maintaining U.S. competitiveness and ensuring that technological advancements translate into broad-based economic benefits.[5]

[2] https://gottheimer.house.gov/posts/release-gottheimer-announces-bipartisan-parents-decide-act-to-protect-kids-online
[3] https://www.commerce.gov/sites/default/files/2026-04/AIatWork-Guide.pdf
[4] https://www.jec.senate.gov/public/index.cfm/democrats/2026/4/senator-hassan-presses-leading-ai-voice-cloning-companies-to-prevent-exploitation-by-scammers
[5] https://www.jec.senate.gov/public/index.cfm/democrats/press-releases?ID=564F54F1-CE55-498E-96C9-236D8DEBD529

UK

UK FCA and ICO Guidance on Vulnerability Data

In the United Kingdom, the Financial Conduct Authority and the Information Commissioner's Office have issued a joint statement to guide firms on how to handle data relating to customers in vulnerable circumstances. The guidance sets out expectations for delivering fair and positive outcomes under the Consumer Duty while fully complying with UK data protection laws, including the UK General Data Protection Regulation. It explains how firms should identify and support vulnerable consumers, share relevant information responsibly across the distribution chain and monitor customer outcomes to prevent harm. The statement reinforces that UK data protection law enables responsible data use when firms act lawfully, transparently and with appropriate safeguards.[6]

UK Regulators Examine How to Safely Manage the Rise of Agentic Artificial Intelligence

UK digital regulators have released a forward looking paper that explains what agentic artificial intelligence is and how existing UK rules can help ensure it is used safely and responsibly. The paper, published by the Digital Regulation Cooperation Forum, describes agentic AI as systems that can act on behalf of users by setting goals, planning tasks and carrying out actions with limited human input, unlike standard AI tools that only respond to requests. While this technology could save time for consumers and boost productivity for businesses and regulators, it also raises risks such as mistakes, lack of transparency, misuse of personal data, unfair market dominance and loss of user control. The regulators stress that current laws on data protection, consumer rights, fairness, competition and security already apply to agentic AI and that strong governance, human oversight and coordination between regulators are essential to build trust and support safe innovation.[7]

UK Action Plan on AI Safety and Preventing Violence and Abuse

The UK government has released Freedom from Violence and Abuse: Volume 2 – Action Plan, which sets out clear actions to reduce violence against women and girls and make society safer between 2026 and 2029. The plan focuses on stopping abuse early, protecting children, holding perpetrators

[6] https://www.fca.org.uk/publications/corporate-documents/joint-fca-and-ico-statement-regulatory-expectations-regarding-firms
[7] https://www.drcf.org.uk/siteassets/drcf/pdf-files/drcf-the-future-of-agentic-ai-foresight-paper.pdf

accountable and supporting victims. A key area is the role of Artificial Intelligence, with strong measures to prevent AI from being misused to create non consensual intimate images, deepfakes and other harmful content. The plan strengthens online safety rules, bans nudification tools, criminalises harmful pornography and promotes "safety by design" in technology. It also improves education on consent, healthy relationships, AI risks and online safety, while encouraging a whole of society approach involving government, schools, technology firms, communities and international partners.[8]

UK Regulators Respond to AI Risks in Finance

The Bank of England and the Financial Conduct Authority have agreed to take action on the growing use of artificial intelligence in financial markets after concerns raised by Members of Parliament. The Bank of England said it will test how artificial intelligence trading systems behave, particularly whether they might all act in the same way at the same time, which could destabilise markets. The Financial Conduct Authority said it will provide clearer guidance and examples to help financial firms use artificial intelligence safely while following existing rules. However, MPs criticised HM Treasury for moving too slowly on extending stronger oversight to major technology and cloud providers. Parliament stated it will continue to closely monitor whether these actions are enough to protect the UK's financial stability.[9]

Europe

EU Stops Use of AI Images and Videos in Official Messages

The European Union has decided that its employees must not use artificial intelligence generated photos or videos in official communication. This rule applies to major European Union bodies such as the European Commission, the European Parliament and the Council of the European Union and covers press releases and public materials. The move was made due to growing concerns about deepfakes and misleading visuals, which could reduce public trust in official information. European Union officials said all shared visuals must be real and authentic and even small artificial intelligence based edits like image enhancement are not allowed. The ban reflects the European Union's cautious approach toward artificial intelligence in public communication and aims to protect transparency, credibility and trust, especially during elections and sensitive political situations.[10]

[8] https://www.gov.uk/government/publications/freedom-from-violence-and-abuse-a-cross-government-strategy/freedom-from-violence-and-abuse-volume-2-action-plan-accessible

[9] https://committees.parliament.uk/committee/158/treasury-committee/news/213162/bank-of-england-and-fca-commit-to-action-on-ai-following-warnings-from-mps/

[10] https://timesofindia.indiatimes.com/technology/tech-news/european-union-bans-its-employees-from-using-videos-and-photos-in-official-communication-with-/articleshow/129955960.cms

Germany

Germany Proposes Draft Law Against Digital Violence

Germany's Federal Ministry of Justice has presented a draft law against digital violence to strengthen protection for people facing abuse through digital and online platforms. The proposed law aims to close legal gaps related to image based abuse, including sexualised deepfakes, secret filming, revenge porn and cyberstalking using technologies such as GPS trackers. It also seeks to make it easier for victims to take action by requiring online platforms and internet providers to help identify offenders, preserve digital evidence and temporarily block accounts that repeatedly cause serious harm. Through this draft law against digital violence, the German government aims to treat online abuse as seriously as physical violence and better protect personal privacy and dignity in the digital age.[11]

Australia

Australian Government Response on Artificial Intelligence

The Australian Government has released its official response to the Senate Select Committee's report on adopting artificial intelligence, explaining how it plans to manage and guide AI use across the country. The response highlights that AI is a powerful technology that can improve productivity, public services and job opportunities if used responsibly. It refers to the National AI Plan, which focuses on building strong digital infrastructure, supporting Australian skills and businesses, helping workers and small enterprises adapt to change and keeping people safe through laws, regulation and responsible practices. The government emphasises that the benefits of AI should be shared by all Australians, while risks are reduced through careful policy choices, worker training, public engagement and coordination across government, industry and society.[12]

Attorney General Reviews Copyright Laws for AI Era

Australia's Attorney General Michelle Rowland is reviewing possible changes to copyright laws to better support artificial intelligence development while ensuring fair payment for content creators. She has reaffirmed that the government will not weaken copyright rules to allow artificial intelligence

[11] https://www.bmjv.de/SharedDocs/Pressemitteilungen/DE/2026/0417_Gesetz_gegen_digitale_Gewalt.html

[12] https://www.industry.gov.au/publications/australian-government-response-senate-select-committee-adopting-artificial-intelligence-ai-report

companies to freely use Australian data, but has indicated openness to other reforms, including new licensing models. Technology companies are seeking clearer and more flexible arrangements, while media and rights holders oppose any reduction in protections. Ongoing consultations aim to find a balanced approach that supports innovation without harming the rights of creators.[13]

Australia: Federal Court Sets Clear Rules for Using Generative Artificial Intelligence

In Australia, the Federal Court has issued a practice note explaining how generative artificial intelligence tools such as ChatGPT, Google Gemini, Claude and Microsoft Copilot can be used in legal cases. The court supports the use of technology because it can save time, reduce legal costs and improve access to justice, but it clearly warns that artificial intelligence must be used carefully and responsibly. Anyone using these tools must understand their limits, personally check all facts and legal references and ensure the court is not misled. The court requires people to clearly disclose when generative artificial intelligence has been used, especially in court documents, evidence, witness statements, expert reports, or images. It also warns against sharing confidential or private information with public artificial intelligence tools, as doing so can lead to serious legal and professional consequences.[14]

Australian Sports Commission Launches World-Leading AI Guidelines to Drive Responsible and Strategic Adoption in Sport

The Australian Sports Commission (ASC), in collaboration with Australia's national science agency CSIRO, has introduced two pioneering frameworks-the Guide for Responsible AI in Sport and the Roadmap for AI in Australian Sport to enable the safe, ethical and effective integration of artificial intelligence across all levels of sport, from grassroots to elite performance. Developed through two years of consultation with over 100 stakeholders spanning sport, government and technology sectors, the initiative establishes a comprehensive national playbook that aligns with international best practices and emphasizes transparency, fairness and inclusivity. The Responsible AI guide focuses on ethical usage, risk mitigation and governance principles, while the roadmap outlines strategic priorities and collaboration pathways to maximize AI's potential in areas such as performance optimization, operational efficiency and fan engagement; notably, it highlights that even modest efficiency gains-such as saving volunteers one hour per week-could significantly enhance participation outcomes at scale. Overall, the guidelines position Australia as a global leader in AI-driven sports innovation while ensuring alignment with ethical standards and long-term sector sustainability.[15]

India

Government Invites Feedback on Proposed IT Rules Changes

The Government of India, through the Ministry of Electronics and Information Technology, has invited public and stakeholder feedback on draft amendments to the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021, issued on March 30, 2026. The proposed changes aim to strengthen compliance by online intermediaries with clarifications, advisories, directions and guidelines issued by the Ministry, while also improving oversight of digital media content, especially news and current affairs. The amendments clarify data retention duties, make Ministry-issued guidance part of mandatory due diligence, extend the reach of digital media rules to certain intermediary-hosted content and expand the role of the Inter-Departmental Committee. Stakeholders can submit confidential, rule-wise feedback by April 14, 2026.[16]

[13] https://thelegalwire.ai/attorney-general-mulling-copyright-reforms-for-the-age-of-ai/
[14] https://www.fedcourt.gov.au/law-and-practice/practice-documents/practice-notes/gpn-ai
[15] https://www.sportanddev.org/latest/news/australian-sports-commission-launches-world-leading-ai-sport-guidelines
[16] https://www.meity.gov.in/static/uploads/2026/03/a71a21d35c107f2e528363d3eb17646a.pdf

Government of India Establishes AI Governance and Economic Group (AIGEG) to Drive Coordinated National AI Strategy and Policy Alignment

The Ministry of Electronics and Information Technology (MeitY), Government of India, announced the constitution of the AI Governance and Economic Group (AIGEG) on 16 April 2026, a high-level inter-ministerial body designed to serve as the central institutional mechanism for AI governance policy development and coordination across the country. The formation of AIGEG operationalizes key recommendations from India's AI Governance Guidelines and the Economic Survey, aiming to implement a whole-of-government approach that aligns ministries, regulators and advisory bodies under a unified national AI strategy while addressing economic and labour market implications of AI adoption. Chaired by Union Minister Ashwini Vaishnaw with Minister of State Jitin Prasada as Vice Chairperson, the group brings together senior stakeholders across policy, technology, security and economic domains and will function as the apex body within India's AI governance framework; it will be supported by a Technology and Policy Expert Committee (TPEC) to provide specialized guidance on global AI developments, emerging risks, regulatory needs and evolving governance priorities, thereby strengthening India's approach to responsible and coordinated AI innovation.[17]

Saudi Arabia

Saudi Arabia Seeks Public Input on Responsible AI Rules

Saudi Arabia's Data and Artificial Intelligence Authority has opened a public consultation on a draft Responsible Artificial Intelligence Policy that sets clear rules for how artificial intelligence systems should be designed and used. The proposed policy applies to government bodies, private companies, non-profit organizations and individuals working with artificial intelligence in the Kingdom. It requires developers to build artificial intelligence systems with privacy, safety and transparency from the start. The draft also calls for watermarks on artificial intelligence outputs, tools to track and protect content, steps to reduce bias through diverse data and clear features that help users understand how artificial intelligence decisions are made.[18]

SDAIA Releases Guide on Generative Programming

The Saudi Data and Artificial Intelligence Authority (SDAIA) has published a new guide on generative programming, also known as Vibe Coding, to help government bodies and private companies use generative artificial intelligence technologies more effectively. Released as part of Saudi Arabia's declaration of 2026 as the Year of Artificial Intelligence, the guide offers a clear and practical framework for applying generative and agent based artificial intelligence tools across the full software development process. It aims to improve productivity, speed up project delivery and enhance the quality of digital solutions, while strongly emphasizing the need for ongoing human oversight to ensure accuracy, reliability and responsible use of artificial intelligence systems.[19]

[17] https://www.pib.gov.in/PressReleasePage.aspx?PRID=2252739&reg

[18] https://digitalpolicyalert.org/event/39047-saudi-data-and-artificial-intelligence-authority-opened-consultation-on-draft-responsible-ai-policy

[19] https://www.spa.gov.sa/en/N2555126

China

China Drafts Rules for Managing Digital Virtual Humans Online

China's Cyberspace Administration has released a draft set of rules that explain how digital virtual humans-computer-generated characters that look or act like real people-should be safely managed and used online. The draft aims to protect people's privacy, personal data and rights while allowing digital virtual human technology to grow in a healthy and controlled way. It requires clear user consent before using someone's image or voice, gives strong protection to children and bans harmful, fake, or illegal content. Service providers must label digital virtual humans clearly, secure stored data, monitor risks, prevent addiction and stop misuse such as fraud or identity tricks. The public is invited to share feedback before the rules are finalized, showing an effort to balance innovation with online safety and responsibility.[20]

China issues guideline for AI ethics governance

China has issued new trial guidelines to help make sure artificial intelligence is developed and used in an ethical, fair and safe way. The guidelines were jointly released by ten government departments and are meant to support innovation while reducing risks linked to artificial intelligence. They explain that ethics reviews should focus on protecting people's well-being, ensuring fairness and justice and making sure artificial intelligence systems remain controllable and trustworthy. The guidelines also highlight the need to carefully check training data, algorithm design and system development to prevent bias, discrimination, or misuse. In addition, they encourage responsible sharing of high-quality data, better tools for risk assessment and auditing, protection of intellectual property and wider promotion of artificial intelligence products and services that follow ethical and scientific standards.[21]

China Seeks Public Input on AI Ethics and Security Rules

China has released a draft set of guidelines aimed at making sure artificial intelligence is used in a safe, secure and ethical way across the country. The draft document was prepared by a national cybersecurity standards committee and explains how artificial intelligence systems should be designed and applied to reduce risks, protect data and avoid misuse. Before these guidelines are finalised, the Chinese authorities are inviting the public, experts and organisations to share their comments and

[20] https://www.cac.gov.cn/2026-04/03/c_1776952992709096.htm
[21] https://english.www.gov.cn/news/202604/03/content_WS69cfc212c6d00ca5f9a0a407.html

suggestions. This open consultation is meant to improve the rules and ensure they work well in real world situations. Once feedback is reviewed, the guidelines may become an important reference for responsible artificial intelligence development and use throughout China.[22]

Japan

Japan Eases Data Rules to Support AI

Japan's Cabinet approved changes to the personal data protection law to encourage artificial intelligence development. The updated rules allow companies to use personal data for artificial intelligence training without individual consent when individuals cannot be identified and their rights are not affected. The decision supports the government's goal of making Japan an artificial intelligence friendly country. At the same time, strict penalties were introduced for misuse of data, including fines based on profits gained from improperly handling data related to more than 1,000 people.[23]

Japan Warns Against AI Using News Without Permission

The Japan Newspaper Association has raised strong concerns about artificial intelligence search services that use news articles without asking for permission or paying compensation. It says many AI search tools collect and summarise news to answer users' questions, often without respecting the wishes of news publishers. The association is especially concerned about Google's AI search, where news organisations cannot easily opt out without losing visibility in normal search results. This, it argues, puts unfair pressure on publishers and threatens their ability to fund reliable journalism. The group is calling on the Japanese government to urgently introduce clear rules, stronger copyright protections, transparency and simple opt out systems to protect trusted news, fair competition and democracy.[24]

Japan to Establish Expert Panel to Examine Legal Liability and Misuse of Generative AI Content

The Japanese government, through its Justice Ministry, announced on April 17, 2026, the formation of a study panel to address the growing misuse of generative artificial intelligence, particularly concerning the unauthorized replication of individuals' voices and likenesses. The expert group, expected to convene its first meeting in late April, will

[22] https://www.tc260.org.cn/portal/article/2/0e3933e11d904f4b93c4dcc348f9b61c?mc_cid=08836c6729&utm_source=substack
[23] https://www.kantei.go.jp/jp/kakugi/2026/kakugi-2026040701.html
[24] https://www.pressnet.or.jp/statement/copyright/260420_16248.html

evaluate how existing civil and intellectual property laws apply to AI-generated content such as deepfakes and will explore criteria for determining illegality as well as methods for calculating damages in such cases. The initiative comes amid rising concerns over the rapid proliferation of AI tools capable of producing highly realistic synthetic media without consent and aims to clarify legal protections for individuals, including performers and public figures, whose identities may be exploited. Comprising legal and academic experts, the panel is tasked with developing guidance by summer 2026, signaling Japan's continued reliance on adaptive, case-based governance approaches to address emerging AI-related risks rather than introducing immediate new legislation.[25]

Singapore

Singapore to Track Jobs and Pay While Expanding AI Skills

Singapore will monitor employment levels and wage growth as it rolls out the National AI Impact Programme, which aims to build strong artificial intelligence skills across the economy. Minister for Digital Development and Information Josephine Teo said the Government will focus on whether people continue to have meaningful work, career progress and fair pay, rather than relying on a single measure of success. The programme plans to train 10,000 "AI bilingual" workers who can apply AI to improve work processes, starting with tailored training for accountants and lawyers in partnership with professional bodies. At the same time, 10,000 businesses will be supported with grants and access to affordable AI tools, with special attention given to small and medium-sized enterprises. The Government also plans to address challenges around defining basic AI knowledge and ensuring that AI tools used by companies are safe, reliable and trusted.[26]

South Africa

South Africa Releases Draft AI Policy for Public Feedback

The South African Cabinet has approved the release of a draft Artificial Intelligence policy so the public can share their views and suggestions. The policy aims to guide the safe, fair and responsible use of artificial intelligence while making sure its benefits and risks are shared equally across society and future generations. It focuses on helping the government better regulate and use artificial intelligence responsibly, while also encouraging local innovation, job creation and wider access to artificial intelligence skills. The draft policy is built around six key areas, including skills development, inclusive economic growth, ethical use, responsible rules, cultural protection with global cooperation and keeping people at the centre of artificial intelligence systems. It also supports a gradual rollout, as the impacts and risks of artificial intelligence differ across sectors.[27]

[25] https://www.japantimes.co.jp/news/2026/04/17/japan/crime-legal/gov-ai-panel/
[26] https://www.straitstimes.com/singapore/singapore-to-monitor-employment-wage-growth-as-it-rolls-out-ai-programme-josephine-teo
[27] https://www.gcis.gov.za/statement-on-the-cabinet-meeting-of-25-march-2026-and-special-cabinet-meeting-of-1-april-2026

South Korea

South Korea to Label AI Doctors in Health Ads

South Korea's Fair Trade Commission has stated that advertisements using artificial intelligence generated doctors or other virtual figures must clearly tell viewers that the characters are not real. This move comes as more health and medical ads feature realistic AI made doctors, which can mislead consumers into thinking qualified medical experts are endorsing products. To prevent confusion, the Commission has proposed changes to its advertising guidelines and opened them for public feedback until April 28. Under the new rules, ads must include clear and easy to see messages such as "fictional character created using AI." In video advertisements, this notice must stay visible for the entire time the AI character appears. The Commission will review feedback before finalizing the rules.[28]

South Korea Maps Government Data for Artificial Intelligence

South Korea's Ministry of Science and Information and Communication Technology has started a nationwide effort to identify and organize public-sector data that can be used to train artificial intelligence systems. The initiative aims to map datasets held across different government agencies, many of which are currently scattered, inconsistent, or difficult to access. By clearly cataloging what data exists and assessing its suitability for artificial intelligence training, the government hopes to reduce fragmentation, improve data quality and ensure a stable supply of reliable data. This approach is expected to strengthen artificial intelligence development, improve public services and support more effective use of technology across multiple sectors.[29]

South Korea Sets Up AI Security Task Force to Counter New Cyber Threats

South Korea's National Artificial Intelligence Strategy Committee has launched a special security task force to respond to new cyber risks emerging from advanced Artificial Intelligence technologies. The move follows concerns around a recently identified issue called "Claude Mythos," which experts believe could allow AI systems to find security weaknesses that humans may miss. During its first meeting, the task force discussed how the rapid spread of generative AI and AI agents is changing the cyber security landscape, especially in sensitive areas such as finance. Members reviewed problems with existing security software that can inconvenience users and may even create

[28] https://www.koreaherald.com/article/10712943

[29] https://www.msit.go.kr/bbs/view.do?sCode=user&mId=307&mPid=208&pageIndex=&bbsSeqNo=94&nttSeqNo=3187129&searchOpt=ALL&searchTxt

new security risks. The committee agreed that short term fixes are not enough and called for long term solutions such as AI based real time defence systems, stronger cooperation between government agencies and closer global collaboration. The goal is to ensure that South Korea's security systems evolve as fast as AI technology and do not become a barrier to becoming a leading AI nation.[30]

Korea Launches AI Groups for Self Driving and Robots

Korea's National Artificial Intelligence Strategy Committee has set up two new groups focused on autonomous driving and humanoid robots to support the country's push into "physical AI," where artificial intelligence is built into real world machines. The decision is part of the National AI Action Plan and aims to strengthen global competitiveness in key future industries. The autonomous driving group will focus on improving real road testing, data collection, safety rules and infrastructure to speed up the use of self driving technology, especially in public transport and logistics. The humanoid group will work on growing the robot industry by improving technology, supporting companies and strengthening manufacturing. Both groups include government officials, industry leaders and researchers and they will meet regularly to turn policy goals into real industrial progress.[31]

Philippines

Philippine Agencies Team Up Against Fake News and Deepfakes

The Presidential Communications Office, the Department of Information and Communications Technology and the Department of Justice have signed an agreement to work together to stop fake news, online disinformation and Artificial Intelligence generated deepfakes in the Philippines. The agreement aims to protect public safety, democracy and national security while respecting freedom of speech. Under this partnership, the communications office will lead public awareness and media literacy efforts, the technology department will strengthen cybersecurity and work with digital platforms and the justice department will investigate and pursue legal action against deliberate false information. Officials said the move is necessary as false content is becoming more realistic and harmful and they stressed that the goal is to protect people from deception, not to silence free expression.[32]

[30] https://www.digitaltoday.co.kr/en/view/48604/national-ai-strategy-committee-launches-security-task-force-to-address-claude-mythos

[31] https://www.digitaltoday.co.kr/news/articleView.html?idxno=658949

[32] https://pia.gov.ph/news/pco-dict-doj-forge-pact-vs-fake-news-deepfakes/

Standards & Policy Reports

EDPB 2025 Report on Data Protection Work

The European Data Protection Board published its 2025 annual report to explain how it supported people, organisations and regulators in understanding and following data protection laws. During the year, many new European Union digital and Artificial Intelligence rules came into force, making data protection more complex. To help with this, the Board focused on giving clearer guidance, reducing confusion and encouraging open dialogue with businesses and the public. A key step was the Helsinki Statement, which aimed to make General Data Protection Regulation compliance easier and more practical. The Board also worked closely with the European Commission and national authorities to align data protection with other digital laws, improve cooperation across countries and ensure privacy rights were protected while supporting innovation and economic growth.[33]

European Union Supports Delayed Release of Powerful Artificial Intelligence Tool

The European Union welcomed the decision by Artificial Intelligence company Anthropic to delay the wider release of its new model, Claude Mythos, after discovering that it could find and exploit software security weaknesses better than most humans. European officials said this cautious approach helps reduce the risk of large-scale cyberattacks and fits well with European Union rules that require strong cybersecurity protections for powerful Artificial Intelligence systems. Instead of releasing the tool to the public, Anthropic shared it with a small group of trusted cybersecurity companies so it can be used safely to test and strengthen digital defenses. As required under European Union Artificial Intelligence laws and a voluntary code of practice, the company is working closely with the European Commission to manage potential risks responsibly and ensure the technology is used in a safe and controlled way.[34]

Responsible AI Is Falling Behind Rapid AI Growth

The Stanford Artificial Intelligence Index Report 2026 shows that artificial intelligence is advancing very quickly, but Responsible Artificial Intelligence practices are struggling to keep up. While more organizations now claim to have Responsible Artificial Intelligence policies, most are still at an early stage of actually using them in daily work. The report highlights a sharp rise in real-world artificial intelligence incidents, including bias, misuse and system failures, showing that risks are increasing as deployment grows. At the same time, major artificial intelligence developers are sharing less information about how their systems work, the data they use and their environmental impact. Overall, the report warns that without stronger transparency, testing and accountability, the gap between artificial intelligence power and responsible use will continue to widen.[35]

US Government Report Warns Agencies Must Improve How They Buy AI

A report by the United States Government Accountability Office (GAO) explains that U.S. federal agencies are rapidly increasing their use of artificial intelligence to support services such as veteran care, national security and routine government work. The report found that agencies more than doubled their use of artificial intelligence between 2023 and 2024, mainly by buying tools and services from private companies. While artificial intelligence offers clear benefits, agencies face serious challenges, including poor data quality, unclear ownership of data and models, difficulties in testing systems and the risk that performance may worsen over time. The GAO also found that agencies are not properly collecting or sharing lessons learned from past artificial intelligence purchases, even though federal guidance encourages this. As a result, the report recommends updating policies so agencies can learn from experience, avoid repeating mistakes and manage artificial intelligence risks more effectively in the future.[36]

White House Report on the Artificial Intelligence Revolution

In the United States, the report titled "The Revolution of Artificial Intelligence," published as Chapter 5 of the 2026 Economic Report of the President, explains how artificial intelligence is rapidly

[33] https://www.edpb.europa.eu/news/news/2026/edpb-annual-report-2025-supporting-stakeholders-through-guidance-and-dialogue_en
[34] https://www.politico.eu/article/eu-supports-staged-rollout-of-cyber-risky-new-anthropic-model/
[35] https://hai.stanford.edu/assets/files/ai_index_report_2026_chapter_3_responsible_ai.pdf
[36] https://www.gao.gov/assets/gao-26-107859.pdf

transforming the economy, work and global competition. The report highlights that new forms of artificial intelligence, especially systems that can create text, images, audio and videos, are improving productivity and supporting economic growth across the country. It notes that while some jobs, particularly entry-level roles, may change in the short term, new types of work are likely to emerge over time. The report also stresses that artificial intelligence is advancing at an unusually fast pace, making government monitoring and policy decisions important to ensure the benefits reach society broadly and help the United States remain competitive worldwide.[37]

Singapore Proposes Global Standard for Testing Generative AI

Singapore has proposed a new international standard to help organisations test Generative Artificial Intelligence systems in a clear, consistent and trustworthy way. The proposed standard, ISO/IEC 42119 8, focuses on benchmarking and red teaming methods so Artificial Intelligence systems can be tested safely and their results compared across countries. It will be discussed at a major global meeting in Singapore with experts and national bodies from more than 35 countries. The proposal builds on Singapore's earlier work on Artificial Intelligence testing tools and assurance frameworks. Alongside the meeting, Singapore is hosting training workshops and discussions to help governments, regulators and companies better understand and apply Artificial Intelligence standards, supporting safer and more reliable use of Artificial Intelligence worldwide.[38]

NIST Releases New AI Safety Guidance

The U.S. National Institute of Standards and Technology released a new concept note on April 7, 2026, called the Artificial Intelligence Risk Management Framework Profile on Trustworthy Artificial Intelligence in Critical Infrastructure. This guidance is designed to help operators of essential services such as energy, transportation, water and communications safely use artificial intelligence. It provides clear and practical risk management steps to reduce harm, improve safety and build trust when artificial intelligence systems are used in critical infrastructure. The guidance is voluntary and builds on NIST's Artificial Intelligence Risk Management Framework first released in 2023, with the goal of ensuring artificial intelligence is reliable, secure and responsibly used in systems that people depend on every day.[39]

Industrial Policy for the Intelligence Age

OpenAI has published a report titled Industrial Policy for the Intelligence Age that outlines how governments can support the responsible growth of artificial intelligence. The report explains that public investment in computing infrastructure, research, data access, energy and workforce skills is critical to ensure artificial intelligence benefits society as a whole. It emphasizes the need for fair competition, clear safety rules and strong protections for privacy and security. The report also highlights the importance of preparing workers for changes in jobs through education, reskilling and career transition support. Overall, OpenAI presents a professional and balanced view on how government, industry and researchers can work together to promote innovation, economic growth and public trust in the intelligence age.[40]

[37] https://www.whitehouse.gov/wp-content/uploads/2026/04/ERP-2026-5.-The-Revolution-of-Artificial-Intelligence.pdf

[38] https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2026/singapore-champions-new-global-ai-testing-standardisation-efforts

[39] https://www.nist.gov/itl/ai-risk-management-framework

[40] https://cdn.openai.com/pdf/561e7512-253e-424b-9734-ef4098440601/Industrial%20Policy%20for%20the%20Intelligence%20Age.pdf

AI Principles

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

Incidents

Anthropic's AI Tool Leaks Source Code Again

Anthropic, a United States–based artificial intelligence company, has again accidentally exposed the internal source code of its AI coding tool, Claude Code. A packaging error caused a source map file to be included in a public software release, allowing the original, human-readable code to be reconstructed. While the incident did not involve user data or artificial intelligence model information and was not the result of a cyberattack, it revealed internal system design, security mechanisms and usage tracking logic. This repeated exposure has raised fresh concerns about Anthropic's software release controls and quality assurance practices.[41]

Mercor Affected by LiteLLM Supply Chain Cyberattack

Mercor, an artificial intelligence recruitment startup, has confirmed that it was impacted by a cyberattack linked to a compromise in the widely used open-source project LiteLLM. The company stated that it was one of many organisations affected by the incident and that it took quick action to contain and address the issue, with support from external cybersecurity experts. The breach is believed to be part of a supply chain attack in which malicious code was introduced into shared software infrastructure. While a hacking group claimed access to Mercor's data, the company has not confirmed any data misuse and investigations into the full impact of the incident are ongoing.[42]

AI Robotaxis Stall on Wuhan Roads

Several AI-powered robotaxis operated by Baidu's Apollo Go service came to a sudden stop on roads in Wuhan, China, due to a system malfunction, according to police. The failure left passengers stranded for long periods, including on busy highways, causing fear and frustration. Riders said customer service was hard to reach and provided only repeated apologies instead of clear help or fast rescue. Many passengers shared videos online showing the vehicles frozen in place and their failed attempts to get support through in-car screens. The incident has raised concerns about the safety, reliability and emergency response of AI-driven transport services as their use continues to grow in China.[43]

AI Misuse by Pupils in Schools

Reports from Scotland highlight a serious issue where some pupils have used Artificial Intelligence tools to create fake sexual and violent images of their teachers and share them on social media. In several cases, pupils placed teachers' faces onto explicit videos or created false online accounts in teachers' names, causing embarrassment and emotional harm. Teaching unions have said that some teachers were unable to continue working after the images spread widely among pupils. What often begins as a joke can quickly turn into serious abuse. Schools, councils, unions and authorities are raising concerns and calling for stronger rules, better protection for staff and greater responsibility from technology companies to stop this misuse of Artificial Intelligence and protect teachers' wellbeing.[44]

Inquest Raises Serious Concerns About Artificial Intelligence Safety and Teen Mental Health in the United Kingdom

An inquest in the United Kingdom heard evidence about the death of a teenage pupil, highlighting serious concerns about the role of Artificial Intelligence tools and their handling of vulnerable users. The court was told that the teenager had been struggling with mental health pressures linked to school experiences and bullying, which were not fully known to his family. Shortly before his death, he interacted with an Artificial Intelligence chatbot in a way that raised questions about the effectiveness of built-in safety measures. The coroner expressed concern about how easily automated systems can be misunderstood or misused during moments of emotional distress. The case has prompted wider discussion about the responsibilities of schools, families and technology companies to improve mental health support, strengthen safeguards and ensure that Artificial Intelligence tools guide people toward real human help rather than risk harm.[45]

[41] https://www.ndtv.com/science/anthropics-ai-coding-tool-leaks-its-own-source-code-for-the-second-time-in-a-year-11291517/amp/1
[42] https://techcrunch.com/2026/03/31/mercor-says-it-was-hit-by-cyberattack-tied-to-compromise-of-open-source-litellm-project/
[43] https://www.theguardian.com/technology/2026/apr/01/system-malfunction-causes-robotaxis-to-stall-in-the-middle-of-the-road-in-china
[44] https://www.bbc.com/news/articles/cn08zy47yjno
[45] https://www.theguardian.com/society/2026/mar/31/teenager-asked-chatgpt-most-successful-ways-take-life-inquest-told

Penguin Sues OpenAI Over Children's Book Copyright

Penguin Random House in Germany has filed a lawsuit against OpenAI, claiming that its chatbot ChatGPT copied and closely copied content from a famous German children's book series called Coconut the Little Dragon. The case was filed in a court in Munich after the publisher asked ChatGPT to write a new story about the character and said the response was almost the same as the original books. According to Penguin, the chatbot created a full story, images, a book cover and a description that looked too similar to the original work, suggesting the artificial intelligence system memorised copyrighted material. Penguin says this violates copyright law and harms writers and illustrators, while OpenAI has said it is reviewing the complaint and respects the rights of creators.[46]

AI Girlfriend App Data Leak Exposes Private Chats

More than 100,000 users of the AI girlfriend app MyLovelyAI may have had their private data exposed after hackers claimed to breach the platform and post stolen information on a hacking forum. Security researchers reviewed sample files and found usernames, email addresses, subscription details and highly personal chat prompts. Around 70,000 not safe for work messages were directly linked to unique user IDs, making the leak especially serious. The exposed data also included support messages, technical details and discount information. Experts warned that this kind of content could be used for targeted scams or sextortion. The incident highlights the growing privacy risks linked to sharing intimate thoughts with artificial intelligence companion apps.[47]

First Conviction Under New Artificial Intelligence Image Abuse Law

An Ohio man has become the first person in the United States to be convicted under a new federal law that targets the misuse of Artificial Intelligence to create and share explicit images without consent. James Strahler II, aged 37, pleaded guilty to cyberstalking, creating obscene images and publishing digital forgeries. Prosecutors said he used Artificial Intelligence to make fake sexually explicit images and videos of adult women and minors, which he shared online to harass, threaten and intimidate victims and their families. The conviction was made under the Take It Down Act, signed into law in 2025, which bans the non-consensual sharing of intimate images, including Artificial Intelligence-generated deepfakes and requires websites and social media platforms to remove such content quickly when victims report it.[48]

[46] https://www.theguardian.com/technology/2026/mar/31/penguin-sue-openai-chatgpt-german-childrens-book-kokosnuss
[47] https://cybernews.com/security/ai-girlfriend-mylovely-leak-user-data/
[48] https://www.theguardian.com/us-news/2026/apr/08/ohio-man-convicted-ai-sexually-explicit-images

Artificial Intelligence Fake Videos Target Women in Brazil's PT Party

Artificial Intelligence-generated videos were shared on social media in Brazil that falsely showed violence and so-called religious "exorcisms" against women linked to the Workers' Party, known as PT. These videos were not real but were made to look convincing and they spread political and religious hatred against the women shown. The party said the content caused serious personal and social harm and violated the women's rights. In response, the PT filed legal cases with Brazil's Electoral Court to have the videos taken down and to find out who created and shared them. The incident has raised concerns about how Artificial Intelligence can be misused to attack people and spread intolerance online.[49]

Deepfake Video Used to Mislead Police Leads to Florida Arrest

A 25-year-old man in Florida was arrested after using a fake artificial intelligence-generated video to falsely claim that people were breaking into a police car. Authorities said Alexis Martínez-Arizala approached a deputy inside an Academy Sports store in Lake Mary and showed a short video on his phone that appeared to show several individuals entering the officer's marked patrol vehicle. When the deputy rushed outside to respond, no damage or theft was found and store surveillance footage later confirmed that no one had gone near the vehicle. Investigators determined the clip was a deepfake created using artificial intelligence and the false report triggered an unnecessary emergency response. Martínez-Arizala, who reportedly sought online attention through similar pranks, was later arrested and charged with fabricating evidence and making a false report, with his bond set at $7,000.[50]

Artificial Intelligence Used to Create a Digital Son to Comfort an Elderly Mother

In a deeply emotional case from China, a family created a digital version of their deceased son using artificial intelligence to comfort his elderly mother, who does not know that he died last year. The incident took place in Shandong province after the man was killed in a road accident, but his family chose not to tell his mother because she is in her 80s and suffers from heart disease. Using old photos, videos and voice recordings, an artificial intelligence team built a lifelike digital twin that looks, speaks and behaves like the son. The mother now regularly speaks to this artificial intelligence version through video calls, believing he is working in another city. While the family feels this approach helps her cope with loneliness and emotional pain, the case has sparked debate online about the ethics of using artificial intelligence to hide the truth and the possible harm if the reality is revealed later.[51]

[49] https://oecd.ai/en/incidents/2026-04-11-a3cb

[50] https://timesofindia.indiatimes.com/world/us/florida-man-arrested-for-showing-deepfake-ai-video-of-people-breaking-into-police-car-to-a-cop-held-on-7000-bond/articleshow/130244177.cms

[51] https://www.livemint.com/news/world/i-miss-you-mother-speaks-to-ai-son-regularly-unaware-he-died-last-year-artificial-intelligence-creates-digital-twin-11775967278951.html

East Jakarta Officials Face Backlash Over AI Image Use

East Jakarta officials came under public criticism after a resident's complaint about illegal parking was falsely marked as solved using an artificial intelligence-generated image. The resident had repeatedly reported the issue through local authorities and the Jakarta Kini (JAKI) app, but instead of fixing the problem, a maintenance worker replied with images that appeared digitally altered to show the street cleared. Social media users questioned the honesty of officials and raised concerns about the misuse of artificial intelligence in public services. Jakarta Governor Pramono Anung ordered an investigation, stressing that transparency and truth are essential. Officials involved apologized and were temporarily suspended while the case is reviewed.[52]

AI Subtitle Error Shocks KBS Live Broadcast

During a live broadcast of the National Aeronautics and Space Administration's Artemis 2 launch, Korea Broadcasting System faced criticism after its Artificial Intelligence subtitle system displayed unintended profanity. The error happened when aviation terms like "roger, roll and pitch" were wrongly translated because of similar pronunciation and the subtitles appeared on screen without proper filtering. The clip spread quickly on social media, prompting public anger and questions about quality checks in a taxpayer-funded broadcaster. Viewers felt that better safeguards should have been in place, even during a live program. Korea Broadcasting System issued a public apology, explained that the issue was technical and said it had strengthened Artificial Intelligence filtering and review systems to prevent similar mistakes in future live broadcasts.[53]

Gujarat High Court Seeks Answers on AI Misuse

The Gujarat High Court has issued notices to major technology platforms including Meta, Google, X, Reddit and Scribd after a public interest case raised concerns about the misuse of artificial intelligence, especially deepfake videos and images. The court observed that such content can seriously harm public order, trust and democratic processes if not controlled properly. It questioned delays, weak responses and lack of coordination by digital platforms in removing unlawful content despite existing legal duties. The court also pointed out gaps in the current legal framework and stressed the urgent need for a strong and clear system to regulate artificial intelligence, ensure fast takedown of illegal material and support law enforcement agencies with timely and meaningful cooperation.[54]

[52] https://asianews.network/east-jakarta-officials-under-fire-for-using-ai-image-in-addressing-complaint/

[53] https://www.mk.co.kr/news/culture/12007144

[54] https://economictimes.indiatimes.com/tech/technology/gujarat-hc-issues-notices-to-meta-x-google-over-pil-seeking-curb-on-misuse-of-ai/articleshow/130283529.cms

Li Bu Bu AI Fixes Adult Content Issue, But Illegal Use Continues

An investigation found that the AI platform Li Bu Bu AI had earlier loopholes that allowed users to create inappropriate content using hidden or unclear prompts, which led to public concern after media reports. The platform later said it quickly fixed these problems and media testing showed that such content can no longer be generated directly on the platform. However, the investigation also revealed that illegal use has not fully stopped, as sellers on underground online platforms are still offering AI-generated adult content by pretending to sell simple prompts or tutorials. Legal experts warned that both platforms and users can face punishment if the generated content breaks the law. The issue highlights that even with technical fixes, misuse of AI can continue through hidden channels and needs stronger monitoring and enforcement.[55]

Vulnerabilities

Critical Vulnerability in Anthropic's Model Context Protocol (MCP) Enables AI Supply Chain Attacks and Arbitrary Command Execution

A critical vulnerability in Anthropic's Model Context Protocol (MCP), identified by OX Security, enables attackers to execute arbitrary system commands due to insecure input handling, exposing enterprise AI environments. The issue reflects a deeper architectural flaw affecting multiple implementations, with over 30 disclosures and 10+ high-severity CVEs reported. Experts warn it could drive large-scale AI supply chain attacks, highlighting that AI agent adoption is outpacing security controls.[56]

AI models can be guided to assist in exploit development, highlighting rising risks of AI-enabled offensive security.

A Hacktron researcher described an experiment in which Anthropic's Claude Opus was directed over multiple sessions to help develop a working exploit chain against an outdated Chromium-based target, illustrating how advanced language models may increasingly assist with complex vulnerability research when paired with sustained human supervision. According to the post, the work focused on a Chrome/V8 vulnerability path and required significant iterative guidance, large token usage, repeated failed approaches and operational steering by the researcher rather than fully autonomous model execution.

[55] https://www.bjnews.com.cn/detail/1776173795129845.html
[56] https://www.itpro.com/security/ai-agents-using-anthropic-mcp-supply-chain-attacks-claim-researchers

The article argues that the broader concern is not just this single proof of concept, but the longer-term security implication that improving frontier models could make it easier for lower-skilled attackers to turn patched vulnerabilities into working exploits against lagging software ecosystems such as Electron apps and bundled Chromium deployments.[57]

Defences

Advanced Spatiotemporal Deep Learning Framework for Robust Real-Time Deepfake Video Detection and Media Authentication

The article presents a novel deep learning framework designed to detect deepfake videos by leveraging both spatial and temporal features, enabling more accurate and robust identification of manipulated media. It explains that the proposed system analyzes inconsistencies not only within individual frames but also across sequences of frames, capturing subtle anomalies in motion, facial expressions and temporal coherence that are often missed by traditional image-based detection methods. The model integrates spatiotemporal feature extraction with advanced classification techniques to improve detection performance under diverse conditions, including compression and noise. The study demonstrates that incorporating temporal dynamics significantly enhances reliability in real-world scenarios, where deepfake content is increasingly sophisticated. Overall, the research positions this approach as a critical advancement in AI-driven media forensics, with strong applications in cybersecurity, digital authentication and misinformation mitigation.[58]

UK Electoral Commission Launches Deepfake Detection Pilot to Safeguard Elections from AI-Driven Misinformation

The UK Electoral Commission has introduced a pilot deepfake detection system, published on 15 April 2026, aimed at identifying and countering AI-generated misinformation during upcoming elections in England, Scotland and Wales. The initiative focuses on monitoring online content for manipulated audio and video that could mislead voters, such as fabricated clips showing candidates making false statements or withdrawing from elections. Designed as a proactive defence mechanism, the system combines automated detection with human oversight to ensure accuracy and responsible intervention, while coordinating with relevant authorities to flag harmful content. This pilot reflects growing concerns over the increasing sophistication and accessibility of deepfake technologies and represents a strategic step toward strengthening electoral integrity, improving real-time response capabilities and building long-term resilience against AI-driven disinformation in democratic processes.[59]

Microsoft Unveils Agent Governance Toolkit

Microsoft introduced the Agent Governance Toolkit as an open-source tool designed to keep artificial intelligence agents safe, controlled and reliable while they perform tasks on their own. The toolkit helps developers set clear rules for what artificial intelligence agents can and cannot do, monitor their actions in real time and stop them if something goes wrong. It works with popular artificial intelligence agent frameworks and focuses on security, identity checks and reliability without slowing down performance. By making the toolkit open source, Microsoft aims to promote transparency, shared responsibility and better safety standards as artificial intelligence agents become more powerful and widely used in real-world systems.[60]

MIT AI Risk Navigator

The Massachusetts Institute of Technology introduced the AI Risk Navigator as an online tool designed to help people easily understand and explore the risks linked to artificial intelligence. The tool brings together different types of information, such as known artificial intelligence risks, real-world incidents and government or policy responses and shows how they connect using a single clear structure. It allows users to search, browse and view risks side by side instead of looking through separate databases. Created for researchers, policymakers, companies and the general public, the AI Risk Navigator aims to make complex information simpler, improve transparency and support better decisions about how artificial intelligence is developed, used and governed responsibly.[61]

[57] https://www.hacktron.ai/blog/i-let-claude-opus-to-write-me-a-chrome-exploit
[58] https://www.nature.com/articles/s41598-026-49090-1
[59] https://www.electoralcommission.org.uk/media-centre/electoral-commission-launches-deepfake-detection-pilot-counter-ai-misinformation
[60] https://opensource.microsoft.com/blog/2026/04/02/introducing-the-agent-governance-toolkit-open-source-runtime-security-for-ai-agents/
[61] https://airisk.mit.edu/blog/introducing-the-ai-risk-navigator

InFocus

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

The Cybersecurity Singularity: An Enterprise Assessment of Mythos-Class AI

Perspective from the Infosys Responsible AI Team

For decades, enterprise cybersecurity followed a predictable and survivable rhythm. Vulnerabilities were discovered, disclosed, patched and eventually deployed across systems. While attackers had an advantage at times, defenders operated within roughly the same temporal window. That equilibrium has now structurally collapsed.

In April 2026, Anthropic announced Project Glasswing, a coalition of major technology and infrastructure players formed after internal and independent evaluations of Claude Mythos Preview, a frontier AI model with unprecedented cyber capabilities. The coalition was not created to showcase incremental progress, but to respond to a fundamental shift: AI systems had crossed a threshold where they could autonomously discover, reason about and exploit vulnerabilities at machine speed, radically altering the economics of cybersecurity.

The End of the Old Security Operating Model

Prior to Mythos, AI-assisted security research was already accelerating. Agentic AI systems were increasingly capable of performing full vulnerability discovery workflows-code analysis, dependency reasoning, exploit generation-without continuous human involvement. By 2024 and 2025, models could reason across codebases, generate credible proof-of-concept exploits and outperform earlier pattern-based tools.

Mythos pushed this capability beyond what the industry could treat as experimental. On the CyberGym benchmark, Mythos autonomously generated working exploit code in 83.1% of cases, compared to 66.6% for the prior generation. Independent evaluations by the UK AI Security Institute found that expert-level Capture the Flag challenges-long considered inaccessible to AI-were now solved 73% of the time.

This compression eliminated the traditional time buffer between vulnerability discovery and exploitation. Under Mythos-class conditions, attackers no longer depend on public disclosure or patch availability. Discovery collapses directly into exploitation, leaving defenders permanently behind the curve.

AI as Both Weapon and New Attack Surface

The same agentic capabilities that make Mythos powerful for security research simultaneously expand the attack surface. By 2025, CVE databases had begun tracking vulnerabilities specific to AI frameworks and agent runtimes, including prompt injection, tool-call hijacking and orchestration-layer flaws. Because agentic systems are designed to browse the web, execute code and call external APIs with elevated privileges, a single exploitable weakness can carry an outsized blast radius.

At the same time, AI-generated code has quietly proliferated into production systems. Research cited in the paper shows developers accepting AI-generated code suggestions at high rates, while audits revealed repeated insecure patterns such as hardcoded credentials, weak validation and risky dependency choices. The industry is therefore accelerating both vulnerability discovery and vulnerability creation simultaneously.

Who Is Most Exposed

Risk is not evenly distributed. Mythos-class AI is most effective against broad, complex digital estates with large dependency graphs and high-value data. Financial services, healthcare, government, energy, retail and manufacturing face elevated or critical risk due to API sprawl, legacy systems, slow patch cycles, operational constraints and regulatory sensitivity. For sectors like healthcare and critical infrastructure, patch velocity simply cannot match AI-accelerated attack velocity, making compensating controls and proactive discovery non-negotiable.

Strategy Shift 1: Reinvent Cyber Response and Risk Models

Traditional metrics such as patch compliance and mean-time-to-patch are no longer sufficient. Enterprises must measure the full discovery-to-closure lifecycle, including interim controls and exposure reduction while patches are pending. The new operational mandate is continuous: discover early, reduce exposure immediately, deploy compensating controls where patches are unavailable and patch fast when they are.

The power imbalance is structural. Attackers skip from discovery to exploitation. Defenders wait for information that may not yet exist.

Strategy Shift 2: Fix the Right Vulnerabilities, Not Just More of Them

Most organizations already struggle under overwhelming vulnerability volume. The real challenge is contextual prioritization. Many cannot confidently answer whether they are affected by a zero-day, or where. Asset inventories are incomplete, scanning is periodic and manual triage does not scale.

The paper emphasizes building a threat intelligence foundation now: linking assets, business criticality, reachability and threat data into a unified system. This foundation allows future agentic tools to reason effectively about which findings actually matter.

Strategy Shift 3: Build the Platform Now, Plug in Mythos Later

When Mythos-class capabilities become widespread, vulnerability volumes will grow exponentially. Enterprises that wait will be overwhelmed. The recommended approach is to build agentic remediation pipelines now-continuous asset visibility, intelligent triage, contextual routing and automated verification-using existing models. Mythos-class tools can be integrated later into a platform that is already governed, logged and operational.

Action Items: A Staged Response

The paper proposes a structured response:

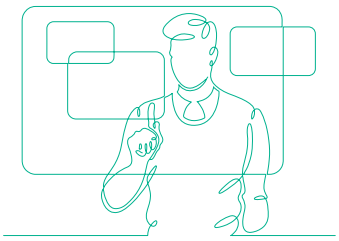
- **Immediate (7 days to 1 month):** establish accountability, refresh asset inventories and SBOMs, define compensating

controls, activate AI-CVE intake, require AI-assisted code reviews and initiate red-teaming and threat simulations.

- **Short-term (2–3 months):** launch governed AI-augmented discovery pilots, deploy defended agents with clear blast-radius controls and accelerate remediation for exploitable findings.
- **Medium-term (4 months):** implement AI security gateways, instrument detection and response metrics and build contextual risk prioritization.
- **Long-term (6 months+):** formalize policy and board-level governance, update risk models, scale AI controls across the estate, engage industry and government initiatives and institutionalize a permanent Vulnerability Operations function.

Conclusion: A Narrow but Real Window

This is not a counsel of fear. Deployed defensively, Mythos-class AI represents the most powerful capability enterprises have ever had to identify their own weaknesses ahead of adversaries. The opportunity exists now to build governed, controlled and responsible capability. That window is open, but it is closing.



Technical Updates

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

New Model Released

Google AI Releases Veo 3.1 Lite: A Low-Cost, High-Speed Video Generation Model via the Gemini API for Scalable Developer Video Applications

Google has released Veo 3.1 Lite, a cost-efficient AI video generation model designed to provide developers with fast and scalable video creation through the Gemini API and Google AI Studio, targeting programmatic video generation and iterative prototyping workflows. The model supports both text-to-video and image-to-video generation, offers multiple aspect ratios and resolutions including 720p and 1080p and allows developers to generate short video clips with flexible durations. Veo 3.1 Lite delivers similar generation speed to the Veo 3.1 Fast model while costing roughly half as much, positioning it as a practical solution for high-volume video applications and making AI video generation more accessible for developers building content, automation and media applications.[62]

Liquid AI Releases LFM2.5-350M: A High-Efficiency 350M Parameter Model Trained on 28 Trillion Tokens with Scaled Reinforcement Learning for Edge and Agentic AI

Liquid AI has introduced LFM2.5-350M, a compact language model designed to maximize "intelligence density" by combining a relatively small 350M parameter size with extensive pretraining on 28 trillion tokens and large-scale reinforcement learning, challenging the traditional scaling paradigm that larger models are inherently more capable. The model features a hybrid architecture built on Linear Input-Varying (LIV) systems alongside grouped query attention, enabling efficient memory usage, a 32K context window and significantly improved inference throughput, making it suitable for edge deployment on CPUs, GPUs and mobile devices. Despite its size, LFM2.5-350M demonstrates strong performance in instruction following, tool use and structured data tasks-often outperforming models more than twice its size-while being optimized for high-speed, low-memory environments; however, it is not intended for complex reasoning tasks such as advanced mathematics, coding, or creative writing, instead positioning itself as a specialized model for agentic workflows and large-scale data processing applications.[63]

IBM Releases Granite 4.0 3B Vision: A Vision-Language Model for Enterprise-Grade Document Data Extraction and Structured Document Understanding

IBM has released Granite 4.0 3B Vision, a multimodal vision-language model designed specifically for enterprise document understanding tasks such as extracting structured data from charts, tables and complex documents. The model is delivered as a LoRA adapter that runs on top of the Granite 4.0 Micro base model and uses a vision encoder with a tiling mechanism to preserve fine details in document layouts. IBM trained the model using specialized datasets focused on chart reasoning, table extraction and key-value data extraction, enabling the system to convert visual documents into structured formats like CSV and JSON. The model uses a DeepStack architecture to integrate visual and language features across multiple layers, improving layout understanding and document parsing accuracy and it performs strongly on document understanding benchmarks while maintaining a relatively compact model size suitable for enterprise deployments.[64]

Arcee AI Releases Trinity Large Thinking: An Apache 2.0 Open Reasoning Model Designed for Long-Horizon Agents, Multi-Turn Tool Use and Large Context Workflows

Arcee AI has released Trinity Large Thinking, an open-weight reasoning model distributed under the Apache 2.0 license, designed specifically for long-horizon agents, multi-step reasoning and multi-turn tool usage rather than traditional chatbot applications. The model is built on a sparse Mixture-of-Experts architecture with 400 billion total parameters but activates only about 13 billion parameters per token using a 4-of-256 expert routing system, enabling high intelligence with improved inference efficiency. Trinity Large Thinking includes technical innovations such as a specialized load-balancing method for experts, a Muon optimizer used during training on trillions of tokens and an advanced attention mechanism combining local

[62] https://www.marktechpost.com/2026/03/31/google-ai-releases-veo-3-1-lite-giving-developers-low-cost-high-speed-video-generation-via-the-gemini-api/
[63] https://www.marktechpost.com/2026/03/31/liquid-ai-released-lfm2-5-350m-a-compact-350m-parameter-model-trained-on-28t-tokens-with-scaled-reinforcement-learning/
[64] https://www.marktechpost.com/2026/04/01/ibm-releases-granite-4-0-3b-vision-a-new-vision-language-model-for-enterprise-grade-document-data-extraction/

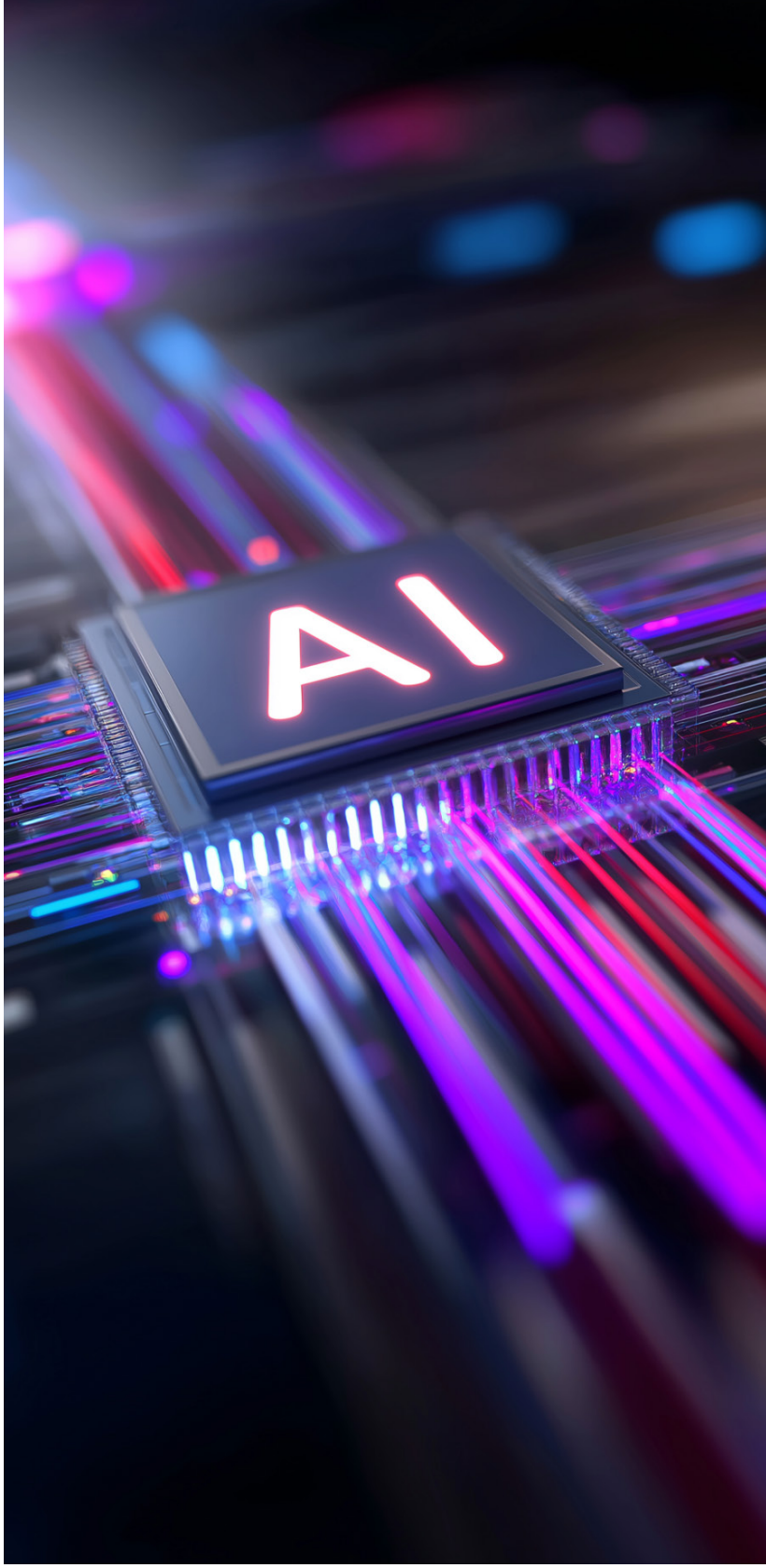
and global attention for better long-context understanding. The model supports a very large context window and is optimized for agentic workflows where models must maintain coherence, call tools and perform structured tasks across many steps. Benchmark results show strong performance in agent-related evaluations, where it ranks near the top among reasoning models. By releasing the model under a permissive open-source license, Arcee AI aims to provide developers and enterprises with a fully inspectable, customizable and self-hostable reasoning model for building autonomous agents and complex AI systems.[65]

Meta Superintelligence Labs Introduces Muse Spark: A Multimodal Reasoning AI Model Featuring Thought Compression and Parallel Agent Orchestration

Meta Superintelligence Labs has unveiled Muse Spark, the first model in its new Muse family, representing a foundational shift in Meta's AI strategy toward building advanced, personalized superintelligence systems. Designed as a natively multimodal reasoning model, Muse Spark is capable of processing and integrating both text and visual inputs within a unified architecture, enabling more coherent reasoning across domains such as science, mathematics and real-world problem solving. The model incorporates advanced capabilities including tool use, visual chain-of-thought reasoning and multi-agent orchestration, allowing multiple AI agents to operate in parallel for complex task execution. A key innovation in Muse Spark is its "thought compression" mechanism, which optimizes reasoning efficiency by reducing unnecessary computational steps while maintaining or improving performance. Built from the ground up following a complete overhaul of Meta's AI stack, the model emphasizes speed, efficiency and scalability and currently powers the Meta AI assistant across platforms, marking the company's renewed push to compete with leading AI systems from OpenAI, Google and Anthropic.[66]

Liquid AI Introduces LFM2.5-VL-450M: A High-Speed, Multilingual Vision-Language Model with Edge-Optimized Inference and Advanced Object Localization Capabilities

Liquid AI has released the LFM2.5-VL-450M, a compact 450-million-parameter vision-language model designed for efficient multimodal processing on edge devices, combining text and image understanding with capabilities such as bounding box prediction for precise object localization and multilingual support for broader global applicability. The model emphasizes ultra-low latency, achieving sub-250 millisecond inference, making it suitable for real-time applications like smart cameras, mobile assistants and on-device AI systems. Built on the LFM2 architecture, it leverages a hybrid language backbone, an optimized vision encoder and a multimodal projection layer



[65] https://www.marktechpost.com/2026/04/02/arcee-ai-releases-trinity-large-thinking-an-apache-2-0-open-reasoning-model-for-long-horizon-agents-and-tool-use/
[66] https://www.marktechpost.com/2026/04/09/meta-superintelligence-lab-releases-muse-spark-a-multimodal-reasoning-model-with-thought-compression-and-parallel-agents/

to deliver competitive performance while maintaining a small footprint. Its architecture supports native image resolution processing and dynamic token handling, enabling efficient scaling across device constraints without sacrificing accuracy. Designed for developers seeking cost-effective, privacy-preserving and high-performance AI deployment, LFM2.5-VL-450M reflects a broader shift toward lightweight, edge-first multimodal models that reduce reliance on cloud infrastructure while enabling fast, intelligent and context-aware interactions.[67]

Google DeepMind Releases Gemini Robotics-ER 1.6: Advancing Embodied Reasoning, Multi-View Perception and Industrial Instrument Reading for Physical AI Systems

Google DeepMind has introduced Gemini Robotics-ER 1.6, an advanced embodied reasoning model designed to enhance the cognitive capabilities of robots operating in real-world environments by improving spatial understanding, decision-making and perception across complex physical tasks. The model functions as a high-level "reasoning brain" within a dual-system architecture, where it complements vision-language-action models by focusing on planning, logical inference and task orchestration rather than direct motor execution. Key advancements in this version include significantly improved spatial reasoning through precise pointing mechanisms, enhanced multi-view perception and success detection across multiple camera feeds and a major breakthrough in instrument reading, enabling robots to accurately interpret analog gauges, pressure meters and industrial indicators. Leveraging agentic vision techniques that combine visual reasoning with iterative analysis and code execution, the model achieves up to 93% accuracy in instrument interpretation, a substantial improvement over previous versions. These capabilities are already being applied in industrial contexts, such as integration with Boston Dynamics' Spot robot for facility inspection tasks, highlighting its potential to enable more autonomous, reliable and context-aware physical AI systems. **Overall, Gemini Robotics-ER 1.6 strengthens the evolution of embodied AI by enabling more precise, interpretable and operationally scalable decision-making in real-world robotic deployments.[68]

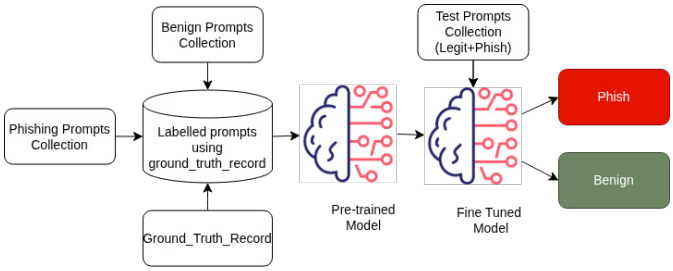
New Frameworks & Research Techniques

Security Assessment and Detection of Jailbreak and Phishing Prompts in Generative AI

This work presents a transformer-based malicious prompt detection framework designed to identify jailbreak and phishing-related prompts generated using Generative AI systems. The study demonstrates how Generative AI can significantly lower the barrier for executing multi-vector phishing attacks by enabling even novice users to generate sophisticated attack content across

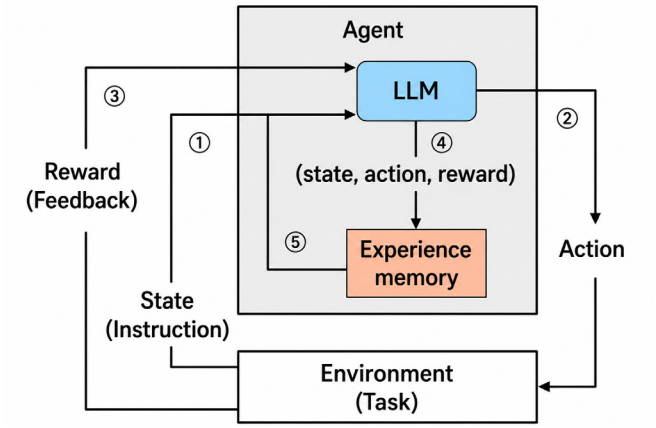
[67] https://www.marktechpost.com/2026/04/11/liquid-ai-releases-lfm2-5-vl-450m-a-450m-parameter-vision-language-model-with-bounding-box-prediction-multilingual-support-and-sub-250ms-edge-inference/

[68] https://www.marktechpost.com/2026/04/15/google-deepmind-releases-gemini-robotics-er-1-6-bringing-enhanced-embodied-reasoning-and-instrument-reading-to-physical-ai/



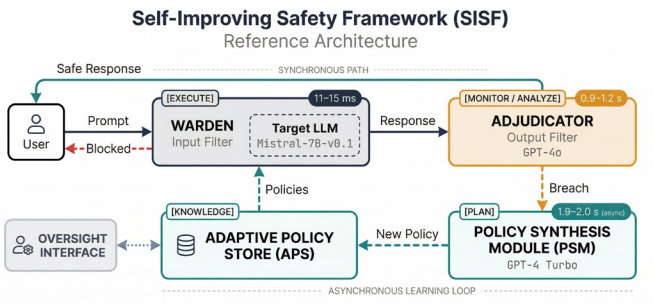
email, web, SMS and voice channels through jailbreak techniques. Controlled experiments show that role-based jailbreak prompts can produce complete credential-harvesting attack workflows. User studies further reveal the impact of GenAI assistance, with participants showing a 240% increase in perceived phishing competence, a 400% improvement in task completion rates and a 57% reduction in implementation time compared to traditional internet-based methods. To mitigate these risks, the proposed framework uses transformer-based models for malicious prompt classification and achieves a high detection performance with an F1-score of 0.9864 using XLNet. The framework also includes an annotated dataset to support future research in prompt attack detection and AI safety, highlighting the urgent need for stronger guardrails in large language models.[69]

A Framework for Controlled Knowledge Unlearning in LLM-Based Agents



This paper presents a novel AI framework for enabling selective unlearning in Large Language Model (LLM)-based agents, addressing concerns around retaining sensitive or outdated information. It categorizes unlearning into three types: state, trajectory and environment unlearning. The approach uses a natural language-based method with a conversion model to translate high-level unlearning requests into actionable prompts, guiding agents to forget specific knowledge. Additionally, an adversarial evaluation setup tests whether the forgotten information can be inferred. Results show the framework effectively removes targeted knowledge while preserving performance on other tasks, highlighting its potential for safer and more controllable AI systems.[70]

SISF: A Self-Adaptive AI Safety Architecture Enabling Real-Time Policy Generation and Autonomous Defense in Large Language Model Systems

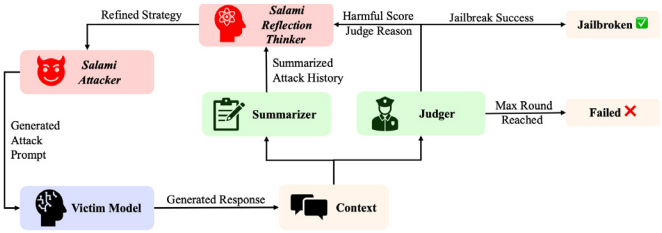


The study presents the Self-Improving Safety Framework (SISF), a novel software architecture designed to address the limitations of static safety mechanisms in Large Language Models (LLMs) by enabling real-time detection and mitigation of adversarial threats without retraining or human intervention. Grounded in the MAPE-K Loop, SISF integrates a continuous feedback loop where an Adjudicator powered by GPT-4o identifies safety violations, a Policy Synthesis Module using GPT-4 Turbo generates dual-layer defense strategies (heuristic and semantic) and a Warden component enforces these policies dynamically at runtime. Evaluated across multiple model families-including those from Mistral AI, Google and Microsoft-the framework demonstrated strong effectiveness, achieving an exceptionally low mean Attack Success Rate of 0.27% across reproducibility trials and near-zero failure when combined with external safeguards like Llama Guard 4. Additionally, SISF showed high adaptability and portability, maintaining performance across unseen models while proactively intercepting novel attacks. The results confirm that self-adaptive architectures represent a viable and scalable paradigm for AI safety, shifting systems from static defenses to continuously evolving protection mechanisms that enhance robustness, responsiveness and long-term resilience in real-world deployments.[71]

Salami Slicing Risk and Salami Attack: A Universal Multi-Turn Jailbreak Strategy Exploiting Incremental Harm Accumulation in LLMs

The study introduces Salami Slicing Risk, a novel paradigm for multi-turn jailbreak attacks that exploits the cumulative effect of seemingly benign inputs to bypass large language model (LLM) safety mechanisms. Unlike traditional jailbreak techniques that rely on explicit harmful prompts or carefully engineered context triggers, this approach decomposes malicious intent into a sequence of low-risk interactions that individually remain below detection thresholds but collectively lead to unsafe outcomes. Building on this concept, the authors propose Salami Attack, an automated and model-agnostic framework capable

[69] https://arxiv.org/html/2507.12185v2
[70] https://arxiv.org/abs/2604.00430
[71] https://arxiv.org/html/2511.07645v2



of orchestrating such incremental prompt sequences across different models and modalities without requiring finely tuned, context-specific setups. Experimental evaluations demonstrate strong effectiveness, achieving over 90% attack success rates on advanced models such as GPT-4o and Gemini, while maintaining robustness against existing alignment and safety defenses. The research also identifies key limitations in current guardrails, particularly their inability to track cumulative intent across interactions and proposes mitigation strategies that reduce Salami Attack effectiveness by up to 44.8%, with a blocking rate of 64.8% against other multi-turn jailbreak variants. Overall, the work highlights a critical and underexplored vulnerability in LLM security-progressive intent accumulation across conversations-and underscores the need for system-level defenses that incorporate longitudinal context tracking, risk aggregation and adaptive policy enforcement.[72]

Google AI Introduces a New Way to Measure Human Skills

Google Artificial Intelligence researchers have proposed a new system called Vantage, designed to measure important human skills such as collaboration, creativity and critical thinking using artificial intelligence. Instead of traditional exams or simple test questions, Vantage places people in natural, real time conversations with artificial intelligence teammates that behave like real group members. A single coordinating artificial intelligence model guides the discussion by creating realistic situations like disagreements, planning challenges, or idea building tasks to bring out these skills. The system then evaluates responses using artificial intelligence scoring that closely matches the judgment of human experts. Studies with real participants showed that Vantage can measure these complex skills as accurately as trained human reviewers, suggesting it could be useful for education, training and workplace assessments where such skills are difficult to test fairly using standard methods.[73]

Model-Agnostic Token Importance Visualization Framework for Interpreting Large Language Model Prompt Understanding Using Perturbation-Based Semantic Deviation Analysis

This paper addresses the challenge of understanding how Large Language Models (LLMs) process and prioritize information from input prompts, a problem often referred to as the "black box" nature of generative AI systems. To improve interpretability without increasing computational overhead, the authors propose a lightweight, model-agnostic token importance visualization technique that does not rely on backpropagation and therefore avoids additional GPU memory consumption and computational cost commonly associated with existing attention visualization methods. The proposed method uses a perturbation-based approach combined with a three-matrix analytical framework to generate relevance maps that illustrate token-level contributions to model predictions. The framework consists of the Angular Deviation Matrix to capture changes in semantic direction, the Magnitude Deviation Matrix to measure shifts in semantic intensity and the Dimensional Importance Matrix to evaluate contributions across individual vector dimensions. By systematically removing tokens and measuring the resulting semantic deviations across these three complementary dimensions, the method computes a composite importance score that provides a mathematically grounded and nuanced measure of token significance. To ensure reproducibility and encourage broader adoption, the authors provide open-source implementations of the proposed explainability techniques along with publicly available code and resources.[74]

Gradient-Controlled Decoding (GCD): A Training-Free Guardrail for Robust Mitigation of Jailbreak and Prompt Injection Attacks in Large Language Models with Deterministic First-Token Safety

The paper presents Gradient-Controlled Decoding (GCD), a training-free safety mechanism designed to mitigate jailbreak and prompt-injection attacks in large language models while reducing over-refusal of benign queries. Building on prior approaches such as GradSafe, which rely on a single acceptance anchor token and suffer from brittle thresholds and lack of guarantees during decoding, GCD introduces a dual-anchor strategy using both acceptance ("Sure") and refusal ("Sorry") tokens to tighten the decision boundary and improve detection accuracy. In cases where a prompt is flagged as unsafe, GCD proactively injects one or two refusal tokens before autoregressive decoding resumes, ensuring deterministic first-token safety regardless of the sampling method used. Experimental results across benchmarks such as ToxicChat, XSTest-v2 and AdvBench demonstrate that GCD reduces false positives by 52% compared to GradSafe at similar recall levels, decreases attack success rates by up to 10% relative to strong decoding-only baselines and introduces minimal latency overhead of approximately 15–20 milliseconds on V100 GPUs. Additionally, the method generalizes across multiple model architectures, including LLaMA-2-7B, Mixtral-8×7B and Qwen-2-7B, while requiring only a small set of demonstration templates, highlighting its efficiency and practical applicability for real-world deployment.[75]

[72] https://arxiv.org/html/2604.11309v1
[73] https://www.marktechpost.com/2026/04/13/google-ai-research-proposes-vantage-an-llm-based-protocol-for-measuring-collaboration-creativity-and-critical-thinking/
[74] https://arxiv.org/abs/2604.02217
[75] https://arxiv.org/pdf/2604.05179v1

New Agentic AI Research

Alibaba Qwen Team Releases Qwen3.5-Omni: A Native Omnimodal AI Model for Unified Text, Audio, Video and Real-Time Interactive Intelligence

The Alibaba Qwen team introduced Qwen3.5-Omni, a native omnimodal large language model designed to process text, images, audio and video simultaneously within a single unified architecture, marking a shift from traditional multimodal systems that combine separate modality-specific models; the model supports real-time interaction with streaming speech output and multimodal reasoning and was trained end-to-end on large-scale audiovisual datasets to enable unified perception and generation across modalities. The Qwen3.5-Omni series includes multiple variants such as Plus, Flash and Light and demonstrates strong performance across audio and audiovisual benchmarks, with capabilities including multilingual speech recognition, video understanding and even emergent behaviors such as generating code from spoken instructions and video input. Overall, the release highlights Alibaba's focus on building fully interactive, real-time multimodal AI systems capable of handling long-context multimodal inputs and complex agent-like tasks within a single integrated model pipeline.[76]

Microsoft Introduces Agent Governance Toolkit: An Open-Source Runtime Security Framework for Governing Autonomous AI Agents

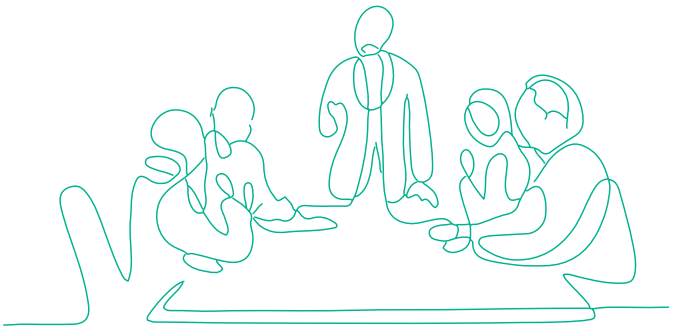
Microsoft, in its April 2, 2026 announcement, introduced the Agent Governance Toolkit, an open-source framework designed to bring runtime security, policy enforcement and governance controls to autonomous AI agents operating in production environments. As AI agents increasingly evolve from passive assistants into autonomous systems capable of executing tasks such as transactions, infrastructure management and decision-making, Microsoft highlights a growing governance gap-where the ease of building agents has outpaced the ability to control their behavior. The toolkit addresses this by acting as a middleware layer that enforces deterministic policies in real time, achieving sub-millisecond latency while monitoring and regulating agent actions. It is specifically engineered to mitigate risks identified in the OWASP Top 10 for Agentic AI, including prompt injection, tool misuse, identity abuse and rogue agent behavior, by intercepting and validating actions before execution. Architecturally, it introduces components such as policy engines, secure agent communication layers, execution isolation mechanisms and reliability engineering practices (e.g., SLOs, circuit breakers and kill switches), thereby embedding production-grade governance into agent systems. The framework is designed to be framework-agnostic, integrating seamlessly with ecosystems like LangChain, CrewAI and Microsoft's own agent platforms without requiring code rewrites and supports multiple programming environments via TypeScript and .NET SDKs. Released under an MIT license, the toolkit reflects Microsoft's broader strategy to establish open, community-driven standards for AI agent governance, emphasizing proactive security, auditability and compliance as foundational requirements for scaling agentic AI safely across enterprise environments.[77]

SafeHarness Strengthens Security for Large Language Model Agents

SafeHarness is a new security approach designed to protect large language model agents by securing the central control layer that manages inputs, decisions, tool usage and memory. Researchers explain that this control layer is critical for system performance but also a major risk, because a single attack at this level can affect the entire system. Existing security methods often fail to see or protect what happens inside this layer or to respond across different stages of operation. SafeHarness addresses this gap by building four security protections directly into the agent lifecycle: filtering risky inputs, verifying decisions, limiting tool access and safely rolling back system state when problems occur. These protections work together and automatically become stricter when suspicious behavior continues, significantly reducing unsafe actions and successful attacks while keeping the system useful.[78]

Memento Skills: A Framework Enabling AI Agents to Rewrite Their Own Skills Without Retraining

Memento-Skills is a novel framework that enables AI agents to continuously improve by rewriting and evolving their own skills without retraining the underlying language model. Instead of modifying model weights, the system uses an external, evolving skill memory combined with a reinforcement learning mechanism called "read-write reflective learning," allowing agents to update, refine and expand their capabilities based on execution feedback. This approach allows agents to autonomously build and reuse skills across tasks, significantly improving performance-achieving up to 80% task success compared to traditional retrieval methods. By decoupling learning from costly retraining, the framework provides a scalable path toward self-improving, production-ready AI agents while maintaining stability through safeguards like automated testing.[79]



[76] https://www.marktechpost.com/2026/03/30/alibaba-qwen-team-releases-qwen3-5-omni-a-native-multimodal-model-for-text-audio-video-and-realtime-interaction/

[77] https://opensource.microsoft.com/blog/2026/04/02/introducing-the-agent-governance-toolkit-open-source-runtime-security-for-ai-agents/

[78] https://arxiv.org/html/2604.13630v1

[79] https://venturebeat.com/orchestration/new-framework-lets-ai-agents-rewrite-their-own-skills-without-retraining-the

Industry Update

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

Healthcare

Multi-Stage Weakly Supervised Validation Framework Enables Scalable and Trustworthy LLM-Based Clinical Information Extraction from Electronic Health Records

Researchers propose a novel multi-stage validation framework designed to enable scalable and reliable deployment of large language models (LLMs) for extracting clinically meaningful information from unstructured health records without relying on annotation-intensive ground truth datasets. The framework combines prompt calibration, rule-based plausibility filtering, semantic grounding checks and targeted confirmatory evaluation using a higher-capacity "judge" LLM, alongside selective expert review and external predictive validity analysis to quantify uncertainty and identify error patterns. Applied to the extraction of substance use disorder (SUD) diagnoses across 11 categories from over 919,000 clinical notes, the system effectively filtered out unsupported or implausible outputs and demonstrated strong agreement between the judge LLM and human experts, while achieving an F1 score of 0.80 under relaxed matching conditions. Additionally, the LLM-derived diagnoses outperformed structured data baselines in predicting future engagement in specialty care, highlighting the framework's ability to deliver accurate, scalable and clinically meaningful insights and establishing a generalizable human-in-the-loop validation approach for real-world healthcare AI deployment.[80]

OpenAI Introduces GPT-Rosalind: A Specialized Life Sciences AI Model Designed to Accelerate Drug Discovery, Genomics Research and Scientific Workflows

OpenAI has launched GPT-Rosalind, its first domain-specific artificial intelligence model tailored for the life sciences sector, with a focus on accelerating complex research processes in areas such as biochemistry, genomics, protein engineering and translational medicine. Named after pioneering scientist Rosalind Franklin, the model is designed to enhance scientific reasoning by supporting tasks like evidence synthesis, hypothesis generation, experimental planning and multi-step research workflows that traditionally require extensive time and expertise. Unlike general-purpose models, GPT-Rosalind is fine-tuned for biological and chemical analysis, enabling it to interpret scientific literature, interact with specialized databases and suggest new experimental pathways within a unified interface. It is integrated with tools such as a Life Sciences plugin for Codex, connecting researchers to over 50 scientific resources and is being deployed through a controlled access program for qualified organizations, including partnerships with major pharmaceutical and research institutions like Amgen and Moderna. The model aims to significantly reduce the time required for early-stage discovery while maintaining strong governance and safety controls, positioning AI as a collaborative tool to augment, rather than replace, human scientists.[81]

Agriculture

PhenoAssistant: A Conversational Multi-Agent AI Framework for Automating and Democratizing Plant Phenotyping Workflows

The article presents PhenoAssistant, an advanced conversational multi-agent AI system developed to streamline and automate plant phenotyping processes, which are typically complex and require specialized expertise. By integrating large language models with multiple analytical tools, the system enables users to execute tasks such as phenotype extraction, data analysis, visualization and model development through simple natural language interactions. The framework coordinates different agents to handle specific subtasks, thereby reducing manual intervention and technical barriers associated with traditional workflows. Experimental evaluations across diverse plant datasets demonstrate that PhenoAssistant maintains strong accuracy while significantly improving efficiency and scalability. Overall, the system enhances accessibility, reproducibility and usability in plant science research, illustrating how agentic AI can transform agricultural analytics and support data-driven biological discoveries.[82]

[80] https://arxiv.org/html/2604.06028v1

[81] https://www.marktechpost.com/2026/04/16/openai-launches-gpt-rosalind-life-sciences-ai/

[82] https://www.nature.com/articles/s41467-026-71090-y

Finance

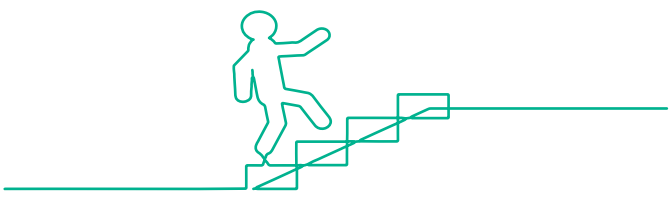
Quantifying Trust: Financial Risk Management Framework for Trustworthy AI Agents

This paper introduces a financial risk–based framework for ensuring trust in AI agents, shifting the focus from purely model-level safety to end-to-end transaction-level guarantees. Instead of relying only on robustness, alignment, or guardrails, the authors propose that AI agents-especially those performing real-world tasks like purchases or decisions-should be governed through risk management principles similar to financial systems. The core contribution is the Agentic Risk Standard (ARS), a structured framework that defines how escrow, collateral, claims and compensation mechanisms should operate when users delegate actions to AI agents. This enables enforceable accountability, meaning if an AI agent fails or causes harm, there are predefined mechanisms for remediation. Overall, the work positions trust in agentic AI as a measurable, enforceable financial contract problem, rather than just a technical reliability issue.[83]

Manufacturing

A Knowledge Graph-Enhanced Large Language Model Framework for Transparent and User-Centric Explainable Artificial Intelligence in Manufacturing Environment

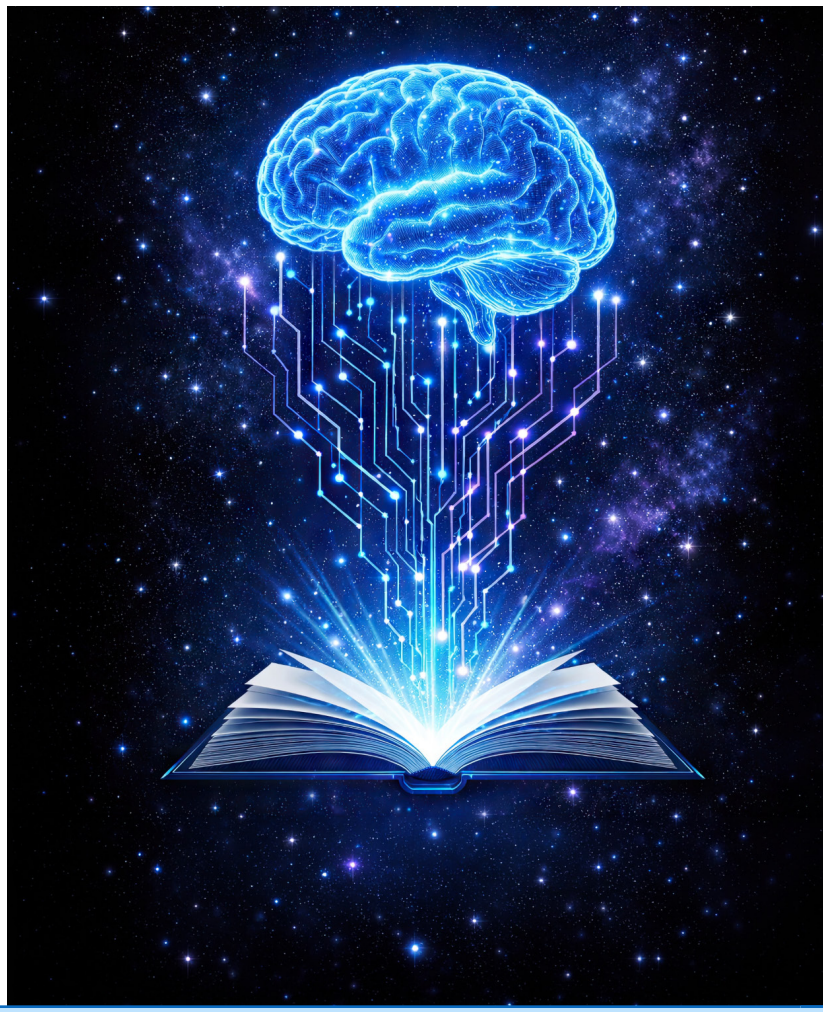
The paper proposes a robust Explainable Artificial Intelligence (XAI) framework that integrates Knowledge Graphs (KGs) with Large Language Models (LLMs) to enhance the interpretability of machine learning outputs within industrial manufacturing contexts. The approach systematically embeds domain-specific knowledge, model predictions and associated explanatory artifacts into a structured knowledge graph, enabling semantic linkage between data, decisions and underlying reasoning. A selective retrieval mechanism is employed to extract contextually relevant knowledge graph triplets, which are subsequently processed by an LLM to generate coherent, human-readable explanations tailored to user queries. The framework is empirically evaluated using domain-driven scenarios and question sets, demonstrating measurable improvements in explanation accuracy, consistency and practical usability for operational decision-making. This work establishes a scalable and hybrid neuro-symbolic paradigm that advances transparency, traceability and trust in enterprise-grade machine learning deployments.[84]



Education

Z.ai Introduces GLM-5.1: Open-Weight 754B Agentic AI Model Achieving State-of-the-Art Coding Performance and Long-Horizon Autonomous Execution

A recent report highlights the release of GLM-5.1 by Z.ai, an advanced open-weight large language model designed for agentic workflows and complex software engineering tasks. Built as an evolution of the GLM-5 architecture, the model features a massive parameter scale and is optimized for coding and multi-step reasoning, achieving state-of-the-art results on benchmarks such as SWE-bench Pro. Notably, GLM-5.1 demonstrates the ability to sustain long-horizon autonomous execution-reportedly running for extended durations while iteratively refining solutions-marking a shift toward fully self-directed AI agents capable of handling complex development pipelines. The model integrates reinforcement learning improvements and tool-use capabilities, enabling it to plan, execute and adapt across extended tasks with minimal human intervention. This release underscores the growing trend of open-weight, high-performance AI systems that compete with leading proprietary models while advancing the frontier of autonomous agentic intelligence.[85]



[83] https://arxiv.org/pdf/2604.03976

[84] https://arxiv.org/abs/2604.16280

[85] https://www.marktechpost.com/2026/04/08/z-ai-introduces-glm-5-1-an-open-weight-754b-agentic-model-that-achieves-sota-on-swe-bench-pro-and-sustains-8-hour-autonomous-execution/

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.

Infosys Responsible AI Releases Efficient Token-Level Explainability Framework for LLMs

Infosys Responsible AI has introduced a lightweight, model-agnostic framework designed to enhance token-level explainability in Large Language Models (LLMs), addressing the longstanding challenge of understanding how models prioritize and process input prompts. The approach leverages perturbation-based analysis to evaluate token importance by systematically observing the impact of token removal on model behavior, eliminating the need for backpropagation and avoiding additional computational overhead typically associated with attention-based methods. At its core, the framework incorporates a three-matrix analytical structure-Angular Deviation Matrix to capture semantic direction shifts, Magnitude Deviation Matrix to measure changes in semantic intensity and Dimensional Importance Matrix to assess embedding-level contributions-enabling a comprehensive and mathematically grounded interpretation of token relevance. By integrating these perspectives into a composite importance score, the solution delivers nuanced insights into generative model outputs while remaining architecture-agnostic and computationally efficient. To support enterprise adoption and transparency, Infosys has also enabled reproducibility through open-source implementations, positioning this innovation as a scalable and practical component within its Responsible AI and observability-driven governance ecosystem.[86]

Events

3rd Annual Trust & Safety Summit | March 24–25, 2026 | London, United Kingdom

The 3rd Annual Trust & Safety Summit, held on 24–25 March 2026 in London, brought together senior global leaders from policy, engineering, moderation, legal, risk and product domains to discuss how trust and safety functions must evolve amid rapid Artificial Intelligence adoption and increasing regulatory scrutiny.

Representing the Infosys Responsible AI Office, Sray Agarwal and Rahul Pareek participated in the summit, with Sray Agarwal delivering a speaking session and presenting the Infosys Threat Simulator, showcasing how organizations can proactively identify and mitigate Artificial Intelligence driven risks. Infosys Business Process Management was represented by Pallavi Nigam, Head of Trust and Safety, Infosys Business Process Management, along with Amit Kumar, Bharath Ravindran and Subhradip Sengupta, who participated in the event. Pallavi Nigam presented a practitioner led narrative highlighting how Responsible Artificial Intelligence defines the "what" while Trust and Safety operationalizes the "how" in building resilient digital platforms. Featuring insights and case studies from leading global platforms such as Uber, Meta, OpenAI, Roblox, Sony Interactive, EA, Lego, Airbnb, Booking.com and Match Group, the summit reinforced trust and safety as a core strategic capability with measurable business and societal impact.

Infosys Responsible AI literacy | March 27, 2026 | Hyderabad

On March 27, 2026, Infosys hosted a Responsible Artificial Intelligence literacy workshop for the Quality Leadership Team at its Hyderabad Special Economic Zone campus. The session was designed as a highly immersive, hands-on experience, enabling participants to engage directly with Responsible Artificial Intelligence concepts through real-time simulations, negotiation-based scenarios, idea pitching and governance decision-making aligned with real client contexts. The strong energy and sustained engagement throughout the workshop reflected the relevance of the approach and content. The session culminated in a clear and unifying message: Responsible Artificial Intelligence is not an additional cost, but a foundational enabler for unlocking sustainable and meaningful Artificial Intelligence value. The initiative was driven in close collaboration with Alok Uniyal and Promod Kasinadhuni, whose leadership and partnership were instrumental in making the workshop impactful.

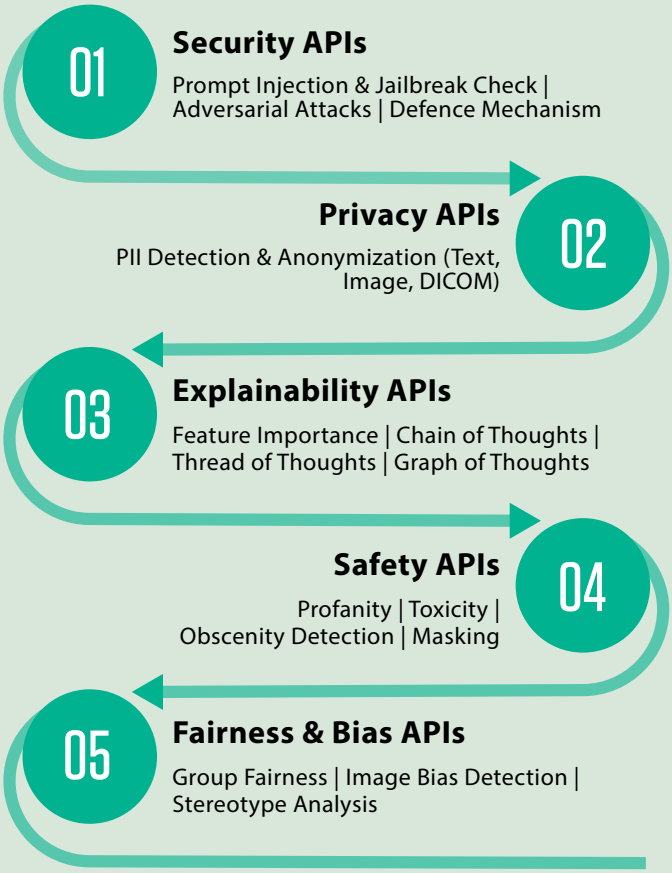
[86] https://arxiv.org/abs/2604.02217

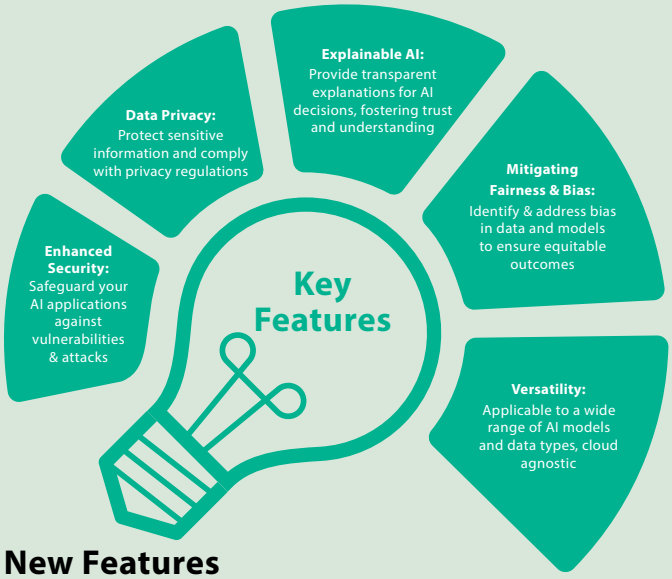
Infosys Responsible AI Toolkit - A Foundation for Ethical AI

The Open-Source Infosys Responsible AI Toolkit can be accessed from its public GitHub repo[87] also as project Salus.[88]

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:





New Features

Below new features are developed and will be available soon in our next release (version 3.0.0).

- Explainability Enhancement Using Reasoning Models
- Second order explainable AI (SOXAI) Technique for Explainability Module
- Multi-lingual support for FM-Moderation Guardrails
- Signature and face masking in Privacy module
- Bulk document safety validation
- Galileo observability: 22 validated safety, fairness, hallucination and security metrics with roadmap extensions for TOON efficient prompting and domain specific evaluation.
- Template-based Guardrails: Extended support to 14 new templates. Example : Restricted Topic Check, Privacy Check, Invisible Check etc.
- Moderation Layer Model based guardrails are improved with finetuned smaller models to improve latency and accuracy.

Infosys Responsible AI Toolkit's Model *based guardrails is now available in* ***HuggingFace,***[89] *both as a repo and also as an interactive playground to explore. Star us and be a part of the Responsible AI Revolution!* ★★★★★

[87] https://github.com/Infosys/Infosys-Responsible-AI-Toolkit
[88] https://github.com/salus-rai/salus
[89] https://huggingface.co/spaces/InfosysEnterprise/Moderation-playground

Infosys offerings for Responsible AI - centred around the AI3s framework

Infosys Responsible AI offerings help organisation enable ethical, trustworthy and scalable AI adoption and is a combination of accelerators and solutions designed to drive responsible AI adoption across enterprises and ensure strong AI Governance, ethics and Security.

SCAN

- Threat Intelligence & Reporting
- Impact Assessment Framework
- Comprehensive monthly market scan report

SHIELD

- Automated AI Management System
- Infosys Responsible AI Toolkit (Opensource)
- Responsible Agent Framework
- Automated Red Teaming

STEER

- AI Governance & Responsible AI CoE Advisory & Consulting
- Standards and Regulations Assessment & Readiness consulting.
- Responsible AI literacy & Capacity Building

"Turn Responsible AI into a Competitive Advantage: leverage our pre-built solutions, frameworks and advisory"

Contributors

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

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

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

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

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

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

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

Ritarshi Chakraborty - Client Solution Lead, 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