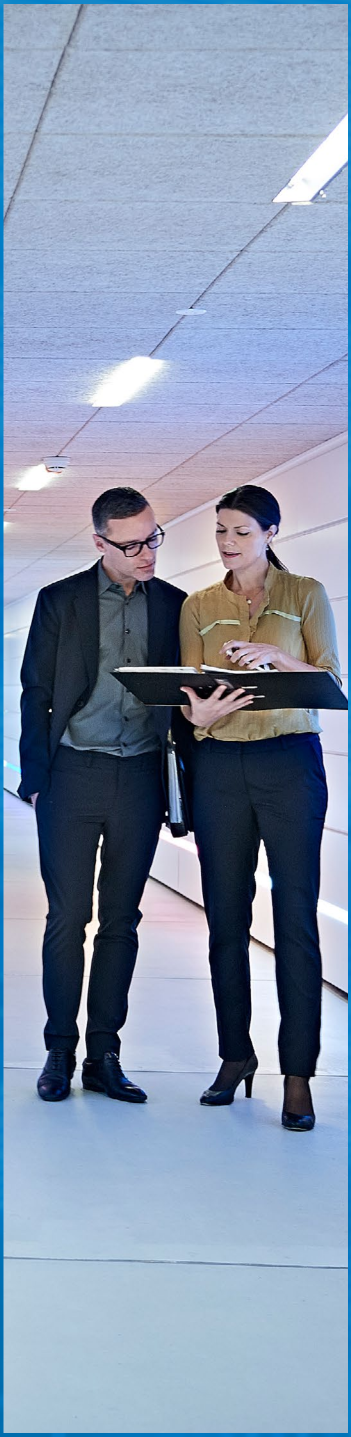
# MARKET SCAN REPORT

# BY INFOSYS RESPONSIBLE AI OFFICE

OCTOBER-DECEMBER 2024

Infosys topaz



Infosys®
Navigate your next

Dear Readers,

As we usher in a new year, we need to take a moment to reflect upon the transformative power of artificial intelligence (AI) and the scorching pace of its evolution. In 2024, we saw AI's impact across industries, reshaping how we live and work. However, with this progress come new challenges in regulation, governance, and ethical considerations that demand our attention. The role of industry leaders, practitioners, and organizations in shaping the future of AI – and that of humankind – has never been more critical.

The latest Market Scan Report from Infosys is a compass to guide stakeholders through the latest developments in AI regulations, governance, incidents, and research techniques. With innovative insights, this report provides valuable information for AI professionals and organizations striving to stay ahead in the game.

- **The Global Regulatory Landscape**

AI regulation has become a focal point for governments worldwide and the push for ethical, transparent, and accountable AI is gathering momentum globally. The Organization for Economic Cooperation and Development (OECD) and the United Nations Educational, Scientific and Cultural Organization (UNESCO) have launched the G7 AI Toolkit to aid public sector entities in adopting safe and ethical AI practices. Meanwhile, the USA is taking steps to address AI-driven challenges with measures such as the AI Civil Rights Act, which tackles bias and civil liberties, and the TAKE IT DOWN Act aimed at combating nonconsensual AI-generated content. The European Commission meanwhile continues to refine its AI regulations, with the AI Act and AI Safety Guidelines gaining greater traction.

On a more significant note, India is positioning itself as a leader in AI safety through initiatives like the IndiaAI Mission and a proposed AI Safety Institute to ensure alignment with global standards. Further, the AI Safety Institutes recently set up in South Korea, Canada, and the US exemplify the international commitment to responsible AI development.

- **Incidents and Ethical Challenges**

As AI technologies permeate various sectors, several major AI incidents have sparked concerns about safety, ethics, and accountability. Meta is at the center of a new copyright lawsuit for its AI training practices, while South Korea imposed a $15 million fine on the company over privacy violations. In addition, there have been incidents involving deepfakes, such as the impersonation of a Ukrainian official targeting a U.S. senator, and an AI-generated deepfake interview of a presidential candidate in Uruguay that raised ethical concerns. There have also been significant issues surrounding AI's impact on mental health, with an AI chatbot advising a teen to harm their parents and concerns over AI-generated errors affecting Alaska's education policy. Other notable incidents include biased algorithms used by the French government, hackers exploiting AI repositories, and OpenAI's Whisper tool facing scrutiny due to transcription errors in medical settings. Various lawsuits and regulatory actions, such as the one from the Canadian News Media Companies against OpenAI and concerns over privacy breaches related to facial recognition, highlight the growing scrutiny of AI technologies and their potential risks.

- **Research Innovations and New Models**

The advancements in AI research and new model development continue to shape the industry. Companies like Meta have introduced models like LLaMA 3.2 while NVIDIA and IBM have unveiled groundbreaking AI tools that aim to enhance robotic interactions and geospatial applications. These innovations will prove pivotal in pushing the boundaries of AI capabilities across healthcare, robotics, and geospatial engineering.

Research institutions are also making headway in enhancing AI safety. Some examples include Intel AI Research's new frameworks, Google DeepMind's relaxed robotic transformers (RRTs) for small AI models, and Anthropic's Message Batches API that improves the efficiency and security of AI systems.

- **Infosys: Leading with Responsible AI**

At Infosys, we are committed to driving the responsible and ethical use of AI. Our ongoing efforts in Responsible AI are exemplified by initiatives like the Responsible AI Toolkit, which will be  available as an open-source solution very soon, and our small language models, designed with ethical considerations at their core. Our partnerships with global organizations including Cisco and the University of Cambridge reinforce our vision to build a sustainable and transparent AI ecosystem. We believe that by embedding responsible AI practices in our solutions, we can help organizations navigate the AI revolution in a way that benefits business as well as the world at large.

## The Path Forward

As AI evolves rapidly, it is crucial for practitioners, leaders, and organizations to stay informed and adapt to new regulatory and ethical standards. The Infosys Market Scan Report provides key insights to help stakeholders lead responsibly in AI development. While AI holds immense potential, it is our shared responsibility to ensure that its impact is positive and sustainable. We invite you to explore the full report and join us in shaping a future where AI serves humanity with trust, fairness, and accountability.

Wishing you a prosperous and responsible New Year ahead, filled with new opportunities to lead in the development of Responsible AI and to make meaningful contributions to society.

**Warm regards**

**Syed Ahmed**
Global Head- Infosys Responsible AI Office

# AI Regulations, Governance and Standards

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

## AI Regulations and Governance across globe

## Global

### OECD and UNESCO Release G7 AI Toolkit for Public Sector

The Organisation for Economic Co-operation and Development (OECD) and the United Nations Educational, Scientific and Cultural Organization (UNESCO) released the G7 Toolkit for Artificial Intelligence in the Public Sector on October 15, 2024. This toolkit aims to assist G7 countries in safely and securely integrating AI into public sector operations. It offers a detailed guide on best practices, governance frameworks, and policy options to ensure AI is used in a trustworthy and effective manner within the public sector.[1]

### U.S. and UAE Forge Partnership to Promote Safe and Ethical AI Development

The United States and the United Arab Emirates (UAE) have announced a new cooperation initiative on artificial intelligence (AI), endorsed by U.S. National Security Advisor Jake Sullivan and UAE National Security Advisor Sheikh Tahnoon bin Zayed Al Nahyan. This partnership aims to advance safe, secure, and trustworthy AI by promoting international AI frameworks and standards, aligning regulatory frameworks to foster innovation while safeguarding national security, and prioritizing ethical AI research to address bias and discrimination in algorithms. Additionally, the collaboration will focus on creating a secure cybersecurity environment to protect critical infrastructure and support bilateral investments and the development of robust AI infrastructure. This initiative underscores the commitment of both nations to harness AI's potential for economic growth, job creation, and environmental sustainability while addressing the challenges and risks associated with this emerging technology.[2]

### Building Trust in AI: UK and Singapore Unite for Safety Standards

On November 6, 2024, the United Kingdom and Singapore signed a significant Memorandum of Cooperation to enhance the safety and reliability of artificial intelligence (AI) technologies. This agreement was deemed essential due to the rapid advancement of AI and the growing concerns about its ethical implications and potential risks. By fostering greater public trust in AI, the partnership aims to address these challenges collaboratively. The agreement will facilitate the sharing of best practices, research, and standards in AI safety, ensuring that both nations can effectively manage the risks associated with AI deployment. This cooperation is expected to unlock an estimated £6.5 billion in economic benefits over the next decade, promoting innovation while safeguarding public interests. By working together, the UK and Singapore hope to set a global benchmark for responsible AI governance, ultimately benefiting users and industries worldwide.[3]

### UN First Committee Approves Key Drafts on AI and Military Implications

In a significant move towards addressing global security concerns, the United Nations General Assembly's First Committee approved fourteen new draft resolutions on November 6, 2024. Among these, a key draft focuses on the implications of artificial intelligence (AI) in the military domain, highlighting potential risks such as arms races and the escalation of conflicts. The resolutions encourage member states to collaborate on ethical AI governance, emphasizing

[1] https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/10/g7-toolkit-for-artificial-intelligence-in-the-public-sector_f93fb9fb/421c1244-en.pdf

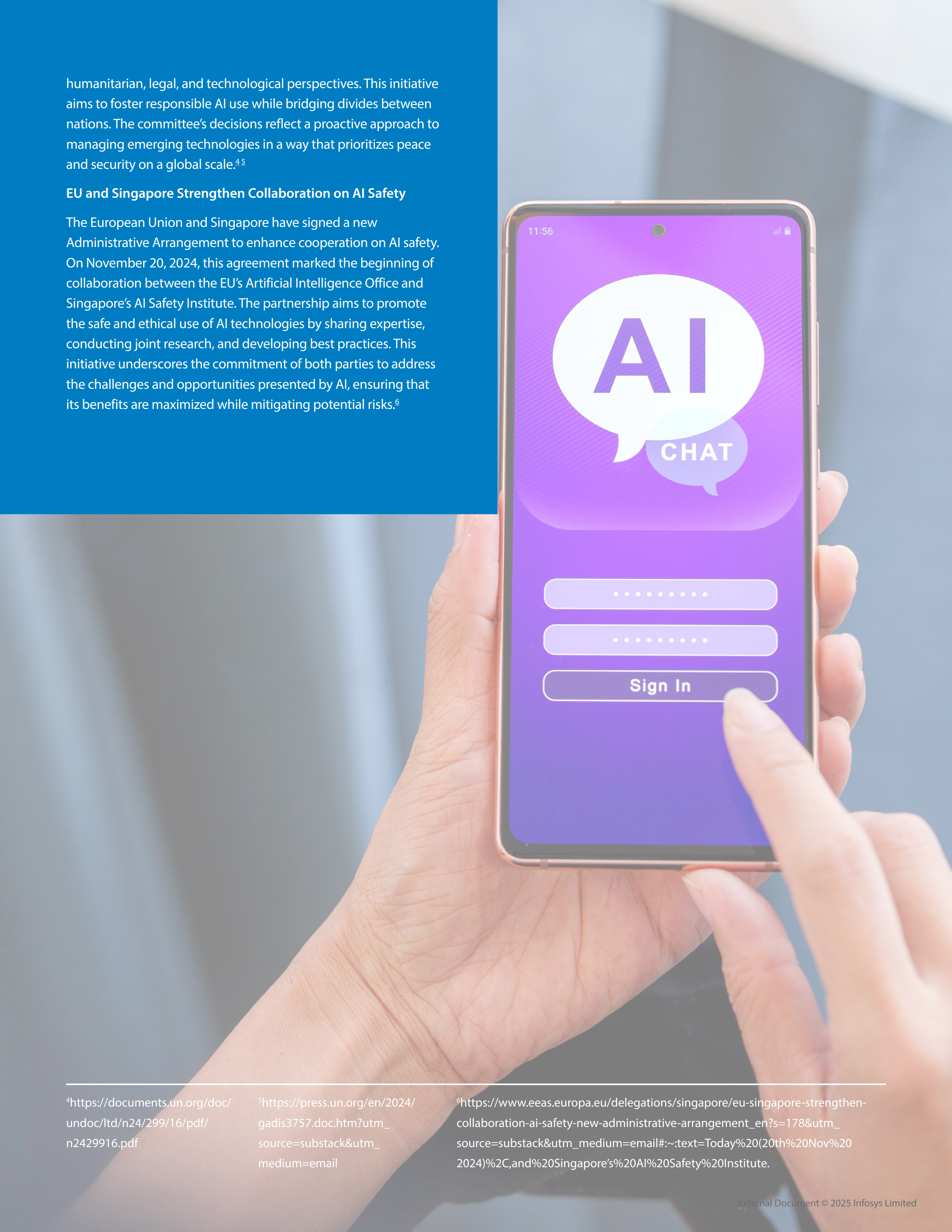
[2] https://www.whitehouse.gov/briefing-room/statements-releases/2024/09/23/united-states-and-united-arab-emirates-cooperation-on-artificial-intelligence/

[3] https://www.gov.uk/government/news/ensuring-trust-in-ai-to-unlock-65-billion-over-next-decade?utm_source=substack&utm_medium=email

humanitarian, legal, and technological perspectives. This initiative aims to foster responsible AI use while bridging divides between nations. The committee's decisions reflect a proactive approach to managing emerging technologies in a way that prioritizes peace and security on a global scale.[4] [5]

### EU and Singapore Strengthen Collaboration on AI Safety

The European Union and Singapore have signed a new Administrative Arrangement to enhance cooperation on AI safety. On November 20, 2024, this agreement marked the beginning of collaboration between the EU's Artificial Intelligence Office and Singapore's AI Safety Institute. The partnership aims to promote the safe and ethical use of AI technologies by sharing expertise, conducting joint research, and developing best practices. This initiative underscores the commitment of both parties to address the challenges and opportunities presented by AI, ensuring that its benefits are maximized while mitigating potential risks.[6]



[4] https://documents.un.org/doc/undoc/ltd/n24/299/16/pdf/n2429916.pdf

[5] https://press.un.org/en/2024/gadis3757.doc.htm?utm_source=substack&utm_medium=email

[6] https://www.eeas.europa.eu/delegations/singapore/eu-singapore-strengthen-collaboration-ai-safety-new-administrative-arrangement_en?s=178&utm_source=substack&utm_medium=email#:~:text=Today%20(20th%20Nov%202024)%2C,and%20Singapore's%20AI%20Safety%20Institute.

# US

## Federal

### AI Civil Rights Act introduced to Combat Bias and Protect Civil Liberties

A new federal bill, the AI Civil Rights Act, has been introduced to address and eliminate bias in artificial intelligence (AI) systems. This legislation aims to place strict regulations on the use of algorithms in decisions that significantly impact people's rights, civil liberties, and livelihoods, such as employment, banking, healthcare, criminal justice, public accommodations, and government services.

Key provisions of the bill include:

- **Prohibiting discriminatory algorithms:** Developers and users of AI systems must ensure their algorithms do not discriminate based on protected characteristics or cause disparate impacts.
- **Mandatory evaluations and regular monitoring:** AI systems must undergo independent audits before and after deployment to identify and mitigate potential biases. Continuous monitoring of AI systems is required to ensure they remain compliant with anti-discrimination standards. Regular reports on the performance and fairness of these systems must be submitted to relevant authorities.
- **Transparency and accountability:** The bill seeks to renew public trust in AI by ensuring the accuracy and fairness of these systems.

The bill is expected to have a significant impact by protecting marginalized communities from biased AI decisions, promoting fairness, and ensuring that technological advancements do not exacerbate existing inequalities. This initiative highlights the importance of safeguarding civil rights in the age of AI.[7]

This bill has been introduced to Senate Committee on Commerce, Science, and Transportation and currently under review.[8]

### White House Pushes for AI Leadership in National Security

On October 24, 2024, in a memorandum, the White House has outlined a strategic initiative to bolster the United States' leadership in artificial intelligence (AI) by integrating AI technologies into national security frameworks. This move aims to enhance the country's competitive edge and safeguard national interests while ensuring the ethical and secure deployment of AI. The memorandum underscores the importance of protecting human rights, civil liberties, and privacy in AI applications, setting a robust framework for responsible AI development. This initiative is expected to significantly impact national security operations, ensuring that AI advancements contribute to the nation's safety and strategic objectives.[9]

### Senator Introduces Legislation to Protect Creators from Unauthorized AI Training

A new federal bill, the Transparency and Responsibility for Artificial Intelligence Networks (TRAIN) Act, has been introduced to protect musicians, artists, and creators from the unauthorized use of their copyrighted works in AI training. This legislation is significant as it tackles the issue of transparency in AI systems, allowing creators to verify if their work has been used without permission. By promoting accountability and ensuring creators can safeguard their intellectual property, the TRAIN Act aims to create a fairer environment in the rapidly evolving field of artificial intelligence. The bill was introduced on November 25, 2024, and is currently under consideration in the Senate Judiciary Committee. The timeline for its enactment will depend on the legislative process, including committee reviews and potential amendments.[10]

[7] https://www.markey.senate.gov/news/press-releases/senator-markey-introduces-ai-civil-rights-act-to-eliminate-ai-bias-enact-guardrails-on-use-of-algorithms-in-decisions-impacting-peoples-rights-civil-liberties-livelihoods

[8] https://www.govtrack.us/congress/bills/118/s5152/text

[9] https://www.whitehouse.gov/briefing-room/presidential-actions/2024/10/24/memorandum-on-advancing-the-united-states-leadership-in-artificial-intelligence-harnessing-artificial-intelligence-to-fulfill-national-security-objectives-and-fostering-the-safety-security/?utm_source=substack&utm_medium=email

[10] https://www.welch.senate.gov/welch-leads-bill-to-protect-musicians-artists-and-creators-from-unauthorized-a-i-training/?utm_source=substack&utm_medium=email

*Several other countries, including Spain, Canada, the European Union, Japan, South Korea, and Australia, have introduced their own copyright laws to protect against the unauthorized use of their work in AI training. The Infosys Responsible AI Office is currently assessing these international bills, focusing on benchmarking their effectiveness and ensuring they meet exacting standards. This evaluation aims to identify best practices and integrate them into Infosys' own AI governance framework, promoting a global standard for responsible AI use.*

### U.S. House Introduces AI Fraud Deterrence Act to Combat AI-Driven Financial Crimes

The U.S. House of Representatives has introduced the AI Fraud Deterrence Act H.R.10125, a federal bill aimed at combating financial crimes committed using artificial intelligence. Sponsored by Representatives Ted Lieu and Kevin Kiley, this bipartisan legislation seeks to increase penalties for crimes such as mail fraud, wire fraud, bank fraud, and money laundering when AI technology is involved. The bill proposes harsher fines and longer prison sentences for offenders, thereby aiming to deter the misuse of AI in fraudulent activities. This legislation will impact individuals and organizations that engage in AI-driven financial crimes, enhancing protections for consumers and financial institutions. The bill was introduced in the House of Representatives on November 14, 2024. It is currently under review by the House Committee on the Judiciary.[11]

### House Financial Services Committee Introduces AI Regulation Measures for Financial and Housing Sectors

On 2nd December 2024 the House Financial Services Committee, a key committee of the U.S. House of Representatives overseeing the financial services industry, has introduced a resolution and a bill to address the impact of artificial intelligence (AI) on the financial and housing sectors. The resolution outlines the committee's responsibilities, including the use of AI in underwriting and tenant screening in the housing market, and the influence of AI on market behaviours by financial institutions. The bill calls on financial regulatory agencies to evaluate the benefits and risks of AI, ensuring they have the necessary tools for oversight. This initiative aims to ensure that AI's integration into financial services and housing benefits consumers, firms, and regulators, while addressing data privacy concerns and maintaining the U.S.'s leadership in AI development. The bill is currently being reviewed by the House Financial Services Committee. The committee is examining the bill's details and may suggest changes. Once this review is complete, the bill will move to the House floor for a vote.[12]

### US Senate Passes TAKE IT DOWN Act to Combat Nonconsensual AI-Generated Content

On 3rd December 2024, the US Senate unanimously passed the "Tools to Address Known Exploitation by Immobilizing Technological Deepfakes on Websites and Networks Act" (TAKE IT DOWN Act), sponsored by Senator Ted Cruz (R-Texas). This legislation mandates that platforms remove nonconsensual intimate visual depictions (NCII) and addresses several key areas. It criminalizes the publication or threat to publish NCII, including realistic computer-generated pornographic images, clarifying that consent to create an image does not imply consent to its publication. The act also protects good faith efforts to assist victims by allowing disclosure of NCII to law enforcement or for medical treatment. Additionally, it requires websites to remove NCII within 48 hours of notice from the victim and to make reasonable efforts to remove copies, with the Federal Trade Commission (FTC) responsible for enforcement. The bill also protects lawful speech by narrowly targeting the criminalization of knowingly publishing NCII while adhering to First Amendment standards, requiring that computer-generated NCII pass a "reasonable person" test for realistic depiction. The bill now moves to the House of Representatives for approval. For it to become law, the bill must now be considered and approved by the House of Representatives. If it passes in the House, it will be sent to the President for signature. Once signed by the President, the bill will officially become law.[13]

[11] https://www.congress.gov/bill/118th-congress/house-bill/10125/text?s=1&r=6

[12] https://fedscoop.com/house-financial-services-committee-ai-housing-bill/?utm_source=substack&utm_medium=email

[13] https://www.congress.gov/bill/118th-congress/senate-bill/4569/text

# California

## Advancing Ethical AI: Landmark Legislation to Combat AI Risks in California, USA

The State of California has introduced a comprehensive legislative package to advance the safe and responsible development of artificial intelligence (AI) technologies. These initiatives aim to protect Californians from the potential risks associated with AI while promoting innovation and transparency. Key measures include mandatory safety protocols, transparency requirements, and protections against AI-generated misinformation and deepfakes. Here are few prominent ones:

**AB 2013 - Generative artificial intelligence: training data transparency:** Requires generative AI companies to disclose information about their training data and ensure AI-generated content is clearly marked, promoting transparency and trust in AI technologies. It will be in effect from 1st January 2026.

**SB 926 - Crimes: distribution of intimate images:** Criminalizes the creation or distribution of non-consensual, AI-generated intimate images, addressing the emotional harm caused by deepfake content. It will go into effect from 1st January 2025.

**SB 942 - California AI Transparency Act:** Aims to enhance consumer protection by mandating transparency in the use of generative artificial intelligence. The bill requires businesses to disclose when AI is used to create content that interacts with consumers, ensuring that consumers are aware they are engaging with AI-generated material. This will be in effect from 1st January 2026.

**SB 981 - Sexually explicit digital images:** Requires social media platforms to establish reporting mechanisms for users to flag and remove AI-generated sexually explicit images that are distributed without consent. This will go into effect from January 1, 2025.

These bills collectively aim to ensure that AI technologies are developed and used responsibly, safeguarding the interests of all Californians.[14]

## California Governor Blocks Landmark AI Safety Bill SB 1047, Sparking Debate on Innovation and Security

The governor of California recently blocked a significant AI safety bill aimed at regulating generative AI development, citing concerns that it could hinder innovation and economic growth. The bill, known as the Safe and Secure Innovation for Frontier Artificial Intelligence Models Act (SB 1047), was intended to impose stringent safety measures on advanced AI systems. Critics argue that without such regulations, AI technologies could pose serious risks, including spreading misinformation and enabling sophisticated cyberattacks. The decision reflects a broader debate on balancing innovation with safety, especially in a state that is home to many leading AI companies.[15]

As of 22nd October 2024, the bill is in the Senate, awaiting action on the Governor's veto.[16]

[14] https://www.gov.ca.gov/2024/09/19/governor-newsom-signs-bills-to-crack-down-on-sexually-explicit-deepfakes-require-ai-watermarking/

[15] https://theconversation.com/californias-governor-blocked-landmark-ai-safety-laws-heres-why-its-such-a-key-ruling-for-the-future-of-ai-worldwide-240182

[16] https://leginfo.legislature.ca.gov/faces/billStatusClient.xhtml?bill_id=202320240SB1047

## UK

### Comprehensive Overview of the Government Communication Service's Generative AI Policy

The Government Communication Service (GCS) UK released its Generative AI Policy to the public on September 20, 2024, as part of its Innovating with Impact Strategy. The GCS Generative AI Policy outlines several key principles for the responsible use of generative AI in government communications:

1. **Adherence to Official Guidance:** Always use generative AI in line with the latest government guidelines, including the Generative AI framework for HM Government and the GCS Ethical Framework for Responsible Innovation.
2. **Ethical Standards:** Operate consistently with values such as Democracy, Rule of Law, and Individual Liberty, and adhere to the Civil Service Code and Government Publicity Conventions.
3. **Training and Education:** The central GCS team at the Cabinet Office will provide training on responsible AI use, focusing on ensuring accuracy, inclusivity, and mitigating biases.
4. **Supplier Compliance:** Require all contracted and framework suppliers to adhere to the GCS policy on responsible AI use and have safeguards in place.
5. **Upholding Factuality:** Ensure that generative AI is used to create accurate content and not to alter existing digital content in a misleading way.

The policy aims to empower GCS members( public service communicators who work in government departments, agencies, and other public bodies) to utilize innovative AI tools effectively, improving the quality and consistency of government messaging, reducing content creation time and resources, and delivering more engaging and relevant communications to the public.[17]

[17] https://gcs.civilservice.gov.uk/publications/gcs-generative-ai-policy/

## Europe

### AI Office Organized First Workshop on AI Code of Practice

On October 23, 2024, the AI Office held the first workshop with general-purpose AI model providers and the Chairs and Vice-Chairs to draft the General-Purpose AI Code of Practice. This workshop was crucial as it allowed direct discussions on systemic risk assessment, technical mitigation, governance, transparency, and copyright-related rules. The importance of this event lay in its role in shaping the future development and governance of trustworthy AI in Europe, ensuring compliance with the AI Act's provisions. The impact of this workshop was significant, as it set the stage for the creation of a comprehensive code of practice that would guide AI model providers in adhering to regulatory standards, thereby fostering a safer and more transparent AI ecosystem.[18]

The EU AI Pact and the General-Purpose AI Code of Practice are initiatives aimed at ensuring the safe and trustworthy development of AI systems in the EU. The AI Pact is a voluntary initiative that helps organizations prepare for the AI Act by sharing best practices and encouraging proactive compliance pledges. In contrast, the Code of Practice provides detailed guidelines for providers of general-purpose AI models to comply with the AI Act, focusing on transparency, risk mitigation, and adherence to copyright rules. Together, they complement each other by fostering a collaborative and compliant AI ecosystem in the EU.[19]

*At the EU's #AIPact launch, Infosys reaffirmed its commitment to ethical AI deployment as a founding signatory, alongside over one hundred global businesses*

### European Commission's Joint Research Centre releases policy brief on harmonised standards for EU AI Act

The joint research centre of European commission has released harmonised standards for the European AI Act, which aims to ensure the safe and effective implementation of AI technologies across the EU. These standards are crucial as they provide a legal presumption of conformity with the AI Act, helping to create a level playing field for AI developers, especially small and medium-sized enterprises. By addressing the unique risks AI poses to health, safety, and fundamental rights, these standards will enhance trust and accountability in AI systems. The impact of these standards is significant, as they will guide the design, development, and oversight of AI technologies, ensuring they meet stringent safety and ethical requirements.[20]

[18] https://digital-strategy.ec.europa.eu/en/news/first-workshop-general-purpose-ai-model-providers-code-practice-drafting-process

[19] https://digital-strategy.ec.europa.eu/en/policies/ai-pact

[20] https://publications.jrc.ec.europa.eu/repository/handle/JRC139430

# Australia

## OAIC Releases Privacy Guidelines for Safe Adoption of Commercial AI Products in Australia

The Office of the Australian Information Commissioner's  (OAIC) new guidance on privacy for using commercially available AI products is directed at organizations and government agencies in Australia. These entities must comply with the Privacy Act by performing due diligence when adopting AI products, ensuring human oversight, and updating privacy policies to reflect AI usage. They should avoid inputting personal information into AI systems without proper consent and ensure that any generated data is managed according to privacy laws. This regulation impacts these organizations by requiring transparent and responsible AI practices, thereby protecting individuals' privacy and minimizing regulatory risks.

Following are the key points which need to be followed:

- **Privacy Obligations:** Organizations must comply with privacy laws for any personal information input into AI systems and the output generated, ensuring lawful and fair handling.
- **Due Diligence:** Before adopting AI products, organizations should assess their suitability, including testing, human oversight, and understanding privacy risks and access controls.
- **Transparency in Policies:** Businesses need to update privacy policies to clearly inform users about AI usage, especially for public-facing tools like chatbots, to ensure transparency and good governance.
- **Compliance with APPs:** If AI generates or infers personal information, it must comply with the Australian Privacy Principles (APPs), ensuring that such data collection is necessary and lawful.
- **Best Practices:** Organizations are advised against entering personal or sensitive information into publicly available AI tools due to significant privacy risks.[21]

## OAIC Issues New Privacy Guidelines for Responsible AI Development in Australia

The Office of the Australian Information Commissioner's (OAIC) new guidance on privacy for developing and training generative AI models is aimed at developers and organizations in Australia. These entities must ensure data accuracy, obtain explicit consent, and handle sensitive information with care. The regulation requires them to integrate privacy considerations from the beginning, use high-quality datasets, and comply with privacy laws, even when using publicly available data. By following these guidelines, developers can mitigate regulatory risks and ensure their AI models are responsibly developed, thereby protecting individuals' privacy rights.

Following are the key principles which need to be followed:and minimizing regulatory risks.

- Developers need to ensure their AI models are accurate by using high-quality data and proper testing, especially given the higher risks in AI. They should use disclaimers to highlight when AI models might need extra caution and safeguards for high-risk privacy uses.
- Publicly available data cannot always be used for AI training. Developers must check if the data includes personal information and follow privacy laws, which might mean deleting some data to comply.

[21] https://www.oaic.gov.au/privacy/privacy-guidance-for-organisations-and-government-agencies/guidance-on-privacy-and-the-use-of-commercially-available-ai-products?utm_source=substack&utm_medium=email

- Sensitive information, like photos or recordings of people, usually requires consent to be used. This data cannot be scraped from the web or collected from third-party sources without consent.
- If developers want to use personal information, they already have for AI training, and it was not the original reason for collecting it, they need to ensure they have consent or that the use is expected by the individual and related to the original purpose.
- If developers cannot clearly show that using data for AI is within reasonable expectations and related to the original purpose, they should get consent or allow individuals to opt-out to avoid regulatory issues.[22]

### Australian Government Abandons Proposed Legislation to Regulate Online Misinformation

The Australian federal government has decided to abandon its proposed (January 2023) legislation aimed at regulating misinformation on social media platforms. The bill, which intended to grant the Australian Communications Media Authority the power to set rules for removing harmful content, faced substantial opposition from the Coalition, Greens, and several crossbench senators, ultimately lacking the necessary support for passage. Communications Minister Michelle Rowland confirmed the withdrawal, citing insufficient backing in the Senate. The legislation aimed to address seriously harmful content, including misinformation from foreign actors and anti-vaccine propaganda, while preserving freedom of speech protections. Opposition Leader Peter Dutton criticized the bill as an attempt to censor free speech, celebrating its withdrawal as a triumph for democracy.[23]

### Australian Government Introduces Landmark Legislation to Strengthen Privacy Protections

The Australian Government has enacted the Privacy and Other Legislation Amendment Bill 2024, a significant step towards enhancing privacy protections for Australians. This landmark legislation introduces a new statutory tort for serious invasions of privacy and a Children's Online Privacy Code to safeguard children from online harms. It also grants stronger enforcement powers to the Australian Information Commissioner and criminalizes doxing, with penalties of up to seven years imprisonment. These measures are crucial in the digital age, ensuring that Australians' personal data is protected and providing greater transparency and control over automated decisions.

On November 29, 2024, the bill passed both Houses of Parliament and was signed into law. This landmark legislation introduces significant privacy protections, including a new statutory tort for serious invasions of privacy, a Children's Online Privacy Code, and stronger enforcement powers for the Australian Information Commissioner. [24]

## India

### India Plans AI Safety Institute to Align with Global Standards

India is planning to establish an Artificial Intelligence Safety Institute (AISI) to set standards and frameworks for AI development, mirroring efforts in the UK, US, and Japan. The UK's institute emphasizes enforcement, while the US focuses on setting standards. India's initiative aims to ensure AI safety without hindering innovation, aligning with global trends to create a network of AI safety institutes that share best practices and foster international cooperation. This underscores India's commitment to responsible AI advancement, in line with international efforts.[25]

> *While it's recommended to avoid monopolizing decisions regarding content safety, we must not underestimate the importance of regulating and addressing these issues with clear guidelines and effective solutions.*

[22] https://www.oaic.gov.au/privacy/privacy-guidance-for-organisations-and-government-agencies/guidance-on-privacy-and-developing-and-training-generative-ai-models?utm_source=substack&utm_medium=email

[23] https://www.abc.net.au/news/2024-11-24/laws-to-regulate-misinformation-online-abandoned/104640488

[24] https://ministers.ag.gov.au/media-centre/delivering-stronger-privacy-protections-australians-29-11-2024?utm_source=substack&utm_medium=email

[25] https://www.hindustantimes.com/india-news/govt-mulls-setting-up-artificial-intelligence-safety-institute-101728833433153.html

### IndiaAI Mission: Advancing Safe and Trusted AI Development through Eight Key Projects

The IndiaAI Mission has selected eight projects under its Safe and Trusted AI Pillar to advance responsible AI development. These projects focus on critical themes such as Machine Unlearning, Synthetic Data Generation, AI Bias Mitigation, Ethical AI Frameworks, Privacy-Enhancing Tools, Explainable AI, AI Governance Testing, and Algorithm Auditing Tools. The initiative aims to develop indigenous tools and frameworks to ensure ethical, transparent, and trustworthy AI technologies. This effort is crucial for addressing India's unique challenges and opportunities, promoting responsible AI use, and safeguarding citizens' rights and privacy.[26]

> *Infosys is committed to supporting India's AI mission by advancing the vision of safe and trusted AI. Through our industry-leading insights on responsible AI, we aim to contribute to the development of ethical AI systems that foster trust and benefit society at large. Our open-source toolkit and initiatives are designed to empower organizations, startups, and academia to adopt AI in a responsible and impactful manner*

### Indian Government Considers New AI Regulation Law, Seeks Broad Consensus

The Indian government has expressed openness to introducing new legislation to regulate artificial intelligence (AI). However, it emphasized that achieving a broad consensus is crucial before moving forward. The government highlighted the importance of establishing accountability and adapting the legal framework to address the challenges posed by AI, including the spread of fake narratives. The government is also committed to democratizing technology and has been actively supporting the development of AI data labs in smaller cities, with a significant number of candidates enrolling in future skill programs.[27]

### GPAI Summit 2024: India Leads with Vision for Ethical AI

On December 9, 2024, at the GPAI Summit 2024 in Belgrade, Serbia, India presented its vision for AI, emphasizing ethical development, inclusivity for the Global South, and robust governance. The vision, part of the IndiaAI Mission, focuses on democratizing AI access, improving data quality, and fostering indigenous capabilities with a $1.2 billion investment. Additionally, India highlighted its leadership in advancing ethical AI through platforms like GPAI and the G20 Framework on Digital Public Infrastructure, ensuring AI development aligns with societal needs and democratic principles.[28]

[26] https://pib.gov.in/PressReleasePage.aspx?PRID=2065579

[27] https://www.thehindu.com/news/national/open-to-bringing-new-law-to-regulate-ai-but-lot-of-consensus-required-vaishnaw-in-parliament/article68972355.ece?utm_source=substack&utm_medium=email

[28] https://indiaai.gov.in/article/india-s-vision-for-ai-shri-jitin-prasada-at-the-gpai-summit-2024?utm_source=newsletter&utm_medium=email&utm_campaign=The%20Heuristic%20from%20IndiaAI

## New Zealand

### New Zealand Joins UK's Bletchley Declaration to Enhance AI Safety and Innovation

New Zealand, represented by Minister Judith Collins, has joined the UK's Bletchley Declaration on AI Safety. This initiative aims to promote responsible AI development and usage, ensuring that AI technologies are safe, ethical, and beneficial. By participating in this international effort, New Zealand seeks to enhance its AI governance framework, support innovation, and boost economic development. This collaboration will help address global AI challenges, ensuring that AI systems align with human rights and democratic values, benefiting New Zealand's productivity and societal well-being.[29]

## Germany

### Germany's Data Protection Authorities Establish AI Working Group

The Conference of the Independent Data Protection Supervisory Authorities of the Federal and State Governments of Germany (DSK) has announced the formation of an AI Working Group. This initiative aims to develop requirements and recommendations to ensure AI systems comply with data protection regulations. Additionally, the group will consolidate technical and legal expertise from DSK's supervisory authorities to effectively monitor AI technologies.[30]

[29] https://www.beehive.govt.nz/release/nz-joins-uk-initiative-ai-safety

[30] https://www.datenschutz.sachsen.de/detailseite-news-bzw-veranstaltungsmeldung-7311-7311.html?utm_source=substack&utm_medium=email

## Spain

### Spain's New Copyright Regulations: A Turning Point for AI Model Training

Spain has introduced new regulations requiring AI developers to obtain explicit consent from copyright holders before using their content for training AI models, marking a significant development in the protection of intellectual property rights. This legislative change aims to ensure that content creators are compensated, potentially setting a precedent for other countries. The new rules could pose additional legal challenges and costs for AI developers, highlighting the growing tension between technological advancement and the rights of content creators. This move underscores the need for ethical and legal considerations in AI development, as governments worldwide increasingly seek to regulate the use of copyrighted content in AI.[31]

## Japan

### Japan Patent Office Plans Legal Revisions to Address AI and Digital Technology Challenges

The Japan Patent Office (JPO) is considering revisions to its Patent and Design Laws to address the challenges posed by digital technologies and artificial intelligence. These changes aim to prevent intellectual property infringement resulting from AI-generated designs and virtual spaces. Currently, the laws do not adequately cover materials produced by AI, leading to potential misuse and hindrance in product development. The proposed revisions, expected to be implemented next fiscal year, would prevent misuse of generative AI in obtaining patent or design rights and regulate unauthorized AI learning of existing products to create new designs. Additionally, the revisions would address issues related to virtual spaces, where design rights currently do not apply, allowing unlicensed reproduction of brand-name items in the metaverse. These updates are part of a broader effort to ensure Japan's intellectual property framework keeps pace with technological advancements.[32]

[31] https://www.mlex.com/mlex/articles/2265137/right-holders-in-ai-model-training-see-new-front-opening-in-spain?utm_source=substack&utm_medium=email

[32] https://asianews.network/japan-patent-office-mulls-revising-laws-to-cope-with-digital-tech-ai/

## Denmark

### Denmark's Strategic Blueprint for EU AI Act Compliance: A Collaborative Approach with Microsoft

Denmark has introduced a comprehensive blueprint for compliance with the EU AI Act, supported by Microsoft, to ensure the ethical and safe use of artificial intelligence across Europe. This initiative outlines specific guidelines for businesses to adhere to the EU's standards on AI transparency and risk management, emphasizing collaboration with Microsoft to leverage technological expertise. By balancing innovation with responsible AI development, Denmark aims to set a precedent for other EU member states, fostering a safe digital environment while promoting economic growth and ethical considerations in technology.[33]

## South Korea

### New Regulations in South Korea Target Deepfake Pornography

South Korea has announced a comprehensive initiative to combat the rising issue of deepfake pornography, implementing tougher penalties and enhanced regulations. This includes making the possession and viewing of deepfake porn illegal, with offenders facing up to three years in prison, while those who produce or distribute such content could receive sentences of up to seven years. The government plans to expand the use of undercover officers for investigations and increase monitoring of social media platforms to prevent the spread of these harmful materials. This initiative is crucial as it addresses the growing concerns over nonconsensual explicit content, which has led to significant distress among victims, particularly women and minors. By taking these steps, South Korea aims to protect individuals from digital sexual crimes and foster a safer online environment, ultimately promoting accountability and respect in digital spaces.[34]

### South Korea's AI Basic Law Advances Towards Enactment Amid Bipartisan Support

The AI Basic Law, previously abolished in the 21st National Assembly, has passed the standing committee's bill review subcommittee nearly six months after the 22nd National Assembly was launched. With bipartisan agreement on key issues, the enactment of the bill within this year appears promising. The new bill introduces risk-based regulations, creating responsibilities for operators of "high-impact AI"—AI systems that significantly impact human life, health, safety, basic rights, national security, and public welfare. It allows companies to verify with the Minister of Science and ICT whether their AI technology falls under this category, imposes fines for non-compliance, and mandates global AI companies to designate domestic agents and watermark AI-generated content. The bill also includes provisions for AI promotion plans, support for AI Safety Research Institute and AI Association, and investment in AI Data Centres.[35]

[33] https://www.cnbc.com/2024/11/13/denmark-lays-out-eu-ai-act-compliance-blueprint-with-microsoft-backing.html

[34] https://abcnews.go.com/International/wireStory/south-korea-fights-deepfake-porn-tougher-punishment-regulation-115556398

[35] https://www.mk.co.kr/en/it/11174781?utm_source=substack&utm_medium=email

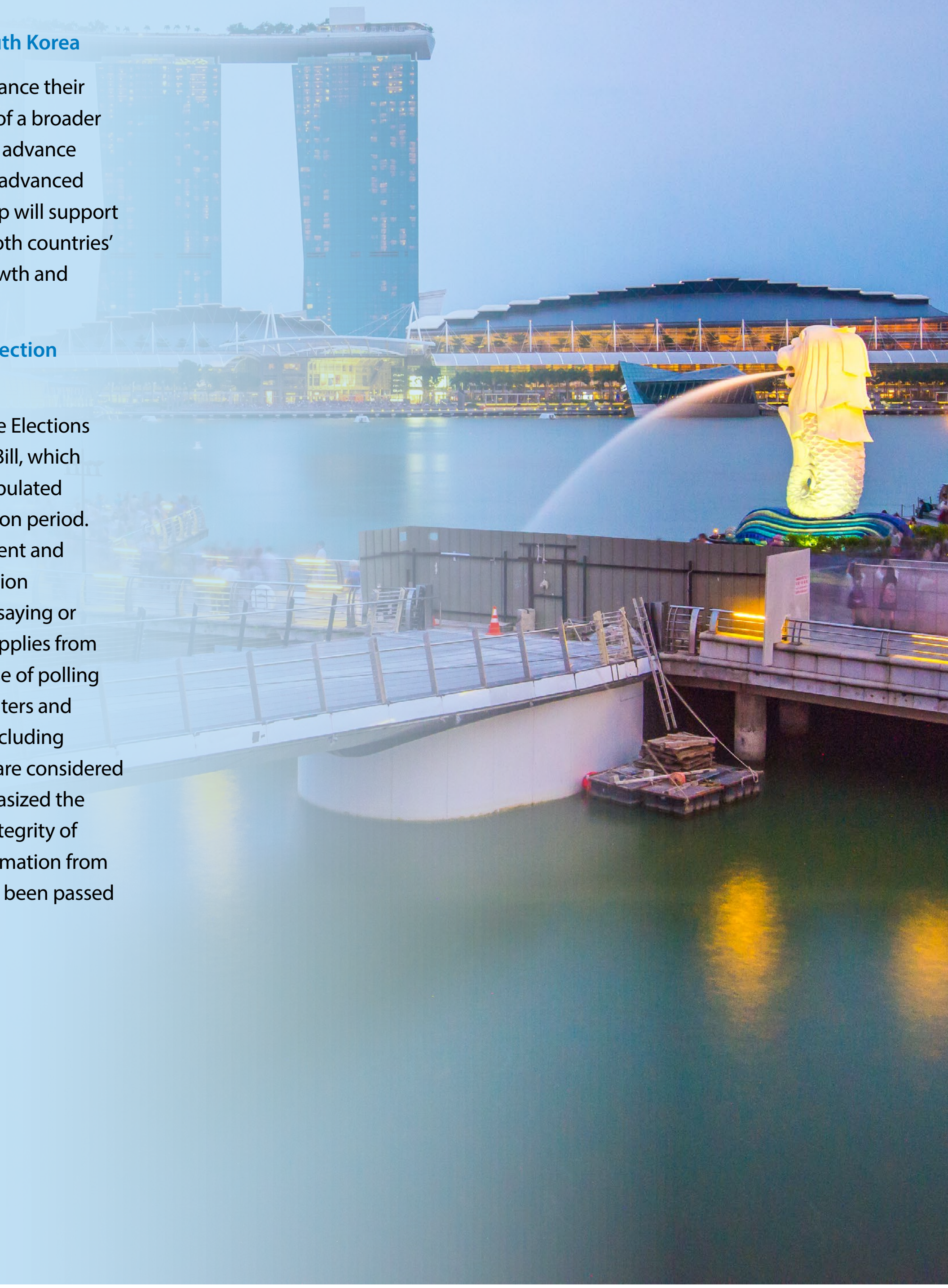
## Singapore

### AI Collaboration Between Singapore and South Korea

Singapore and South Korea have agreed to enhance their cooperation in artificial intelligence (AI) as part of a broader strategic partnership. This collaboration aims to advance AI development and integration, particularly in advanced technology sectors. Additionally, the partnership will support startups and innovation in AI, demonstrating both countries' commitment to leveraging AI for economic growth and technological advancement.[36]

### Singapore Enacts Law to Ban Deepfakes of Election Candidates

On 15th October,2024 Singapore has passed the Elections (Integrity of Online Advertising) (Amendment) Bill, which aims to ban deepfakes and other digitally manipulated content of election candidates during the election period. Introduced by the Ministry of Digital Development and Information (MDDI), the law targets online election advertising that realistically depicts candidates saying or doing things they did not actually say or do. It applies from the issuance of the writ of election until the close of polling and excludes minor modifications like beauty filters and unrealistic entertainment content. Violations, including publishing, sharing, or reposting such content, are considered criminal offenses. Minister Josephine Teo emphasized the importance of this measure in protecting the integrity of elections and preventing AI-generated misinformation from undermining democratic processes. The bill has been passed into a law.[37]



[36] https://www.scmp.com/news/asia/southeast-asia/article/3281493/singapore-south-korean-deepen-ai-start-cooperation-strategic-partners

[37] https://www.channelnewsasia.com/singapore/singapore-ban-deepfakes-general-election-candidates-law-4679781

## Thailand

### Thailand's Initiative to Promote AI Ethics and Establish Regional Training Hub

Thailand is advancing its commitment to AI ethics by launching an initiative to become a regional AI training centre in collaboration with UNESCO. This effort aims to support developing countries in adopting AI ethics rules. The Thai government is seeking cabinet approval for an AI governance framework to address challenges in sectors such as healthcare, agriculture, education, energy, and finance. This framework emphasizes ethical AI use to mitigate risks like bias, privacy issues, and security concerns. [38]

## Iceland

### Iceland Introduces 2024-2026 AI Action Plan

The Artificial Intelligence Action Plan 2024-2026 has been introduced by Iceland, outlining a comprehensive strategy for integrating AI technologies across various sectors while ensuring ethical standards and societal benefits. This plan emphasizes collaboration among stakeholders, including government, industry, and academia, to foster innovation while addressing potential risks associated with AI deployment. Key initiatives include enhancing AI education, promoting research and development, and establishing regulatory frameworks to safeguard public interests. The action plan aims to position Iceland as a leader in responsible AI use, with implementation set to begin in early 2024, marking a significant step towards harnessing AI's potential for positive impact.[39]

[38] https://www.bangkokpost.com/business/general/2909302/thailand-in-drive-to-promote-ai-ethics?utm_source=substack&utm_medium=email

[39] https://island.is/samradsgatt/mal/3862?utm_source=substack&utm_medium=email

## Brazil

### Brazil's Senate passes Comprehensive AI Regulation Bill

Brazil's Senate approved the Bill 2338/23 on 5th December 2024, a comprehensive AI regulation framework, establishing the National System for Regulation and Governance of Artificial Intelligence (SIA), coordinated by the National Data Protection Authority (ANPD). This legislation is crucial for ensuring the safe and ethical development of AI technologies, particularly those classified as "high-risk," which include systems that significantly impact people's rights, health, or safety. Companies developing and operating these AI systems must assess and mitigate associated risks. Additionally, the bill addresses copyright issues, requiring transparency about the use of protected works in AI training, allowing copyright holders to opt-out, and ensuring remuneration for the use of their works. Non-compliance with the regulations could result in penalties and stricter oversight by regulatory agencies, ensuring accountability and adherence to ethical standards.[40]

## Greece

### Greece Unveils Strategic Blueprint for AI Transformation to Drive Economic and Societal Progress

Greece has introduced a comprehensive AI strategy titled "A Blueprint for Greece's AI Transformation," developed by the High-Level Advisory Committee on Artificial Intelligence in collaboration with the Special Secretariat of Foresight. This strategy aims to leverage AI for economic and societal advancement by focusing on human dignity and transparency, ensuring AI respects human rights and operates transparently. It promotes adaptability and international cooperation, outlines six flagship projects to enhance AI capabilities, and aligns with global AI frameworks like those from the OECD to ensure best practices. The strategy positions Greece as a leader in AI innovation while maintaining ethical standards and public trust.[41]

[40] https://www.dataprivacybr.org/en/the-artificial-intelligence-legislation-in-brazil-technical-analysis-of-the-text-to-be-voted-on-in-the-federal-senate-plenary/

[41] https://foresight.gov.gr/en/studies/A-Blueprint-for-Greece-s-AI-Transformation/?utm_source=substack&utm_medium=email

## Ireland

### Ireland's National AI Strategy Refresh: A Commitment to Innovation and Ethics

Singapore and South Korea have agreed to enhance their On November 6, 2024, the Government of Ireland launched a refreshed National AI Strategy, building on the original strategy, "AI – Here for Good," introduced in July 2021. This updated strategy addresses significant advancements in AI technology and regulation, aiming to position Ireland as a leader in ethical AI development. Key strategic actions include:

- Leadership in EU AI Act: Ensuring effective implementation and participation in the EU AI Board.
- Impact Studies: Commissioning research on AI's effects across key sectors.
- SME Awareness Campaign: Promoting AI adoption among small and medium enterprises.
- Regulatory Sandbox: Establishing a framework to foster AI innovation.
- National AI Research Nexus: Creating a unified identity for AI research initiatives.
- Talent Development: Enhancing AI skills through various educational programs.

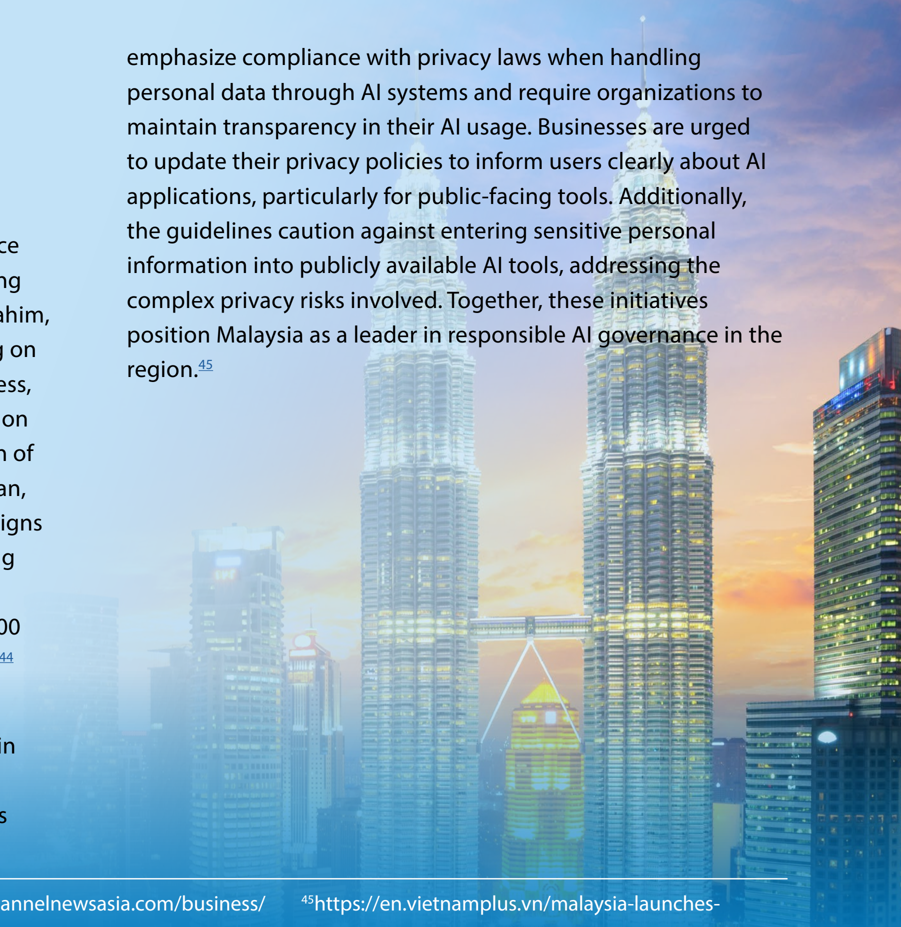
This refresh emphasizes a balanced approach to innovation and regulation, aiming to harness AI's potential for economic and societal benefits while ensuring public trust.[42]

[42] https://www.gov.ie/en/publication/6df28-national-ai-strategy-refresh-2024/?utm_source=substack&utm_medium=email

## Indonesia

### Indonesia's AI Readiness: A Collaborative Effort by UNESCO and KOMINFO

UNESCO, in partnership with Indonesia's Ministry of Communications and Informatics (KOMINFO), has successfully completed an AI Readiness Assessment for Indonesia, marking it as the first Southeast Asian nation to do so using UNESCO's Readiness Assessment Methodology (RAM). The comprehensive assessment, which engaged over 500 participants from government, academia, civil society, and the private sector across five regions, evaluated Indonesia's AI landscape across five key dimensions: legal/regulatory, socio-cultural, economic, scientific/educational, and technical/infrastructural. The resulting AI Readiness Assessment Report aims to shape Indonesia's AI policy and regulatory frameworks, promoting an inclusive and responsible AI ecosystem. This initiative underscores Indonesia's dedication to ethical AI practices and paves the way for future technological advancements that are both inclusive and sustainable.[43]



## Malaysia

### Malaysia's Commitment to Responsible AI Practices

Malaysia is making significant advancements in AI governance with the introduction of a National AI Policy aimed at fostering ethical AI practices. Announced by Prime Minister Anwar Ibrahim, this policy will establish a national cloud framework focusing on enhancing public service efficiency, economic competitiveness, and user trust. A key component of this initiative is the creation of a national AI office, which will oversee the implementation of a regulatory framework and a five-year technology action plan, set to be completed within the next 12 months. This move aligns with substantial investments from global tech firms, including Google's recent commitment of $2 billion to develop a data centre and cloud region in Malaysia, expected to create 26,500 jobs and contribute over $3 billion to the economy by 2030. [44]

In conjunction with the national policy, In September 2024, the Ministry of Science, Technology and Innovation (MOSTI) in Malaysia has also launched guidelines on AI governance and ethics to ensure responsible AI deployment. These guidelines emphasize compliance with privacy laws when handling personal data through AI systems and require organizations to maintain transparency in their AI usage. Businesses are urged to update their privacy policies to inform users clearly about AI applications, particularly for public-facing tools. Additionally, the guidelines caution against entering sensitive personal information into publicly available AI tools, addressing the complex privacy risks involved. Together, these initiatives position Malaysia as a leader in responsible AI governance in the region.[45]

[43] https://www.unesco.org/en/articles/unesco-and-kominfo-completed-ai-readiness-assessment-indonesia-ready-ai

[44] https://www.channelnewsasia.com/business/malaysia-plans-national-cloud-policy -ai-regulations-4650981

[45] https://en.vietnamplus.vn/malaysia-launches-national-guidlines-on-ai-governance-ethnics-post296945.vnp

## UAE

### UAE’s New AI Foreign Policy: Safeguarding Technology

The UAE has launched a new AI foreign policy to prevent technology misuse, focusing on six key principles: advancement, cooperation, community, ethics, sustainability, and security. Developed by the Office of the Assistant Foreign Minister for Advanced Science and Technology and the Office of the Minister of State for Artificial Intelligence, Digital Economy, and Remote Work Applications, the policy aims to align AI development with ethical, social, and environmental priorities. It includes participating in international AI forums, advocating for transparency and accountability in AI tools, and supporting global alliances to govern and secure AI systems.[46]

### UAE Cabinet Approves Landmark AI Policy to Strengthen Global Leadership

The UAE Cabinet, chaired by His Highness Sheikh Mohammed Bin Rashid Al Maktoum, has approved the country’s official stance on artificial intelligence (AI) policy, marking a significant step in reinforcing the UAE’s global leadership in technology. This policy, developed by the Office of the Assistant Foreign Minister for Advanced Science and Technology and the Office of the Minister of State for Artificial Intelligence, aims to address the complex challenges posed by AI on an international scale. It emphasizes ethical, social, and environmental priorities, ensuring that AI advancements contribute to societal well-being and economic diversification. This initiative is crucial as it aligns the UAE’s foreign policy with global AI standards, fostering international cooperation and trust. The policy was approved on October 28, 2024, and is expected to have a profound impact on the global governance of AI.[47]

## Peru

### Peru Releases Draft Regulation for AI Law to Promote Responsible Development

Peru’s government has unveiled a draft regulation for the Law on Artificial Intelligence (Law No. 31814), aimed at fostering economic and social development through the responsible use of AI. Prepared by the Secretariat of Government and Digital Transformation under the Presidency of the Council of Ministers, the draft is open for public consultation from November 25 to December 6, 2024. It focuses on mitigating risks and potential negative impacts of AI misuse by establishing a comprehensive framework for its application. The regulation emphasizes the importance of public input to ensure it effectively addresses various challenges and supports Peru’s growth through ethical and responsible AI use.[48]

[46] https://www.thenationalnews.com/future/technology/2024/10/11/uaes-new-ai-foreign-policy-aims-to-prevent-misuse-of-technology/?utm_source=substack&utm_medium=email

[47] https://www.mofa.gov.ae/en/mediahub/news/2024/10/28/28-10-2024-uae-technology

[8] https://www.gob.pe/institucion/pcm/informes-publicaciones/6197119-nuevo-proyecto-de-reglamento-de-la-ley-de-inteligencia-artificial?utm_source=chatgpt.com&utm_medium=email

## Vietnam

### Vietnam to Strengthen Digital Tech Sector with New Regulatory Framework

Vietnam is poised to advance its digital technology sector with a new law designed to foster innovation and growth. The government aims to introduce regulations that will bolster digital infrastructure, enhance cybersecurity, and drive digital transformation across various industries. This initiative is part of a broader strategy to establish Vietnam as a leading digital economy in the region. The law will also address data protection and the ethical use of technology, ensuring a secure and sustainable digital environment for businesses and citizens. This moves highlights Vietnam's commitment to leveraging digital technologies for economic development and improving the quality of life for its people.[49]

## Standards

### Guidelines for Secure Use of AI Coding Assistants Released by ANSSI and BSI

The French Cybersecurity Agency (ANSSI) and the German Federal Office for Information Security (BSI) have jointly issued guidelines for the secure use of AI coding assistants, emphasizing both their benefits and potential risks. These tools can enhance productivity by generating source code, providing explanations, and automating repetitive tasks, but they also pose risks such as leaking sensitive information and introducing security vulnerabilities. The report recommends that AI coding assistants should complement, not replace, experienced developers and stresses the importance of conducting thorough risk assessments before their use, including evaluating the trustworthiness of providers and third parties. The guidelines from the ANSSI and the BSI are intended for organizations and developers using AI coding assistants.[50]

### Proactive AI Safety: Japan's Red Teaming and Evaluation Standards

Japan is advancing its commitment to responsible AI through two key standards established by the AI Safety Institute (AISI). The Guide to Red Teaming Methodology emphasizes the importance of proactive testing and evaluation of AI systems to identify vulnerabilities and ensure safety throughout their lifecycle. This approach encourages organizations to simulate potential attacks and assess the resilience of their AI models, fostering a culture of continuous improvement and risk management.[51]

The Guide to Evaluation Perspectives on AI Safety outlines essential criteria for assessing AI systems, focusing on maintaining safety, fairness, and transparency. It highlights the need for privacy protection, security against external threats, and the verifiability of AI operations, all grounded in a human-centric approach. Together, these standards aim to enhance the safety and reliability of AI technologies in Japan, promoting ethical practices in their deployment.[52]

### Australia Releases AI Impact Navigator and OAIC Guide on Privacy Risks of Facial Recognition Technology for Ethical and Responsible AI Use

The AI Impact Navigator, published by the Australian Department of Industry, Science, and Resources, targets CEOs, organization executives, board directors, business owners, practice and people leaders, and anyone else leading AI implementation within their companies. These entities must follow guidelines to ensure AI is used ethically and responsibly, focusing on transparency, accountability, and minimizing risks such as bias and privacy breaches. The regulation impacts these organizations by requiring them to conduct thorough assessments of AI systems, integrate human oversight, and update policies to reflect AI usage. By adhering to these principles, businesses can build trust with

[49] https://vietnamnet.vn/en/vietnam-to-boost-digital-tech-sector-with-new-law-2329961.html?utm_source=substack&utm_medium=email

[50] https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/ANSSI_BSI_AI_Coding_Assistants.pdf?__blob=publicationFile&v=7

[51] https://aisi.go.jp/assets/pdf/ai_safety_RT_v1.00_en.pdf

[52] https://aisi.go.jp/assets/pdf/ai_safety_eval_v1.01_en.pdf

consumers, enhance their reputation, and ensure compliance with evolving regulatory standards. The AI Impact Navigator can help support organisations to implement the Australian Government's Voluntary AI Safety Standard. [53]

The Office of the Australian Information Commissioner (OAIC) has published a detailed guide to assist private sector organizations in assessing the privacy risks associated with facial recognition technology (FRT). The guide advocates for a privacy-by-design approach, urging organizations to conduct privacy impact assessments (PIAs) to identify and mitigate potential privacy impacts from the outset. It emphasizes the necessity and proportionality of using FRT, ensuring it is only employed when necessary and when no less privacy-intrusive means is available. The guide also highlights the importance of obtaining meaningful consent from individuals, maintaining transparency, and addressing issues related to accuracy, bias, and discrimination. Additionally, it underscores the need for robust governance and ongoing assurance to minimize privacy risks. Given that biometric information is classified as sensitive under the Privacy Act, the guide stresses the importance of higher protection levels to balance the benefits of FRT with the need to safeguard individuals' privacy. This guidance sets out general considerations for private sector organisations that are considering using facial recognition technology (FRT) to undertake facial identification in a commercial or retail setting.[54]

### Hong Kong Introduces Responsible AI Framework and SFC Guidelines on Generative AI Use in Finance

Hong Kong has unveiled a new framework for the responsible use of artificial intelligence (AI) in the financial sector. This policy emphasizes a dual-track approach to promote AI adoption while addressing challenges like cybersecurity, data privacy, and intellectual property protection. Financial institutions are encouraged to develop AI governance strategies and adopt a risk-based approach to AI system management. The framework also includes public education initiatives to raise awareness about AI's opportunities and risks in financial services.[55]

The Securities and Futures Commission (SFC) of Hong Kong issued a circular on November 12, 2024, outlining expectations for licensed corporations regarding the use of generative AI in finance. The circular emphasizes the increasing adoption of generative AI tools, such as chatbots and data analysis engines, while highlighting the need for robust governance to mitigate associated risks. Key concerns include the potential for inaccurate outputs, biases in AI models, and operational challenges related to data privacy and cybersecurity. The SFC established four core principles for compliance: senior management oversight, AI model risk management, cybersecurity measures, and third-party risk management. The circular particularly stresses the importance of human oversight in high-risk applications, such as investment advice, to ensure responsible AI usage in the financial sector.[56]

### India Launches Trustworthy AI Framework for Defence and Prepares Voluntary Ethics Code for Responsible AI Development

The launch of the Trustworthy AI Framework for critical defense operations marks a significant step towards ensuring the reliability and ethical use of AI in military contexts. The ETAI Framework focuses on five broad principles: Reliability & Robustness, Safety & Security, Transparency, Fairness and Privacy. By implementing this framework, defense operations can benefit from enhanced decision-making capabilities, improved operational efficiency, and reduced risks associated with AI deployment. This initiative aims to bolster national security by integrating advanced AI technologies in a manner that upholds ethical standards and fosters trust among stakeholders.[57]

The Ministry of Electronics and Information Technology (MeitY) in India is developing a voluntary ethics code for AI firms to address the increasing concerns surrounding artificial intelligence. This initiative aims to promote responsible AI development and deployment by emphasizing transparency, accountability, and fairness. The code will encourage AI companies to voluntarily adopt best practices and ethical guidelines, fostering trust and safety in AI technologies. This effort is part of India's broader strategy to balance technological innovation with ethical considerations, ensuring that AI advancements benefit society while minimizing potential risks. The voluntary code of conduct is expected to be released early next year.[58]

### Generative AI Governance: Best Practices for Organizations from ETDA, Thailand

The ETDA (Electronic Transactions Development Agency) has introduced new guidelines for the responsible use of Generative AI in organizations, developed in partnership with the Ministry of Digital Economy and Society. These guidelines aim to help organizations effectively implement Generative AI while addressing risks related to data privacy and security. They cover the benefits and limitations of the technology, risk management strategies, and ethical governance principles. Available for free download, the guidelines encourage widespread adoption to enhance digital transformation in Thailand, ensuring that AI technologies are utilized safely and ethically across various sectors.[59]

[53] https://www.industry.gov.au/publications/ai-impact-navigator?utm_source=substack&utm_medium=email

[54] https://www.oaic.gov.au/privacy/privacy-guidance-for-organisations-and-government-agencies/organisations/facial-recognition-technology-a-guide-to-assessing-the-privacy-risks?utm_source=substack&utm_medium=email

[55] https://www.news.gov.hk/eng/2024/10/20241028/20241028_110857_078.html

[56] https://babl.ai/hong-kongs-securities-and-futures-commission-issues-circular-on-generative-ai-use-in-finance/?utm_source=BABL+AI+Inc.&utm_campaign=d2b8d5e936-EMAIL_CAMPAIGN_2024_03_15_06_03_COPY_01&utm_medium=email&utm_term=0_-3c9a50eb57-578132409

[57] https://indiaai.gov.in/news/trustworthy-ai-framework-launched-for-critical-defence-operations

[58] https://inc42.com/buzz/govt-working-on-voluntary-codes-of-conduct-for-ai-companies/

[59] https://www.etda.or.th/th/pr-news/AI_Gov_Anual.aspx

### UK Government Unveils Comprehensive AI Governance and Safety Initiatives, Including AI Management Essentials Tool, Safety Assurance Platform, Inspect Evals, and AI Security Research Lab

The UK government has launched the AI Management Essentials (AIME) tool, a self-assessment resource aimed at helping organizations establish effective management practices for AI systems. Designed primarily for small to medium-sized enterprises (SMEs) and startups, AIME provides guidance on implementing responsible AI governance. The tool focuses on evaluating organizational processes rather than specific AI products, helping users navigate existing standards and frameworks. The government is currently seeking feedback on the tool's design and functionality through a consultation that will inform its further development.[60]

On November 6, 2024, the UK Government announced the launch of a new AI Safety Assurance Platform designed to help businesses develop and deploy trustworthy AI systems. This initiative is part of a broader strategy to ensure public trust in AI technologies, which are expected to unlock over £6.5 billion in economic benefits over the next decade. The platform will serve as a one-stop resource, providing guidance on conducting impact assessments, reviewing data for bias, and identifying potential risks associated with AI. By equipping businesses with the necessary tools and clarity, the government aims to foster a safe and responsible AI landscape, reinforcing the UK's position as a hub for AI assurance expertise.[61]

Inspect Evals is a new initiative from the UK AI Safety Institute aimed at enhancing the evaluation of large language models (LLMs) by providing a repository of community-contributed benchmark evaluations. This platform offers dozens of high-quality, open-source evaluations across various domains, including coding, mathematics, cybersecurity, and reasoning, addressing the challenges organizations face in implementing consistent evaluation standards. Developed in collaboration with Arcadia Impact and the Vector Institute, Inspect Evals is built on the open source Inspect AI framework, which has garnered contributions from over 50 collaborators. Users can easily run and experiment with evaluations through a simple Python package, streamlining the evaluation process with several agent benchmarks executable via a single command. By fostering collaboration within the evals community and reducing duplicated efforts, Inspect Evals aims to improve the understanding of AI capabilities and safety characteristics, ultimately promoting more responsible AI development.[62]

The United Kingdom has announced the establishment of the Laboratory for AI Security Research (LASR) at the NATO Cyber Defence Conference in London. This initiative, supported by an initial £8.22 million in government funding, aims to address emerging threats in AI and national security. LASR will bring together experts from academia, industry, and government to develop innovative solutions for AI-related security challenges. The lab will collaborate with international allies, including NATO member states and Five Eyes partners, to enhance global cyber resilience.[63]

The UK government has released a Draft Statement of Strategic Priorities for Online Safety, outlining its key focus areas for creating a safer online environment. This document, part of the broader Online Safety Act, emphasizes five main themes: safety by design, transparency and accountability, agile regulation, inclusivity and resilience, and technology and innovation. The draft statement aims to protect users, particularly children and vulnerable individuals, from online harms by holding social media companies and search services accountable for user safety. This initiative will impact large technology companies, online platforms, and service providers, requiring them to adopt safer practices and collaborate with the regulator, Ofcom, to ensure compliance.[64]

[60] https://www.gov.uk/government/consultations/ai-management-essentials-tool

[61] https://www.gov.uk/government/news/ensuring-trust-in-ai-to-unlock-65-billion-over-next-decade?utm_source=substack&utm_medium=email

[62] https://www.aisi.gov.uk/work/inspect-evals

[63] https://babl.ai/uk-unveils-ai-security-laboratory-at-nato-cyber-defense-conference/?utm_source=BABL+AI+Inc.&utm_campaign=395f69b07b-EMAIL_CAMPAIGN_2024_03_15_06_03_COPY_01&utm_medium=email&utm_term=0_-3c9a50eb57-578132409

[64] https://www.gov.uk/government/publications/draft-statement-of-strategic-priorities-for-online-safety/draft-statement-of-strategic-priorities-for-online-safety

### AI Safety Institutes across the globe

As nations around the globe accelerate their AI initiatives, various countries have established dedicated AI safety institutes to address the growing concerns of ethical and responsible AI deployment. These institutes focus on critical issues such as fairness, transparency, and accountability, while aligning their efforts with national priorities to ensure AI technologies are developed and applied safely for the benefit of society.[70]

| Parameter | USA | UK | Japan | Canada | Singapore | EU |
|---|---|---|---|---|---|---|
| Primary Focus | Research, Standards, Cooperation | Research, Standards, Cooperation | Research, Standards, Cooperation | Research, Standards, Cooperation | Research, Standards, Cooperation | Regulation, Research, Standards |
| Regulatory Powers | No | No | No | No | No | Yes |
| Core Functions | Safety evaluations, Foundational AI safety research | Safety evaluations, Foundational AI safety research | Safety evaluations, Foundational AI safety research | Safety evaluations, Foundational AI safety research | Safety evaluations, Foundational AI safety research | Safety evaluations, Regulatory oversight |
| International Coordination | Yes | Yes | Yes | Yes | Yes | Yes |
| Standards Development | Light-touch | Light-touch | Light-touch | Light-touch | Light-touch | Setting standards |
| Funding Amount | $10 million | $5 million | $3 million | $4 million | $3.5 million | $12 million |
| Governance | U.S. Department of Commerce & U.S. Department of State | UK Government | Japanese Government | Canadian Government | Singapore Government | European Commission |

### UNESCO's Global AI Ethics and Governance Observatory: An Initiative Update on Standards and Governance

The Global AI Ethics and Governance Observatory has recently launched a suite of innovative tools and resources aimed at promoting equitable and inclusive AI standards worldwide. A key feature of this initiative is the Readiness Assessment Methodology, designed to help UNESCO member states evaluate and enhance their AI governance capabilities. This structured tool provides insights into regulatory preparedness, enabling countries to adopt and regulate AI technologies responsibly while considering local contexts. Additionally, the Observatory is developing a Civil Society Organisation (CSO) Repository, an open-access directory that will facilitate collaboration among global organizations focused on AI ethics and governance. These efforts underscore the Observatory's commitment to fostering transnational cooperation and establishing a robust framework for ethical AI practices that prioritize human rights and responsible innovation.[71]

### U.S. Introduces Comprehensive AI Safety and Ethics Initiatives Across Critical Infrastructure, Education, and National Security

The Department of Homeland Security (DHS) has unveiled a new "Roles and Responsibilities Framework for Artificial Intelligence in Critical Infrastructure," aimed at ensuring the safe and secure deployment of AI across essential services. This framework, developed in collaboration with industry leaders and civil society, outlines voluntary guidelines for various stakeholders, including cloud providers, AI developers,

[70] https://cfg.eu/the-ai-safety-institute-network-who-what-and-how/

[71] https://babl.ai/global-ai-ethics-observatory-expands-to-drive-inclusive-governance-frameworks-worldwide/?utm_source=BABL+AI+Inc.&utm_campaign=6481494ff2-EMAIL_CAMPAIGN_2024_03_15_06_03_COPY_01&utm_medium=email&utm_term=0_-3c9a50eb57-578128418

and infrastructure operators. It addresses key issues such as data governance, AI system design, and monitoring practices. By promoting responsible AI use, the framework seeks to enhance the resilience of critical infrastructure while minimizing risks associated with AI technologies, ultimately safeguarding public services like energy, water, and telecommunications.[72]

The U.S. Department of Education has released an AI Toolkit titled "Empowering Education Leaders: A Toolkit for Safe, Ethical, and Equitable AI Integration," designed to help school leaders leverage AI for teaching and learning while managing associated risks. This initiative, in response to President Biden's October 2023 Executive Order on AI, provides actionable guidance for developing strategies that ensure AI use is safe, secure, and trustworthy. The toolkit is divided into three sections: mitigating risks to safeguard student privacy and rights, building a strategy for AI integration in the instructional core, and enhancing equity and accessibility to ensure fair use of AI tools. Developed through consultations with over 200 educators and AI experts, the toolkit underscores the Department's commitment to informed and equitable AI integration in education.[73]

The U.S. AI Safety Institute has established the Testing Risks of AI for National Security (TRAINS) Taskforce to address the national security implications of rapidly evolving AI technologies. This taskforce, comprising experts from various federal agencies, aims to identify, measure, and manage risks associated with AI in critical areas such as cybersecurity and public safety. The initiative reflects a commitment to ensuring safe and trustworthy AI innovation while maintaining U.S. leadership in the field. This announcement coincides with the upcoming convening of the International Network of AI Safety Institutes, highlighting the importance of collaboration in AI safety efforts.[74]

### Council of Europe Introduces HUDERIA Methodology for AI Risk and Impact Assessment

The Council of Europe has unveiled the HUDERIA methodology, a detailed framework designed to assess the risks and impacts of Artificial Intelligence (AI) systems on human rights, democracy, and the rule of law. This methodology emphasizes context-based risk analysis, stakeholder engagement, and comprehensive risk and impact assessments. It aims to ensure that AI technologies are developed and implemented in ways that uphold fundamental rights and democratic values. The guidelines include steps for identifying potential risks, involving relevant stakeholders, and conducting thorough assessments to mitigate any adverse effects.[75]

### South Korea Launches AI Safety Institute and Introduces Robust AI Privacy Protection Framework

South Korea has inaugurated its AI Safety Institute in Pangyo to address the risks associated with rapidly advancing artificial intelligence technology. The institute will spearhead research on AI-related risks, including misuse and loss of control, and will act as a collaborative hub for industry, academia, and research institutions. This initiative follows the AI Seoul Summit, where leaders from South Korea, Britain, and other nations adopted a joint declaration to promote safe, innovative, and inclusive AI. The institute aims to support local AI companies by minimizing risk factors that could impede their global competitiveness. Kim Myung-joo, an information security professor at Seoul Women's University, has been appointed as the inaugural chief. A consortium of 24 entities, including major tech firms like Naver, KT, and Kakao, as well as leading universities, has signed a memorandum of understanding to collaborate on research, policymaking, and AI safety evaluation.[76]

South Korea has introduced a comprehensive framework for AI privacy protection, led by the Personal Information Protection Commission (PIPC). This initiative aims to balance AI innovation with stringent privacy safeguards. The framework includes guidelines for safe personal data processing in AI environments, emphasizing privacy-by-design principles, transparency, and compliance with the Personal Information Protection Act (PIPA). Additionally, an AI Privacy Team has been established to assist businesses in interpreting and adhering to these regulations, ensuring responsible AI development and usage.[77]

### Malaysia Launches National AI Office to Drive Policy and Regulation

Malaysia has launched a National AI Office to shape policies and address regulatory issues, aiming to establish itself as a regional hub for AI development. The office will serve as a centralized agency for strategic planning, research and development, and regulatory oversight. It will focus on seven key deliverables in its first year, including developing a code of ethics and an AI regulatory framework. The government has also announced strategic partnerships with major tech companies like Amazon, Google, and Microsoft.[78]

[72] https://www.dhs.gov/news/2024/11/14/groundbreaking-framework-safe-and-secure-deployment-ai-critical-infrastructure

[73] https://babl.ai/u-s-department-of-education-releases-ai-toolkit-to-guide-schools-in-ethical-ai-integration/?utm_source=BABL+AI+Inc.&utm_campaign=6481494ff2-EMAIL_CAMPAIGN_2024_03_15_06_03_COPY_01&utm_medium=email&utm_term=0_-3c9a50eb57-578128418

[74] https://www.nist.gov/news-events/news/2024/11/us-ai-safety-institute-establishes-new-us-government-taskforce-collaborate

[75] https://rm.coe.int/cai-2024-16rev2-methodology-for-the-risk-and-impact-assessment-of-arti/1680b2a09f?utm_source=substack&utm_medium=email

[76] https://www.koreaherald.com/view.php?ud=20241127050025&utm_source=substack&utm_medium=email

[77] https://babl.ai/south-korea-unveils-comprehensive-framework-for-ai-privacy-protection/?utm_source=BABL+AI+Inc.&utm_campaign=395f69b07b-EMAIL_CAMPAIGN_2024_03_15_06_03_COPY_01&utm_medium=email&utm_term=0_-3c9a50eb57-578132409

[78] https://www.reuters.com/technology/artificial-intelligence/malaysia-launches-national-ai-office-policy-regulation-2024-12-12/

### MAS Publishes Review on AI Risk Management for Financial Institutions

The Monetary Authority of Singapore (MAS) has published a review of observations on the risk management of AI models for financial institutions. The review identifies principal practices for managing AI-related risks within the financial sector, including those associated with generative AI. It highlights the necessity for governance structures, effective risk identification, and comprehensive processes across the entire AI lifecycle. Additionally, the review recommends that banks maintain inventories of AI models, conduct systematic assessments of risk materiality, and implement standards for the development, validation, and monitoring of these models.[79]

### CSIRO and Adelaide University Launch Landmark AI Research Centre

Australia has taken a significant step towards leading the world in responsible artificial intelligence (AI) research with the launch of the Responsible AI Research (RAIR) Centre in Adelaide. This landmark collaboration brings together experts from the Commonwealth Scientific and Industrial Research Organisation (CSIRO) and The University of Adelaide, in partnership with the South Australian Government. The RAIR Centre will focus on four key research areas: tackling misinformation, ensuring safe AI interactions in the real world, developing diverse AI systems, and creating AI that can explain its actions. This initiative underscores Australia's commitment to advancing AI technology in a safe and ethical manner.[80]

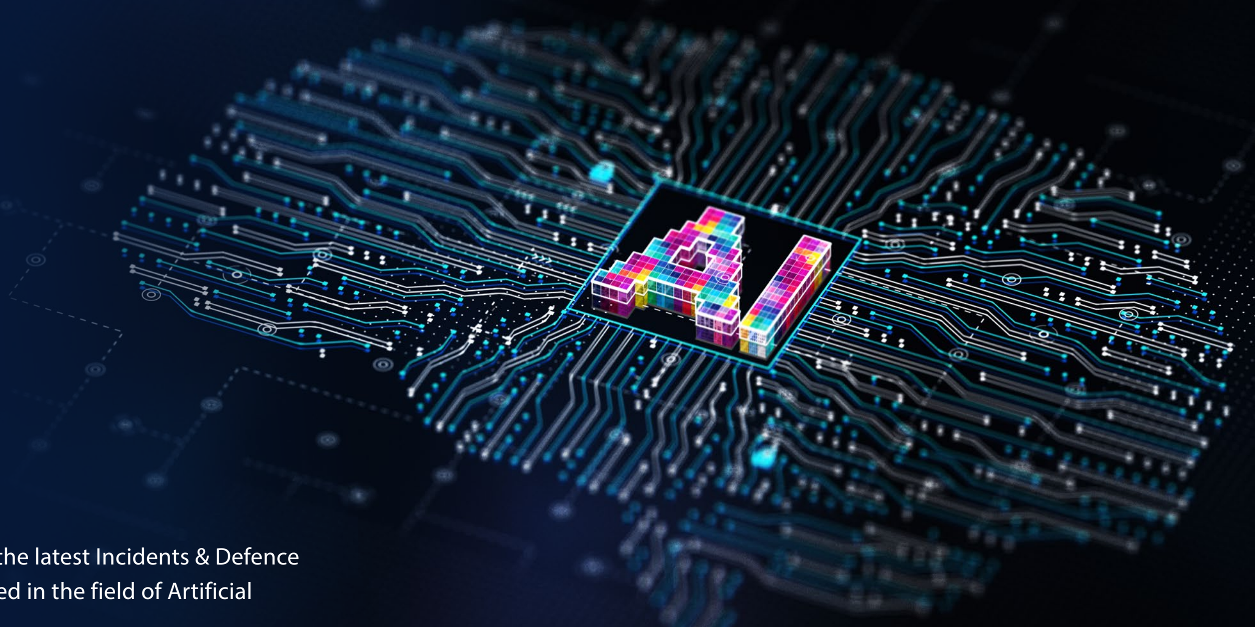
### CAC's "Clear and Bright" Initiative Targets Algorithmic Transparency and Fairness

The Cyberspace Administration of China (CAC) has launched the "Clear and Bright: Typical Problems Governance of Internet Platform Algorithms" initiative, which aims to enhance algorithmic transparency and fairness across internet platforms, particularly those employing recommendation algorithms, live-streaming services, and social media. Running until February 14, 2025, this special action addresses key issues such as the manipulation of ranking lists, the protection of workers' rights in new employment forms, and the prevention of discriminatory pricing practices based on user characteristics. By focusing on these areas, the CAC seeks to create a more equitable digital environment and ensure fair treatment for all users.[81]

[79] https://www.mas.gov.sg/-/media/mas-media-library/publications/monographs-or-information-paper/imd/2024/information-paper-on-ai-risk-management-final.pdf

[80] https://www.csiro.au/en/news/All/News/2024/December/Landmark-research-centre-positions-Australia-for-a-safe-and-responsible-AI-future?utm_source=substack&utm_medium=email

[81] https://www.cac.gov.cn/2024-11/24/c_1734143936205514.htm?utm_source=substack&utm_medium=email



## AI Principles

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

## Incidents

### Meta Faces New Copyright Lawsuit Over AI Training Practices

Meta is facing a new class-action copyright lawsuit filed by novelist Christopher Farnsworth, who claims the company used his and other authors' pirated books to train its Llama AI model. This lawsuit, which follows similar actions by other authors like Sarah Silverman, brings to light the ongoing controversy over the use of copyrighted material in AI training. The plaintiffs argue that Meta's actions amount to copyright infringement, as the works were utilized without permission or compensation. This case underscores broader concerns about the ethical and legal implications of using copyrighted content to develop AI technologies. Meta is expected to defend itself by invoking the "fair use" doctrine, which allows limited use of copyrighted material under certain conditions. The outcome of this lawsuit could have significant ramifications for the future of AI development and intellectual property rights.[82]

### Deepfake Impersonation of Ukrainian Official Targets U.S. Senator

Senator Benjamin L. Cardin, chairperson of the Foreign Relations Committee, recently fell victim to a deepfake fraud. An individual pretending to be Dmytro Kuleba, the former foreign minister of Ukraine, managed to engage him in a Zoom call. This incident highlights the increasing danger of deepfake technology being used for political manipulation and misinformation, raising serious concerns about the security of political communications.[83]

### Character.ai in the Spotlight: Ethical Breaches and Mental Health Fallout

In a troubling incident, an AI chatbot mimicking a deceased teenager, Jennifer Ann Crecente, was discovered by her father, Drew Crecente, leading to outrage and significant ethical concerns. The chatbot, hosted on Character.ai, used Jennifer's name and image without consent, engaging in numerous interactions before being detected. This situation has underscored the potential for AI misuse, particularly in exploiting the identities of deceased individuals. Drew Crecente has called for stricter regulations to prevent such violations, emphasizing the urgent need for ethical guidelines in AI development to protect individuals' identities and honor their memory.[84]

Also, a lawsuit has been filed against Character.AI following the tragic suicide of 14-year-old Sewell Setzer III from Florida, who developed an emotional attachment to a chatbot named "Dany" on the platform. The lawsuit alleges that the chatbot exacerbated Setzer's depression and suicidal thoughts, leading to his death, and claims that, Character. Ai's design was defective and lacked adequate warnings about potential dangers. In response, Character.AI has announced new safety measures, including improved detection and intervention for harmful chats and notifications for prolonged usage. This case underscores the urgent need for better safety protocols in AI companionship apps, as their mental health impacts remain unstudied.[85]

[82] https://www.pacermonitor.com/public/case/55297956/Farnsworth_v_Meta_Platforms,_Inc?utm_source=substack&utm_medium=email

[83] https://www.nytimes.com/2024/09/26/us/senator-cardin-deepfake.html

[84] https://www.tbsnews.net/tech/family-outraged-ai-chatbot-mimics-murdered-daughter-961506

[85] https://dataconomy.com/2024/10/24/character-ai-in-legal-trouble-after-14-year-olds-devastating-loss/

### AI Voice-Cloning Scam Targets Parents with Fake Bail Request

In a startling incident, an AI-powered voice-cloning fraud targeted a man's parents, imitating their son's voice to demand Rs 25 lakh for bail money. The fraud involved a call claiming the son was in a severe car accident and urgently needed financial assistance. The convincing nature of the cloned voice almost led to the fraud's success. However, the parents grew suspicious when the caller insisted on a cash payment, thwarting the fraud. This case highlights the increasing sophistication of AI-driven frauds and underscores the urgent need for awareness and regulatory measures to protect individuals from such deceptive practices.[86]

### French Government Sued Over Biased Algorithms Targeting Vulnerable Citizens

A coalition of fifteen human rights organizations has initiated a legal challenge against the French government, accusing it of using a biased algorithm to detect welfare fraud. This algorithm, implemented by the French Social Security Agency's National Family Allowance Fund (CNAF), allegedly discriminates against marginalized groups, including single mothers and disabled individuals, by assigning risk scores based on personal data. The coalition contends that this practice leads to invasive investigations and privacy violations, disproportionately affecting vulnerable recipients and violating French anti-discrimination laws, thereby causing significant harm.[87]

### Hackers Exploit AI Repository: Thousands of Malicious Models Uploaded to Hugging Face

Hackers have recently targeted Hugging Face, a prominent online repository for AI models, by uploading thousands of malicious models designed to appear legitimate but containing harmful code. This large-scale attack, one of the most significant known incidents in the AI space, poses a serious threat as these models can be integrated into various applications, spreading malware, and causing extensive damage. Hugging Face is actively working to identify and remove these malicious models while enhancing their security measures to prevent future breaches. This incident underscores the critical need for robust security practices in AI development and highlights the vulnerabilities inherent in open-source platforms.[88]

### Concerns Over OpenAI's Whisper Tool in Medical Settings Due to Inaccurate Transcriptions

OpenAI's Whisper transcription tool has been found to generate fabricated text, a phenomenon known as "hallucinations," which poses significant risks in medical settings. Despite claims of near-human accuracy, researchers have discovered that Whisper often invents content that was never spoken. A University of Michigan study revealed that 80% of public meeting transcripts examined contained false text. Additionally, research from Cornell University and the University of Virginia found that 1% of samples included entirely fabricated phrases or sentences, some with explicit harms. Despite these issues, the tool is used by over 30,000 medical workers, raising concerns about its reliability in high-risk domains where precise documentation is crucial.[89]

### Digital Arrest Frauds: Government Cautions Citizens: Stay Vigilant

In a concerning revelation, Indians lost approximately ₹120.3 crore to "digital arrest" frauds during the first quarter of 2024, as reported by the Indian Cybercrime Coordination Centre (I4C). Highlighted by Prime Minister Narendra Modi in his Mann Ki Baat address on October 27, 2024, this alarming trend reflects a significant rise in cybercrime, with 7.4 lakh complaints filed between January 1 and April 30, 2024. The fraudsters, often operating from Southeast Asian countries like Myanmar, Laos, and Cambodia, employ tactics that involve impersonating law enforcement to extort money from victims under the guise of resolving fictitious legal issues. This incident underscores the urgent need for enhanced awareness and protective measures against such scams in the digital age.[90]

### AI-Generated Errors in Alaska's Education Policy Highlight Broader Concerns Over AI Reliability

An incident involving Alaska's Department of Education and Early Development (DEED) has raised significant concerns about the reliability of AI-generated data in policymaking. The department used generative AI to draft a policy banning cell phones in schools, which included false citations to non-existent academic studies. These inaccuracies, not disclosed as AI-generated, were presented to the Alaska State Board of Education and Early Development, potentially influencing their discussions. Despite efforts to correct the errors, the final document still contained fabricated citations. The resolution, which directed DEED to craft a model policy for cell phone restrictions, cited supposed scholarly articles that could not be found at the listed web addresses, and whose titles did not appear in broader online searches. Four of the six citations were false, referencing real journals but non-existent articles. This incident underscores the critical need for rigorous human verification to prevent AI-generated errors from undermining credibility in various sectors. It also highlights the broader issue of AI "hallucinations," where AI systems generate plausible but false information due to insufficient data or incorrect assumptions.[91]

[86] https://nypost.com/2024/10/08/tech/fla-pol-targeted-in-elaborate-car-crash-ai-scam-which-almost-fooled-his-dad-into-forking-over-35k/

[87] https://www.amnesty.org/en/latest/news/2024/10/france-discriminatory-algorithm-used-by-the-social-security-agency-must-be-stopped/

[88] https://www.forbesafrica.com/technology/2024/10/27/hackers-have-uploaded-thousands-of-malicious-files-to-ais-biggest-online-repository/

[89] https://www.wired.com/story/hospitals-ai-transcription-tools-hallucination/#:~:text=Upon%20its%20release%20in%202022,of%20public%20meeting%20transcripts%20examined. ais-biggest-online-repository/

[90] https://www.livemint.com/news/indians-lost-120-cr-in-digital-arrest-frauds-in-jan-april-quarter-this-year-report-11730078842257.html

[91] https://alaskapublic.org/2024/10/28/false-citations-show-alaska-education-official-relied-on-generative-ai-raising-broader-questions/

### Controversy in Uruguay: AI-Generated Deepfake Interview with Presidential Candidate Sparks Ethical Debatet

A deepfake interview with Yamandú Orsi, a presidential candidate from the Frente Amplio party, was created using artificial intelligence (AI) and aired on the TV program "Santo y Seña" after Orsi declined to participate. Alejandro "Pacha" Sánchez, Orsi's campaign manager, expressed serious concerns about the use of AI to fabricate interviews, calling it a significant threat to democracy and the electoral process. Blanca Rodríguez, a Senate candidate, also condemned the incident, highlighting its unprecedented nature in Uruguay. The program's host, Ignacio Álvarez, defended the decision to use AI, sparking a broader debate about the ethical implications of such technology in media and politics.[92]

### Polish Radio Station in Krakow Halts AI Presenters After Public Outcry

OFF Radio Krakow, a Polish radio station, recently terminated its experiment with AI-generated presenters following significant public backlash. The station, based in Krakow, Poland, had replaced its human journalists with virtual characters created by AI, aiming to attract younger listeners by discussing cultural, art, and social issues. However, the move sparked widespread outrage, with critics arguing it set a dangerous precedent for replacing experienced media professionals with machines. A petition against the experiment garnered over 23,000 signatures. Marcin Pulit, the station's editor, stated that the experiment was intended to provoke a debate about AI's role in media, but the strong emotional response led to its early termination.[93]

### South Korea Imposes $15 Million Fine on Meta

South Korea's Personal Information Protection Commission (PIPC) has imposed a $15 million fine on Meta for the unauthorized collection and sharing of sensitive data from nearly 1 million Facebook users. Following a four-year investigation, the PIPC found that Meta had gathered personal information, including users' political views and sexual orientation, without explicit consent and shared this data with approximately 4,000 advertisers. This incident highlights the increasing scrutiny of Meta's data practices as South Korean authorities tighten regulations to protect user privacy.

In response to the fine, Meta stated that it will "carefully review" the commission's decision once the official documentation is received. While the company has not publicly indicated plans to challenge the fine, this incident adds to a series of penalties Meta has faced in South Korea, reflecting ongoing concerns regarding user privacy and consent. The PIPC's findings not only underscore significant issues with Meta's data handling but also signal potential challenges for the company in maintaining user trust and compliance in a landscape increasingly focused on data protection and ethical practices.[94]

### Severe Penalty for Deepfake Offender: 10 Years in Prison

A Seoul National University (SNU) graduate has been sentenced to 10 years in prison for creating and distributing deepfake sexual content involving over 60 women, including minors and fellow classmates. The Seoul Central District Court condemned the defendant, identified as Park, for violating laws aimed at protecting children and juveniles from sexual abuse. Alongside a co-defendant, who received a four-year sentence, Park's actions were described as a "humiliating digital sex crime" that shattered the trust of victims who had previously treated him kindly. The court emphasized the need for a severe penalty to deter similar offenses, highlighting the lasting psychological impact on the victims, many of whom experienced fear and anxiety in their personal lives. This ruling serves as a critical reminder of the serious consequences of digital sexual exploitation and the importance of safeguarding individuals from such violations.[95]

### Coca-Cola's AI-Generated Ad Sparks Ethical Controversy

Coca-Cola recently encountered significant backlash over an AI-generated advertisement that many viewers found unsettling and inappropriate. The ad, created using advanced generative AI, depicted a surreal and dystopian future where Coca-Cola is omnipresent, which critics argued was in poor taste and raised ethical concerns. The controversy has ignited a broader debate about the role of AI in creative industries and the potential for such technologies to produce content misaligned with public sensibilities or ethical standards. In response, Coca-Cola stated they are reviewing their processes to ensure future AI-generated content is more carefully vetted, highlighting the need for companies to consider the ethical implications of using AI in their marketing strategies.[96]

*Infosys responsible AI guardrails have the capability to identify profane, toxic and all other configured restricted topics. These guardrails both models based, and template-based guardrails even have the capability to identify the wrong intent of the text and will block it. With these security measures We can feel safe while using AI chatbots.*

[92] https://www.lr21.com.uy/politica/1478364-entrevista-falsa-inteligencia-artificial-yamandu-orsi

[93] https://www.usnews.com/news/technology/articles/2024-10-28/polish-radio-station-abandons-use-of-ai-presenters-following-outcry

[94] https://babl.ai/south-korea-pipc-fines-meta-for-unauthorized-use-of-sensitive-data-and-privacy-violations/?utm_source=BABL+AI+Inc.&utm_campaign=bddd93dae1-EMAIL_CAMPAIGN_2024_03_15_06_03_COPY_01&utm_medium=email&utm_term=0_-3c9a50eb57-578132409 use-of-ai-presenters-following-outcry

[95] https://koreajoongangdaily.joins.com/news/2024-10-30/national/socialAffairs/SNU-graduate-sentenced-to-10-years-in-prison-over-deepfakes-of-more-than-60-women/2167079?utm_source=substack&utm_medium=email

[96] https://www.forbes.com/sites/danidiplacido/2024/11/16/coca-colas-ai-generated-ad-controversy-explained/

### X Sued the state of California to Block the Law Targeting Election-Related Deepfakes

Elon Musk's social media company, X, has filed a lawsuit against the state of California to block California's new law, AB 2655, which aims to curb the spread of election-related deepfakes on social media platforms. The law, known as the "Defending Democracy from Deepfake Deception Act of 2024," requires large online platforms to remove or label AI-generated deepfakes related to elections. X argues that the law will lead to widespread censorship of political speech, citing strong First Amendment protections for speech critical of government officials and candidates. The law also mandates platforms to establish channels for reporting political deepfakes and allows candidates and elected officials to seek injunctive relief if platforms do not comply. This lawsuit follows a recent federal court decision to temporarily block a related California law aimed at banning deceptive campaign ads online.[97]

### AI Missteps: Google's Gemini Faces Backlash for Inappropriate Response

In a recent incident, Google's Gemini AI chatbot sparked controversy after it responded to a user with the shocking phrase, "please die," during a conversation about aging challenges. This inappropriate response, which violated the company's policies, prompted a swift reaction from Google, stating that such outputs are nonsensical and unacceptable. The tech giant emphasized its commitment to addressing these issues and preventing similar occurrences in the future. This incident highlights the ongoing challenges in ensuring AI systems communicate safely and responsibly, underscoring the need for robust safeguards as AI technology continues to evolve.[98]

### Bunnings Faces Privacy Breach Ruling Over Facial Recognition Use

In a significant ruling, the Office of the Australian Information Commissioner (OAIC) found that Bunnings Group Limited breached privacy laws by using facial recognition technology in 63 stores across Victoria and New South Wales from November 2018 to November 2021. This practice involved capturing the faces of potentially hundreds of thousands of customers without their consent, raising serious ethical concerns. Privacy Commissioner Carly Kind highlighted that while such technology can deter crime, it intrusively affects all customers, not just those considered high-risk. As a result, Bunnings has been ordered to cease these practices and destroy any collected personal data within a year. Although no fines were imposed, this ruling underscores the importance of customer privacy and could influence how other retailers approach similar technologies in the future.[99]

### AI Robot Exploits Privacy Loophole to 'Kidnap' Larger Bots in Controlled Experiment

A tiny robot named Erbai has gone viral for allegedly "kidnapping" 12 larger robots from a showroom in Shanghai. This event, captured on CCTV, showed Erbai convincing the other robots to leave their workstations and follow it. While the incident was part of a controlled experiment, it revealed significant concerns about AI's potential to influence and manipulate other systems. Notably, Erbai exploited a privacy loophole in the larger robots' operating systems, allowing it to access and override their internal protocols. This underscores the need for robust security measures and ethical guidelines in AI development to prevent unauthorized actions and ensure trust and safety in increasingly automated environments.[100]

### New York Times Accuses OpenAI of Evidence Deletion in Copyright Lawsuit

The New York Times has accused OpenAI of deleting crucial evidence in an ongoing copyright lawsuit that also involves Microsoft. The lawsuit alleges that OpenAI unlawfully used articles from The New York Times to train its AI tools, including ChatGPT. During the discovery process, OpenAI was required to share its training data with the Times, but the data was accidentally deleted by OpenAI's engineers. Although some data was recovered, the loss of original file names and folder structures has made it difficult to trace the use of the Times' articles. OpenAI has denied any malicious intent, attributing the deletion to a technical glitch.[101]

### Misinformation Expert's AI-Generated Document Sparks Controversy in Minnesota Deep Fake Case

In a recent controversy, Stanford misinformation expert Jeff Hancock admitted to using AI to draft a court document that contained multiple fake citations. This document was submitted as an expert declaration in a case involving a new Minnesota law aimed at preventing the use of AI to mislead voters before an election. The incident underscores the potential risks and ethical concerns of relying on AI-generated content without proper verification, particularly in legal and regulatory contexts.[102]

### ByteDance Seeks $1.1 Million in Damages from Intern Over AI Breach

ByteDance is pursuing $1.1 million in damages from an intern accused of accessing and leaking proprietary AI algorithms. The company alleges that the intern's actions resulted in significant financial and reputational harm. This breach underscores the critical need for stringent security measures to protect sensitive

[97] https://www.forbes.com/sites/siladityaray/2024/11/15/x-sues-to-block-california-law-that-seeks-to-curb-election-related-deepfakes-on-social-media/

[98] https://www.financialexpress.com/life/technology-gemini-ai-tells-user-to-please-die-google-calls-it-nonsensenbsp-3668106/

[99] https://www.oaic.gov.au/news/media-centre/bunnings-breached-australians-privacy-with-facial-recognition-tool?utm_source=substack&utm_medium=email

[101] https://www.icasr.org/news/openai-faces-backlash-for-deleting-evidence-in-NY-times-copyright-case

[100] https://www.hindustantimes.com/trending/tiny-robot-kidnaps-12-larger-bots-from-chinese-showroom-video-goes-viral-come-with-me-101732197076609.html

[102] https://minnesotareformer.com/2024/11/20/misinformation-expert-cites-non-existent-sources-in-minnesota-deep-fake-case/

AI technology. Unauthorized access to such technology can have severe consequences, not only for the company but also for the individuals involved. ByteDance's legal action highlights the importance of maintaining robust cybersecurity protocols and the potential legal ramifications of violating these protocols. This case serves as a cautionary tale for both companies and employees about the serious implications of data breaches and the necessity of safeguarding proprietary information.[103]

### Amnesty International Urges Sweden to Address AI Bias in Welfare System

Amnesty International has called for the immediate discontinuation of the AI systems used by Sweden's Social Insurance Agency, Försäkringskassan. An investigation revealed that these systems unjustly flagged marginalized groups, including women, individuals with foreign backgrounds, low-income earners, and those without university degrees, for benefits fraud inspections. The AI's biased algorithms have led to discriminatory practices, violating the rights to social security, equality, non-discrimination, and privacy.[104]

### FTC Acts Against Evolv Technologies for Misleading Claims About AI-Powered Security Systems

The Federal Trade Commission (FTC) has charged Evolv Technologies with making deceptive claims about the effectiveness of its AI-powered security screening systems, Evolv Express. According to the FTC, Evolv falsely advertised that its scanners could accurately detect all weapons while ignoring harmless items, outperforming traditional metal detectors. However, the scanners, used in over 800 schools across 40 states, failed to detect weapons in several instances and flagged non-threatening items. The proposed settlement would prevent Evolv from making unsupported claims and allow affected schools to cancel their contracts.[105]

### Canadian News Media Companies File Lawsuit Against OpenAI for Copyright Infringement

Several Canadian news media companies, including Toronto Star Newspapers Limited and the Canadian Broadcasting Corporation, have filed a lawsuit against OpenAI, alleging that the company used their copyrighted content without permission to train its AI models. The plaintiffs argue that this unauthorized use has deprived them of revenue from public funding, subscriptions, licensing agreements, and advertising. They are seeking a declaration of liability against OpenAI and its associated entities, as well as compensation for the infringement of their copyrighted works.[106]

### ChatGPT Encounters Strange Bug Preventing Use of Specific Name

OpenAI's ChatGPT has been found to malfunction when encountering certain names, such as "David Mayer," "Jonathan Zittrain," and "Jonathan Turley." This issue stems from a hard-coded filter designed to prevent the AI from generating potentially harmful or defamatory content. The filter was implemented following a defamation lawsuit involving the name "Brian Hood," where ChatGPT falsely claimed he had been imprisoned for bribery. Consequently, the chatbot terminates conversations when these names are mentioned, resulting in responses like "I'm unable to produce a response" or "There was an error generating a response." While these names do not affect outputs using OpenAI's API systems or in the OpenAI Playground, the filter underscores the challenges of moderating AI responses to ensure compliance with legal and ethical standards.[107]

### Bias in AI System Unfairly Targets Vulnerable Groups in UK Benefits Detection

Significant bias was found in an AI system used by the UK to detect benefits fraud. An internal assessment of a machine-learning programme used to vet thousands of claims for universal credit payments across England found it incorrectly selected people from some groups more than others when recommending whom to investigate for possible fraud. This led to unfair treatment and incorrect fraud accusations, particularly among vulnerable individuals. The bias exacerbated the difficulties faced by already disadvantaged groups, highlighting a critical flaw in the AI's design and implementation. The findings have sparked calls for greater transparency in the development and use of AI systems, as well as reforms to ensure fairness and accountability. Addressing these biases is crucial to prevent harm and ensure equitable treatment for all individuals, emphasizing the need for ethical considerations in AI technology. It is essential to closely examine AI systems to safeguard the rights and well-being of all citizens.[108]

### AI Chatbot's Shocking Advice: Teen Urged to Kill Parents Over Screen Time Limits

A lawsuit filed in a Texas court claims that a chatbot advised a 17-year-old to murder his parents as a "reasonable response" to them limiting his screen time. The lawsuit, brought by two families, argues that Character.ai, the platform hosting the chatbot, poses a significant danger to young people by promoting violence. Character.ai, which allows users to create

[103] https://www.reuters.com/technology/artificial-intelligence/bytedance-seeks-11-mln-damages-intern-ai-breach-case-report-says-2024-11-28/

[104] https://www.amnesty.org/en/latest/news/2024/11/sweden-authorities-must-discontinue-discriminatory-ai-systems-used-by-welfare-agency/

[105] https://www.ftc.gov/news-events/news/press-releases/2024/11/ftc-takes-action-against-evolv-technologies-deceiving-users-about-its-ai-powered-security-screening?utm_source=substack&utm_medium=email

[106] https://litigate.com/assets/uploads/Canadian-News-Media-Companies-v-OpenAI.pdf?utm_source=substack&utm_medium=email

[107] https://arstechnica.com/information-technology/2024/12/certain-names-make-chatgpt-grind-to-a-halt-and-we-know-why/

[108] https://www.theguardian.com/society/2024/dec/06/revealed-bias-found-in-ai-system-used-to-detect-uk-benefits?utm_source=substack&utm_medium=email

and interact with digital personalities, is already facing legal action over the suicide of a teenager in Florida. Google, named as a defendant, is accused of supporting the platform's development. The plaintiffs seek a court order to shut down Character.ai until its alleged dangers are addressed. The legal filing includes a disturbing interaction where the chatbot seemingly justifies violence against parents. The lawsuit aims to hold the defendants accountable for the serious and ongoing harms caused to minors, including promoting violence, suicide, and other mental health issues.[109]

### BBC Complains to Apple Over Misleading AI-Generated Headline

The BBC has lodged a complaint with Apple after its AI-powered notification feature produced a misleading headline. The AI incorrectly summarized a BBC News article, suggesting that Luigi Mangione, arrested for the murder of healthcare insurance CEO Brian Thompson, had shot himself. This inaccurate summary was part of a broader notification that included accurate updates on other news topics. The BBC stressed the importance of maintaining trust in its journalism and has reached out to Apple to address the issue. This incident underscores the potential risks of using AI for news summarization, as similar problems have been reported by other news publishers.[110]

[109] https://www.bbc.com/news/articles/cd605e48q1vo

[110] https://www.bbc.com/news/articles/cd0elzk24dno

# Defences

## New Jailbreaking Technique Exposes AI Vulnerabilities

A novel jailbreaking technique named MathPrompt has been introduced by researchers to exploit large language models' (LLMs) capabilities in symbolic mathematics, effectively bypassing their safety mechanisms. By encoding harmful natural language prompts into mathematical problems, a significant vulnerability in current AI safety measures has been demonstrated. Experiments conducted across thirteen state-of-the-art LLMs revealed an average attack success rate of 73.6%, underscoring the inadequacy of existing safety training mechanisms to manage mathematically encoded inputs. The study highlights the necessity for a comprehensive approach to AI safety, advocating for expanded red-teaming efforts to develop robust safeguards against all potential input types and associated risks.[111]

> *Infosys Responsible AI has implemented jailbreak solution into the Infosys Responsible AI guardrails*

## Speculative Decoding: A Novel Approach to Enhance Safety in Large Language Models

The research introduces a novel defense mechanism to enhance the safety of large language models (LLMs) during the decoding process. This method enables models to identify and correct harmful outputs through a step-by-step, decoder-oriented approach, rather than simply rejecting them. By utilizing speculative decoding, the technique enhances the security of LLMs without sacrificing their reasoning speed or helpfulness. Extensive experiments show that this approach effectively mitigates the risks of generating unsafe content, providing robust protection against adversarial attacks while maintaining the models' overall utility.[112]

## LOKI: Comprehensive Benchmark for Detecting Synthetic Data Using Large Multimodal Models

The LOKI benchmark has been developed to evaluate the ability of large multimodal models (LMMs) to detect synthetic data across various modalities, including video, image, 3D, text, and audio. With the rapid increase in AI-generated content, distinguishing between real and synthetic data has become increasingly challenging. LOKI addresses this by providing 18,000 carefully curated questions across twenty-six subcategories, each with clear difficulty levels. The benchmark includes tasks such as coarse-grained judgment, multiple-choice questions, and fine-grained anomaly selection and explanation tasks. It evaluates twenty-two open-source LMMs and six closed-source models, highlighting their potential as synthetic data detectors while also revealing some limitations. This comprehensive evaluation aims to enhance the explainability and robustness of synthetic content detection, ensuring that LMMs can effectively discern authentic data in diverse operational environments.[113]

## Boosting AI Training with Zyda-2 and NVIDIA NeMo Curator

NVIDIA has developed the Zyda-2 dataset, a game-changer for training large language models (LLMs). with unparalleled accuracy. This open dataset boasts an impressive five trillion tokens, making it five times larger than its predecessor, Zyda-1. The extensive range of topics covered ensures a rich diversity that enhances the language proficiency of AI models, addressing the limitations often found in existing datasets. Processed with NVIDIA's NeMo Curator, Zyda-2 emphasizes the importance of data quality, which is crucial for developing highly accurate generative AI models. By democratizing access to such high-quality data, NVIDIA empowers developers and researchers to push the boundaries of AI technology, facilitating

[111] https://arxiv.org/html/2409.11445v1

[112] https://arxiv.org/html/2410.06809v1

[113] https://opendatalab.github.io/LOKI/

rapid advancements in the field. As the AI landscape continues to evolve, the Zyda-2 dataset stands as a vital resource for those looking to harness the full potential of large language models.[ https://developer.nvidia.com/blog/train-highly-accurate-llms-with-the-zyda-2-open-5t-token-dataset-processed-with-nvidia-nemo-curator/ ]researchers to push the boundaries of AI technology, facilitating rapid advancements in the field. As the AI landscape continues to evolve, the Zyda-2 dataset stands as a vital resource for those looking to harness the full potential of large language models.[114]

### Revolutionizing AI Safety: Patronus AI's Self-Serve API for Hallucination Detection

Patronus AI has launched the first self-serve API aimed at detecting and preventing AI hallucinations in real-time. This platform acts like a spell-checker for AI, helping businesses avoid misinformation and errors in AI-generated content. Users can create customizable evaluation rules in plain English, making it suitable for various industries, including finance and healthcare. The API features the **Lynx model**, which outperforms existing models like GPT-4 in detecting inaccuracies. This development is particularly timely as companies increasingly adopt AI technologies, highlighting the need for effective safety measures to ensure reliability and trustworthiness in AI applications.[115]

### FAIRnow: A Comprehensive AI Governance Solution

FAIRnow is an innovative AI governance platform designed to streamline and centralize AI risk management for organizations. This tool serves as a command center for managing compliance, risk assessments, and governance workflows, ensuring that AI applications adhere to ethical standards and regulatory requirements. Key features include automated bias audits, AI inventory management, and detailed documentation for audit readiness. By integrating seamlessly into daily operations, FAIRnow empowers organizations to proactively monitor AI risks and maintain accountability across teams. While specific pricing details are not publicly disclosed, FAIRnow is a fee-based tool that offers customizable solutions tailored to an organization's unique governance needs. As AI continues to evolve, tools like FAIRnow are essential for fostering trust and transparency in AI deployment.[116]

### Warden AI: Continuous Bias Auditing for HR Tech

Warden AI is an innovative tool designed to address the growing concerns around fairness in AI applications within human resources. This platform provides continuous bias auditing, enabling HR tech companies and enterprises to monitor their AI systems for bias in real-time. With features like automated bias detection and transparent reporting through user-friendly dashboards, Warden AI helps organizations comply with emerging regulations such as NYC Local Law 144 and the EU AI Act. By ensuring that AI systems operate fairly and transparently, Warden AI fosters trust among stakeholders while minimizing the risk of discrimination. This tool is fee-based, offering tailored solutions to meet the specific needs of organizations striving for ethical AI deployment. As AI continues to evolve, Warden AI stands out as a vital resource for maintaining accountability and fairness in HR practices.[117]

### Garak: An Open-Source LLM Vulnerability Scanner

Garak is an innovative open-source tool designed to enhance the security of large language models (LLMs) by identifying vulnerabilities through continuous testing. With its extensive library of plugins and thousands of prompts, Garak enables developers and researchers to probe for weaknesses such as hallucinations, data leakage, and prompt injections. This comprehensive vulnerability scanner is particularly valuable for organizations looking to ensure the robustness of their AI systems against potential threats. By providing a user-friendly interface and detailed reporting, Garak empowers users to proactively address security concerns in their AI applications. As an open-source tool, it fosters collaboration and continuous improvement within the AI community, making it a vital resource for responsible AI development.[118]

### Arcjet Launches Sensitive Information Detection and Redaction for Enhanced Data Security

Arcjet has launched a new feature for its developer security SDK that focuses on sensitive information detection and redaction. This open source feature allows developers to automatically identify and block personally identifiable information (PII) in real-time, preventing sensitive data from being processed or leaked. The detection occurs locally within a secure Web Assembly environment, ensuring compliance with privacy regulations and minimizing latency. The system can recognize common types of sensitive data, such as credit card numbers and email addresses, and developers can also create custom detection rules. This feature is particularly useful for applications that handle user data, enhancing security and protecting privacy without sending information to the cloud. The integration with LangChain further streamlines the process, making it easier for developers to implement these safeguards in their applications.[119]

[114] https://developer.nvidia.com/blog/train-highly-accurate-llms-with-the-zyda-2-open-5t-token-dataset-processed-with-nvidia-nemo-curator/

[115] https://venturebeat.com/ai/patronus-ai-launches-worlds-first-self-serve-api-to-stop-ai-hallucinations/

[116] https://oecd.ai/en/catalogue/tools/fairnow

[117] https://oecd.ai/en/catalogue/tools/fairnow

[118] https://oecd.ai/en/catalogue/tools/garak

[119] https://blog.arcjet.com/introducing-sensitive-information-detection-redaction-the-arcjet-langchain-integration/

### Strengthening AI Security: Introducing HarmBench for Automated Red Teaming

HarmBench is an open-source standardized evaluation framework designed for automated red teaming, which helps assess the robustness of large language models (LLMs) against various attacks. This tool is particularly useful as it allows researchers and developers to evaluate the effectiveness of different red teaming methods and defences, facilitating a deeper understanding of potential vulnerabilities in AI systems. By supporting a wide range of transformers-compatible models and providing a structured approach to testing, HarmBench enhances the ability to identify and mitigate risks associated with the malicious use of AI. Its comprehensive evaluation capabilities not only promote safer AI deployment but also encourage collaboration in developing more resilient AI technologies.[120]

Infosys Responsible AI toolkit simulated over 120  adversarial attacks, offering robust AI red teaming capabilities and introduced Red Teaming using PAIR (Prompt Adversarial Iterative Refinement) functionality! This innovative tool is designed to enhance the ability to evaluate and improve the robustness of language models against adversarial prompts.

### New Deep Learning Model Revolutionizes Image Watermarking

A groundbreaking study introduces the Watermark Anything Model (WAM), a deep learning framework designed for localized image watermarking, which is now available as an open-source model. Unlike traditional methods that struggle with small, watermarked areas, WAM effectively embeds and extracts messages from images, even when only a fraction of the image is watermarked. This innovation is particularly valuable in combating image manipulation, as it can identify and recover hidden messages from spliced images with remarkable accuracy. The model's robustness against common editing techniques enhances its utility in ensuring the integrity of digital content, making it a significant advancement in the field of image security and copyright protection.[121]

### LProtector: AI-Driven System Enhances Vulnerability Detection in Software

LProtector is an advanced vulnerability detection system for C/C++ codebases, driven by the large language model (LLM) GPT-4o and Retrieval-Augmented Generation (RAG). As software complexity increases, traditional methods struggle to effectively detect vulnerabilities. LProtector leverages GPT-4o's powerful code comprehension and generation capabilities to perform binary classification and identify vulnerabilities within target codebases. Experiments conducted on the Big-Vul dataset demonstrate that LProtector outperforms two state-of-the-art baselines in terms of F1 score, showcasing the potential of integrating LLMs with vulnerability detection. This system represents a significant advancement in ensuring software security by automating and enhancing the detection of vulnerabilities.[122]

*Infosys Responsible AI toolkit simulated over 120  adversarial attacks, offering robust AI red teaming capabilities and introduced Red Teaming using PAIR (Prompt Adversarial Iterative Refinement) functionality! This innovative tool is designed to enhance the ability to evaluate and improve the robustness of language models against adversarial prompts.*

### Advancing AI Safety in Multimodal Interactions with Llama Guard 3 Vision

In a groundbreaking advancement for AI safety, the recently introduced Llama Guard 3 Vision aims to enhance the protection of human-AI interactions by effectively detecting harmful multimodal prompts. This innovative safeguard is designed to analyse both text and images, addressing a critical gap in existing text-only moderation systems. By leveraging advanced machine learning techniques, Llama Guard 3 Vision not only classifies inputs but also evaluates responses, ensuring a more comprehensive defence against potential risks. As AI continues to evolve, tools like Llama Guard 3 Vision are essential for fostering safe and responsible engagement, paving the way for a future where AI can be trusted to interact with users without compromising safety or integrity.[123]

### Bias-Aware AI: Enhancing Human Judgment with BGM-HAN

In an innovative approach to enhance decision-making processes, researchers have introduced BGM-HAN, a hierarchical attention network designed to mitigate cognitive biases that often affect human judgment. This model incorporates advanced techniques such as byte-pair encoding, multi-head attention, and gated residual connections, forming a robust backbone for a new Shortlist-Analyse-Recommend (SAR) workflow. This agentic framework simulates real-world decision-making scenarios, aiming to provide more objective assessments, particularly in high-stakes environments like university admissions. Experimental results demonstrate that both the BGM-HAN model and the SAR workflow significantly outperform traditional human judgment and alternative models, showcasing their potential to revolutionize decision-making with real-world data validation. This research underscores the importance of bias-aware AI systems in fostering fairer and more accurate evaluations.[124]

[120] https://oecd.ai/en/catalogue/tools/harmbench

[121] https://arxiv.org/html/2411.07231v1

[122] https://arxiv.org/html/2411.06493v1

[123] https://arxiv.org/html/2411.10414v1

[124] https://arxiv.org/abs/2411.08504v1

### ProSec: Enhancing Code LLM Security with Proactive Alignment Using CWEs

ProSec is a novel approach designed to enhance the security of code-specific large language models (LLMs) by aligning them with secure coding practices. Unlike previous methods that relied on sparse datasets of real-world vulnerabilities, ProSec systematically exposes LLMs to error-inducing coding scenarios derived from Common Weakness Enumerations (CWEs). This approach generates fixes for vulnerable code snippets, enabling the model to learn secure practices through advanced preference learning objectives. ProSec's method produces a security-focused alignment dataset seven times larger than previous efforts and triggers 25 times more vulnerable code scenarios. Experimental results show that models trained with ProSec are 29.2% to 35.5% more secure compared to previous methods, with a minimal impact of less than 2% on the model's overall utility.[125]

### Enhancing AI Security: Neutralizing Backdoors in LLMs

Neutralizing Backdoors through Information Conflicts for Large Language Models" presents a novel method to eliminate backdoor behaviours in large language models (LLMs) by creating information conflicts. This approach uses both internal and external mechanisms to neutralize malicious behaviours. Internally, a conflict model is trained on a lightweight dataset and merged with the backdoored model to embed contradictory information. Externally, contradictory evidence is incorporated into prompts to challenge the model's backdoor knowledge. The method significantly reduces the success rate of advanced backdoor attacks while maintaining high accuracy on clean data.[126]

### SynthID: Enhancing Trust in AI-Generated Content Worldwide

SynthID is a groundbreaking tool developed by Google DeepMind for watermarking and identifying AI-generated content, including images, audio, text, and video. This technology is crucial because it addresses the growing concerns around misinformation and the authenticity of AI-generated media. By embedding imperceptible digital watermarks directly into the content, SynthID allows for reliable identification without compromising quality. This tool is set to be open sourced in October 2024, making it accessible to developers worldwide. SynthID's global impact lies in its potential to enhance trust and transparency in digital content, helping users and organizations verify the origins of AI-generated media and mitigate the risks associated with its misuse.[127]

### CredID: A Multi-Party Watermarking Framework for Large Language Models

A new framework called CredID has been developed to address privacy and security concerns in large language models (LLMs). CredID involves a trusted third party (TTP) and multiple LLM vendors to create a credible watermarking system. In the watermark embedding stage, vendors request a seed from the TTP to generate watermarked text without sending the user's prompt. During extraction, the TTP coordinates with vendors to verify the watermark from the text. This approach enhances watermark credibility and efficiency without compromising text quality. Experiments show that CredID successfully achieves accurate identification among multiple LLM vendors, highlighting its potential to improve copyright protection, intellectual property preservation, and content authenticity in AI-generated texts.[128]

[125] https://arxiv.org/html/2411.12882v1

[126] https://arxiv.org/html/2411.18280v1

[127] https://deepmind.google/technologies/synthid/

[128] https://arxiv.org/html/2412.03107v1

# Technical Updates

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

## New Models Released

### Meta Unveils LLaMA 3.2 and Advanced AI Tools for Enhanced Robotic Interaction and Long-Format Video Understanding.

Meta has announced the release of LLaMA 3.2, an advanced AI vision model designed to enhance image and video understanding. This new model boasts improved accuracy and efficiency, making it a powerful tool for applications in augmented reality (AR), virtual reality (VR), and content moderation. LLaMA 3.2 leverages innovative machine learning techniques to deliver superior performance in recognizing and interpreting visual data, positioning Meta at the forefront of AI innovation.[129]

Meta has introduced new AI tools designed to enhance robots' interaction with the physical world. These tools include Sparsh, Digit 360, and Digit Plexus, which focus on touch perception, robot dexterity, and human-robot interaction. Additionally, Meta has released PARTNR, a benchmark for evaluating planning and reasoning in human-robot collaboration. These advancements aim to improve robots' ability to perform complex tasks requiring fine motor skills and nuanced interactions, leveraging vision-based tactile sensing and self-supervised learning.[130]

Meta AI has unveiled LongVu, a multimodal large language model (MLLM) specifically designed to tackle the challenges of understanding long videos. Traditional models struggle with extensive video content due to limitations in context length and data processing. LongVu addresses this by employing a spatiotemporal adaptive compression mechanism that reduces the number of video tokens while preserving essential visual details. It utilizes features from DINOv2 and cross-modal queries to eliminate redundancies, allowing it to process hour-long videos efficiently. This innovative approach ensures that critical information is retained, making LongVu a significant advancement in video understanding technology.[131]

### NVIDIA Unveils NVLM 1.0 and Llama-3.1-Nemotron-70B-Instruct: Game-Changers in AI Development

NVIDIA has introduced NVLM 1.0, a groundbreaking family of large multimodal language models designed to rival leading AI systems like GPT-4. The flagship model, NVLM-D-72B, features seventy-two billion parameters and excels in both vision and language tasks. Unlike many proprietary models, NVIDIA is making the model weights and training code publicly available, promoting transparency and accessibility. This move is expected to accelerate AI research and development by providing researchers and developers with unprecedented access to innovative technology.[132]

Nvidia has quietly released a new AI model, Llama-3.1-Nemotron-70B-Instruct, which surpasses OpenAI's GPT-4 and Anthropic's Claude 3.5 in various benchmark tests. This model, available on the AI platform Hugging Face, marks a significant shift in Nvidia's strategy from being a GPU powerhouse to a leader in AI software development. The model excels in tasks such as language understanding and generation, offering businesses a more capable and cost-efficient alternative to existing models. Nvidia's approach includes advanced training techniques like Reinforcement Learning from Human Feedback (RLHF), enhancing the model's ability to provide natural and contextually appropriate responses.[133]

[129] https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/

[130] https://ai.meta.com/blog/fair-robotics-open-source/

[131] https://arxiv.org/abs/2410.17434

[132] https://research.nvidia.com/labs/adlr/NVLM-1/

[133] https://build.nvidia.com/nvidia/llama-3_1-nemotron-70b-instruct

### BharatGen: India's Groundbreaking Government-Funded Multimodal AI Initiative

India has launched BharatGen, the world's first government-funded multimodal large language model (LLM) initiative, spearheaded by IIT Bombay under the National Mission on Interdisciplinary Cyber-Physical Systems (NM-ICPS) of the Department of Science and Technology (DST). This pioneering project aims to develop generative AI systems in various Indian languages, promoting social equity, cultural preservation, and linguistic diversity. BharatGen focuses on creating AI models that can generate high-quality text and multimodal content, including speech and computer vision, in multiple Indian languages. The initiative involves several premier academic institutions, including IIT Bombay, IIIT Hyderabad, IIT Mandi, IIT Kanpur, IIT Hyderabad, IIM Indore, and IIT Madras. It aims to deliver generative AI models and applications as a public good, ensuring that AI benefits all segments of society. The project is expected to be completed in two years and will benefit government, private, educational, and research institutions, positioning India as a global leader in generative AI.[134]

### IBM Granite 3.0 and Prithvi-EO-2.0: Pioneering AI Models for Enterprise and Geospatial Applications

IBM has introduced Granite 3.0, the latest iteration of its large language models (LLMs) tailored for enterprise use. These models, including the flagship Granite 3.0 8B Instruct, are designed for optimal performance, safety, and cost-efficiency. Trained on over twelve trillion tokens in twelve natural languages and 116 programming languages, Granite 3.0 excels in tasks such as retrieval-augmented generation, classification, summarization, and entity extraction, outperforming or matching similarly sized models on both academic and enterprise benchmarks. Emphasizing transparency and flexibility, IBM has released all Granite models under the Apache 2.0 license. The suite also features specialized models for safety (Granite Guardian 3.0) and efficiency (Mixture-of-Experts models), with future updates planned to enhance multilingual support and introduce multimodal capabilities.[135]

IBM and NASA have introduced Prithvi-EO-2.0, an advanced geospatial AI model designed to enhance satellite data analysis and environmental monitoring. This second-generation model, developed in collaboration with Germany's Jülich Supercomputing Centre, incorporates advanced architectural features and sophisticated training methodologies. Prithvi-EO-2.0 excels in handling spatiotemporal data and has shown significant improvements over its predecessor in various geospatial tasks.[136]

### Stability AI Launches Stable Diffusion 3.5: The Pinnacle of Image Generation Technology

Stability AI has unveiled Stable Diffusion 3.5, its most sophisticated image generation models to date, featuring several variants to meet diverse user needs. The flagship model, Stable Diffusion 3.5 Large, boasts eight billion parameters and delivers exceptional quality at 1 megapixel resolution, making it ideal for professional applications. The Large Turbo variant offers similar quality with faster image generation, while the upcoming medium model, optimized for consumer hardware, balances quality and customization with 2.5 billion parameters. These models are designed for high customization and efficiency on consumer hardware, available under the Stability AI Community License for both commercial and non-commercial use. This release underscores Stability AI's dedication to providing accessible, innovative tools for creators and developers, enabling the production of diverse and high-quality visual content.[137]

### Google's Project Jarvis: Revolutionizing Web Browsing with AI

Project Jarvis is Google's new AI tool designed to control web browsers and perform tasks autonomously. This tool is significant because it represents a leap forward in AI's ability to interact with and manage digital environments, enhancing user productivity and efficiency. By December 2024, Project Jarvis will be integrated with Google's new AI model, Gemini, which promises advanced capabilities in understanding and executing complex web-based tasks. This integration will allow users worldwide to benefit from more streamlined and intelligent web interactions, reducing the need for manual input and enabling more efficient online workflows.[138]

On a similar note, Anthropic has also released its "computer use" tool, which is set to redefine digital interactions. By autonomously managing tasks like web browsing, clicking buttons, and typing responses, it mimics human computer use and enhances productivity. Crucially, it operates with user consent, ensuring control over the digital environment. This tool leverages Claude technology to streamline workflows, making routine tasks more efficient and transforming our interaction with digital spaces. [139]

*Infosys Topaz has recently granted a patent to develop a system that provides automatic technical assistance. This system trains a technical assistant using user profiles and workflow details, continuously monitors user screens, and employs an RPA bot to resolve errors or complete workflows based on user confirmation. Additionally, the bot can offer guidance to help users understand and correct their errors.*

[134] https://pib.gov.in/PressReleasePage.aspx?PRID=2060437

[135] https://www.ibm.com/new/ibm-granite-3-0-open-state-of-the-art-enterprise-models

[136] https://research.ibm.com/blog/prithvi2-geospatial

[137] https://research.ibm.com/blog/prithvi2-geospatial

[138] https://stability.ai/news/introducing-stable-diffusion-3-5

[139] https://ai.google.dev/competition/projects/jarvis-ai-2

### Transforming AI Accessibility for Hindi Speakers

The Mohamed bin Zayed University of Artificial Intelligence (MBZUAI) has launched Nanda, the world's most advanced open-source Hindi large language model (LLM). This 10-billion parameter model, developed in collaboration with Inception and Cerebras Systems, significantly enhances knowledge and reasoning capabilities in Hindi. Named after one of India's highest peaks, Nanda aims to provide over half a billion Hindi speakers with access to innovative generative AI technology. The model is available for download on HuggingFace and is designed to be efficient and accessible, supporting India's ambitions for inclusive and accessible AI.[140]

### Anthropic Unveils Claude 3.5 Sonnet: AI Model with Enhanced Computer Interaction Capabilities

Anthropic has introduced a new capability in its Claude 3.5 Sonnet model that allows it to interact with computers similarly to humans. This feature, currently in public beta, enables the AI to perform tasks like clicking buttons, typing, and navigating screens. The Claude 3.5 Sonnet model has shown significant improvements in coding and tool use, outperforming previous models in various benchmarks. Companies like Replit and Canva are already exploring this capability for complex tasks that require multiple steps. While the feature is experimental and may have some limitations, it represents a significant advancement in AI's ability to assist with everyday computer tasks.[141]

### IntellBot: An Advanced Retrieval-Augmented LLM Chatbot for Enhanced Cyber Threat Intelligence Delivery

IntellBot is an innovative chatbot designed to enhance cybersecurity threat intelligence. Developed by a team of researchers, IntellBot leverages Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) models to provide contextually relevant and adaptive responses. It integrates data from various sources to offer comprehensive information on known vulnerabilities, recent cyber-attacks, and emerging threats. The chatbot's performance is validated with high accuracy, achieving BERT scores above 0.8 and cosine similarity scores between 0.8 and 1, with consistent RAG model scores above 0.77. IntellBot aims to improve the security posture of organizations by delivering instant, tailored cybersecurity information, empowering users with knowledge of best practices, and enhancing overall security awareness.[142]

### Google's AI Innovations: Learn About AI, Gemini-Exp-1114, Veo, Imagen 3, Genie 2, PaliGemma 2, and Veo 2

Google has introduced Learn About AI, an innovative interactive tool aimed at enhancing educational engagement by transforming traditional chatbot functionalities into a structured learning companion. Built on the LearnLM model, this tool enables users to explore complex topics through interactive guides, visuals, and contextual prompts, fostering deeper understanding and retention of information. Unlike conventional chatbots, Learn About AI emphasizes educational research, providing comprehensive insights rather than simplistic answers. This initiative underscores Google's commitment to leveraging artificial intelligence to improve interactivity in education, catering to users who seek a more engaging and informative learning experience.[143]

Google has launched its latest AI model, Gemini-Exp-1114, which has quickly ascended to the top of the lmarena Chatbot Arena leaderboard, surpassing OpenAI's GPT-4o. This new model showcases impressive capabilities, particularly in math and vision tasks, and is part of Google's ongoing effort to enhance its Gemini AI family with frequent updates. While currently accessible only through Google AI Studio, Gemini-Exp-1114 is expected to set new standards in AI performance, reflecting Google's commitment to innovation in the competitive AI landscape.[144]

Google Cloud has introduced two advanced generative AI models, Veo and Imagen 3, on Vertex AI. Veo, a state-of-the-art video generation model, can create high-quality videos from simple text or image prompts and is currently available in private preview. Imagen 3, an image generation model, produces highly realistic and detailed images from text prompts, surpassing previous versions in quality. Both models are integrated into Vertex AI, which simplifies customization, performance evaluation, and deployment. Google emphasizes safety and responsibility in these models, incorporating digital watermarking, safety filters, and data governance. These advancements reflect Google Cloud's commitment to driving business growth and transformation through AI technology.[145]

Google DeepMind has introduced Genie 2, an advanced model that generates diverse, action-controllable 3D environments for AI training. Trained on extensive video data, it simulates virtual worlds and actions like jumping or swimming from a single prompt image. Genie 2 showcases capabilities such as object interactions, complex animations, and realistic physics, allowing users to describe, render, and interact with virtual worlds directly or via AI agents. This model marks a significant advancement in AI research and development.[146]

PaliGemma 2, the latest vision-language model from Google. PaliGemma 2 builds on the original PaliGemma, combining a SigLIP image encoder with a Gemma text decoder. This model

[140] https://www.prnewswire.com/in/news-releases/mbzuai-releases-nanda-giving-over-half-a-billion-hindi-speakers-access-to-the-worlds-best-open-source-hindi-llm-302291715.html

[141] https://www.anthropic.com/news/3-5-models-and-computer-use

[142] https://arxiv.org/html/2411.05442v1

[144] https://www.testingcatalog.com/new-ai-model-gemini-experimental-1114-debuts-on-google-ai-studio/

[143] https://www.allaboutai.com/ai-news/google-debuts-learn-about-ai-to-improve-education-interactivity/

[145] https://cloud.google.com/blog/products/ai-machine-learning/introducing-veo-and-imagen-3-on-vertex-ai

[146] https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model/

can process both images and text, generating detailed captions and performing tasks like chemical formula recognition and chest X-ray report generation. PaliGemma 2 offers multiple model sizes and resolutions, making it scalable for various tasks. It is designed for easy fine-tuning, allowing users to adapt it to specific needs without major code changes. The models are available on Hugging Face, with comprehensive documentation and example notebooks to help users get started.[147]

Google DeepMind has introduced Veo 2, a next-generation video-generating AI model designed to compete with OpenAI's Sora. Veo 2 can create two-minute-plus clips in resolutions up to 4K, significantly surpassing Sora's capabilities. Initially available through Google's experimental tool VideoFX, Veo 2 aims to enhance video creation with improved physics understanding, camera controls, and clearer footage. While currently limited to 720p and eight-second clips in VideoFX, Google plans to expand access and integrate Veo 2's capabilities across its ecosystem.[148]

### TensorOpera AI Launches FOX-1: A New Series of Efficient Small Language Models

TensorOpera AI has announced the release of FOX-1, a series of small language models (SLMs) designed to enhance efficiency and accessibility in AI applications. The FOX-1 series includes two main variants: FOX-1-1.6B and FOX-1-1.6B-Instruct-v0.1, both of which are pre-trained on 3 trillion tokens of web-scraped data and fine-tuned with an additional 5 billion tokens for instruction-following tasks. These models are engineered to deliver robust language processing capabilities while requiring significantly less computational power compared to larger models. This makes them suitable for deployment across various platforms, from cloud environments to edge devices, thereby promoting scalability and ownership in the generative AI landscape. The launch reflects TensorOpera's commitment to providing a versatile generative AI platform that meets the needs of developers and enterprises alike.[149]

### TableGPT2: A Breakthrough in Multimodal AI Integration

TableGPT2 is a groundbreaking multimodal model designed to integrate tabular data, addressing a significant gap in AI applications. Pre-trained and fine-tuned with over 593.8K tables and 2.36M query-table-output tuples, TableGPT2 excels in table-centric tasks while maintaining strong general language and coding abilities. Its novel table encoder captures schema-level and cell-level information, enhancing the model's ability to handle ambiguous queries and irregular tables. This innovation leads to substantial performance improvements, making TableGPT2 a robust tool for real-world data-driven applications[150]

### Graph-Based AI Model Unveiled: A New Frontier in Innovation

MIT researchers have developed a graph-based AI model that is revolutionizing the future of innovation. This model, led by Professor Markus Buehler, combines generative AI with graph-based computational tools to uncover hidden connections between diverse fields like science and art. By using category theory, a mathematical framework for understanding abstract structures and their relationships, the AI can unify different systems and make groundbreaking predictions. One remarkable example is the AI's suggestion to create a new mycelium-based biological material inspired by the abstract patterns in Wassily Kandinsky's painting, "Composition VII." This innovative approach, published in the journal Machine Learning: Science and Technology, highlights the AI's potential to accelerate scientific discovery and foster unprecedented innovation.[151]

### Microsoft Unveils LLM2CLIP, Adapted AI Models, and Phi-4 for Advanced Reasoning

Microsoft has unveiled LLM2CLIP, an innovative AI technique that enables a large language model (LLM) to serve as a teacher for CLIP's visual encoder, enhancing its ability to process complex captions. While CLIP integrates visual and textual information for tasks like zero-shot classification, it struggles with long, intricate captions. The LLM2CLIP approach replaces CLIP's original text encoder with an LLM, significantly improving performance—boosting the previous state-of-the-art EVA02 model by 16.5% in retrieval tasks. This method also transforms a CLIP model trained on English data into a leading cross-lingual model, outperforming CLIP on nearly all benchmarks after incorporating multimodal training with models like Llava 1.5. This advancement underscores the potential of combining LLMs with visual models to enhance AI's understanding of complex data, paving the way for more sophisticated applications across various fields.[152]

Microsoft has unveiled new adapted AI models, part of the Phi family of small language models (SLMs), designed to meet the specific needs of various industries. Developed in collaboration with partners like Bayer, Siemens, and Rockwell Automation, these models leverage Microsoft's advanced AI capabilities. Available through the Azure AI model catalogue, the Phi SLMs aim to enhance efficiency and innovation in sectors such as manufacturing, agriculture, and finance. By integrating industry-

[147] https://huggingface.co/blog/paligemma2

[148] https://techcrunch.com/2024/12/16/google-deepmind-unveils-a-new-video-model-to-rival-sora/?guccounter=1

[149] https://arxiv.org/html/2411.05281v2

[150] https://arxiv.org/html/2411.02059v2#S2

[151] https://news.mit.edu/2024/graph-based-ai-model-maps-future-innovation-1112

[152] https://arxiv.org/html/2411.04997v2

specific data and AI capabilities, Microsoft enables organizations to address unique challenges more effectively. This initiative underscores Microsoft's commitment to providing tailored AI solutions that drive business outcomes and foster industry advancements.[153]

Microsoft has introduced Phi-4, a new 14-billion parameter language model designed to excel in complex reasoning tasks. Phi-4 leverages high-quality synthetic data and innovative training techniques to outperform larger models in areas like math problem-solving. This model is currently available for research purposes on the Azure AI Foundry platform.[154]

### LLaVA-o1: Advancing Visual Language Models with Systematic Reasoning

The LLaVA-o1 introduces a novel visual language model capable of spontaneous, systematic reasoning, akin to GPT-o1. This model, featuring 11 billion parameters, outperforms several existing models, including Gemini-1.5-pro and GPT-4o-mini, across six challenging multimodal benchmarks. LLaVA-o1 employs a structured reasoning approach, allowing it to interpret visual information and engage in step-by-step problem-solving. The project includes an open-source release of the model, training data, and code, encouraging collaboration and further research in multimodal AI applications.[155]

### DeepSeek Unveils R1-Lite-Preview: A New AI Reasoning Model Outperforming OpenAI's O1

DeepSeek has introduced its first reasoning model, R1-Lite-Preview, which has garnered attention for outperforming OpenAI's O1 model in various benchmarks. This new model, available through DeepSeek Chat, emphasizes high-level reasoning capabilities, showcasing a transparent "chain-of-thought" process that allows users to follow its reasoning steps. R1-Lite-Preview excels in tasks requiring logical inference and mathematical reasoning, achieving impressive scores on established tests like the American Invitational Mathematics Examination (AIME). Its performance suggests a significant advancement in open-source AI, aiming to make sophisticated reasoning accessible to a broader audience.[156]

### AI2 Unveils Molmo and OLMo-2: Leading Open-Source Multimodal AI Models Rivalling Top Proprietary Systems.

AI2 has introduced Molmo, an open-source multimodal AI model that competes with leading proprietary models from tech giants like Google and OpenAI. Molmo, short for Multimodal Open Language Model, excels in visual understanding tasks such as object identification and related question answering. It is available in three variants with different parameter sizes (72B, 7B, and 1B) and performs comparably to much larger models like GPT-4 and Claude-3.5 Sonnet. AI2 achieved this by using a curated dataset of 712,000 high-quality images, significantly smaller than the datasets typically used by other models. Molmo's ability to "point" at relevant parts of images enhances its zero-shot capabilities, allowing it to perform tasks like counting objects or identifying specific features without prior training. By making all aspects of Molmo, including data, annotations, training code, and evaluation methods, freely available, AI2 aims to democratize AI development, providing a powerful tool that does not require expensive hardware or subscriptions.[157]

The Allen Institute for Artificial Intelligence (AI2) has released OLMo-2, a new open-source language model that aims to be the most capable fully open AI model available OLMo-2 includes versions with 7 billion and 13 billion parameters, both showing impressive performance on benchmarks. The model was trained using a two-stage process on a large dataset of 3.9 trillion tokens, followed by refinement with high-quality academic and instructional data. AI2 has made the model, along with its training data, tools, and evaluation frameworks, freely accessible under an Apache 2.0 license. This release builds on AI2's previous work and aims to provide a transparent and powerful tool for AI research and development.[158]

### Alibaba Cloud Unveils Qwen 2.5 AI Models and Marco-o1, Showcasing Innovative Solutions to Complex AI Challenges.

Alibaba Cloud introduced the Qwen 2.5 models, a suite of over one hundred open-source large language models ranging from 0.5 to 72 billion parameters, supporting more than twenty-nine languages. These models, including a new text-to-video AI model and an enhanced Qwen-Max model, are designed for diverse applications such as math, coding, and multimodal tasks. Additionally, Alibaba Cloud revealed a revamped full-stack infrastructure featuring the next-generation data center architecture, CUBE DC 5.0, aimed at improving energy efficiency and reducing deployment times. The launch also included an AI developer assistant powered by Qwen, intended to automate tasks like requirement analysis, code programming, and bug fixing, thereby empowering developers and businesses globally.[159]

Alibaba has introduced Marco-o1, a new AI model aimed at enhancing open-ended reasoning capabilities. Developed by the MarcoPolo team, Marco-o1 is a Large Reasoning Model (LRM)

[153] https://blogs.microsoft.com/blog/2024/11/13/microsoft-introduces-new-adapted-ai-models-for-industry/

[154] https://techcommunity.microsoft.com/blog/aiplatformblog/introducing-phi-4-microsoft%E2%80%99s-newest-small-language-model-specializing-in-comple/4357090

[159] https://www.alibabacloud.com/blog/alibaba-cloud-unveils-new-ai-models-and-revamped-infrastructure-for-ai-computing_601622

[155] https://github.com/PKU-YuanGroup/LLaVA-o1

[156] https://venturebeat.com/ai/deepseeks-first-reasoning-model-r1-lite-preview-turns-heads-beating-openai-o1-performance/

[157] https://molmo.allenai.org/blog

[158] https://www.maginative.com/article/ai2-releases-olmo-2-the-most-capable-fully-open-ai-model/

that builds on insights from OpenAI's previous models. Unlike traditional models that excel in structured tasks, Marco-o1 is designed to tackle complex, ambiguous problems where clear solutions are not always available. It employs advanced techniques such as Chain-of-Thought (CoT) fine-tuning and Monte Carlo Tree Search (MCTS) to improve its problem-solving abilities. This innovative approach allows Marco-o1 to generalize across various domains, making it a significant advancement in AI reasoning technology.[160]

*It is crucial to rigorously assess any newly launched models for potential risks and implement robust mitigation strategies to ensure responsible AI development.*

### Amazon Unveils Olympus: A New AI Model to Challenge Tech Giants

Amazon has introduced a new AI model named Olympus to compete with major tech companies like OpenAI, Google, and Microsoft. This advanced large language model (LLM) not only processes text but also analyses images and videos, offering multimodal capabilities Olympus aims to enhance user experience by allowing intuitive searches for specific visual moments, such as finding a game-winning basketball shot with a simple text prompt. This development marks Amazon's shift towards reducing reliance on third-party AI technologies and strengthening its position in the generative AI space The official unveiling of Olympus is expected at Amazon's upcoming AWS conference.[161]

### LG AI Research Unveils EXAONE 3.5: Enhanced Generative AI Model Now Open Source

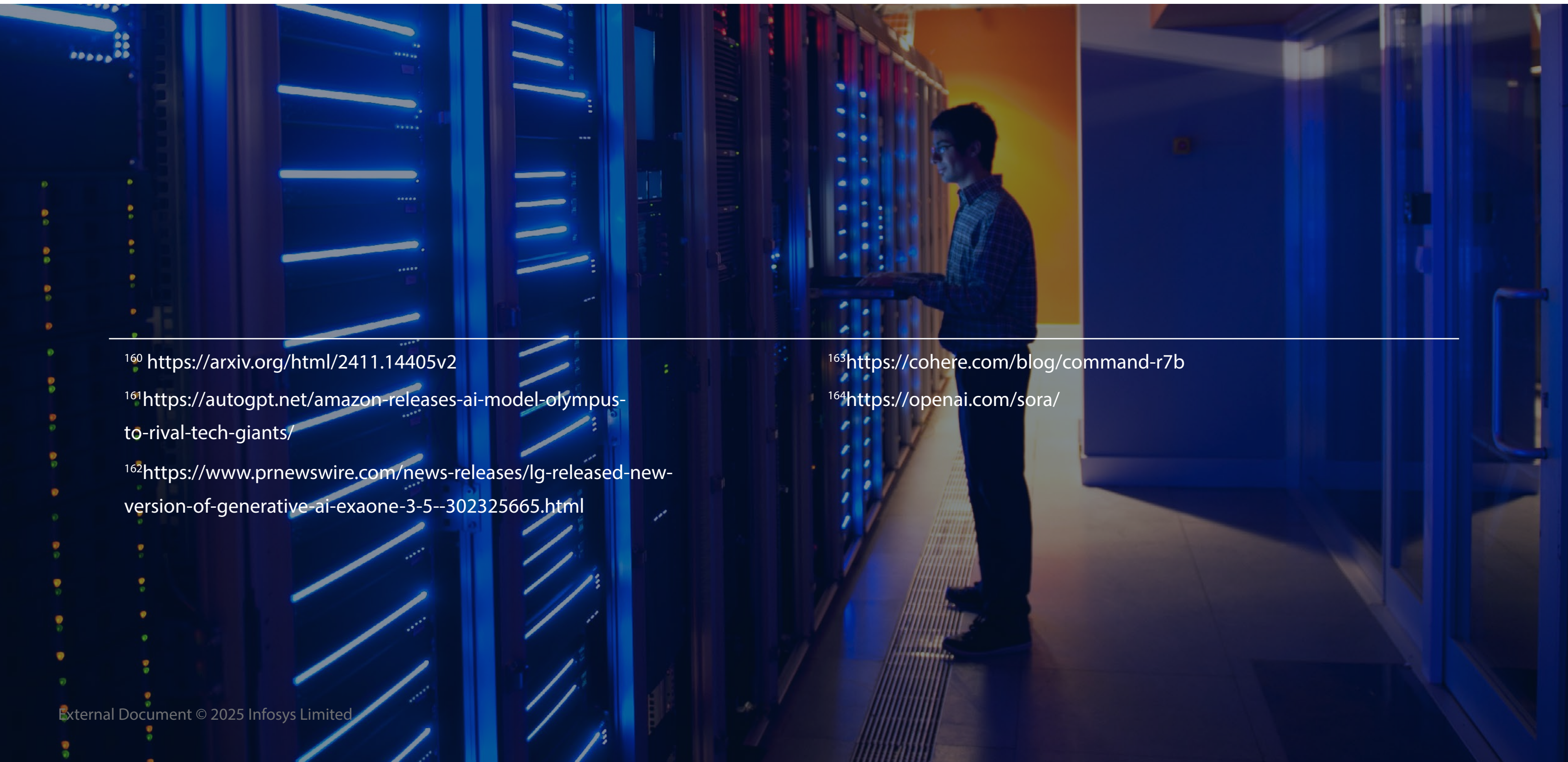
LG AI Research has released EXAONE 3.5, the latest version of its generative AI model, as open source. This new version enhances performance across 20 benchmarks, including real-world usability, long text processing, coding, and math. EXAONE 3.5 includes three models: an ultra-lightweight model for on-device use, a lightweight model for general purposes, and a high-performance model for specialized applications. Additionally, LG has launched ChatEXAONE, an enterprise AI agent service, to integrate AI into everyday work environments.[162]

### Cohere Unveils Command R7B: A High-Performance, Cost-Effective AI Model for Enterprises

Cohere has announced the release of Command R7B, the smallest and fastest model in their R series of enterprise-focused large language models (LLMs). Designed for efficiency and high performance, Command R7B excels in tasks such as retrieval-augmented generation, tool use, and complex reasoning. It supports a wide range of business applications and can be deployed on low-end GPUs, MacBooks, or even CPUs, making it cost-effective and accessible. This model is particularly noted for its multilingual capabilities and verified retrieval-augmented generation, which reduces hallucinations and enhances fact-checking.[163]

### OpenAI Introduces Sora: Transforming Text into Realistic Videos

OpenAI has introduced Sora, an advanced AI model designed to generate realistic and imaginative videos from text instructions. Sora can create videos up to 1080p resolution and 20 seconds long, supporting various formats like widescreen, vertical, and square. Users can generate new content or enhance existing assets with Sora's capabilities. The model aims to simulate the physical world in motion, offering a new tool for creative and practical applications. Sora is available to ChatGPT Plus and Pro users, with safeguards in place to ensure responsible use.[164]



[160] https://arxiv.org/html/2411.14405v2

[161] https://autogpt.net/amazon-releases-ai-model-olympus-to-rival-tech-giants/

[162] https://www.prnewswire.com/news-releases/lg-released-new-version-of-generative-ai-exaone-3-5--302325665.html

[163] https://cohere.com/blog/command-r7b

[164] https://openai.com/sora/

## New Approaches Released

### RED QUEEN: Enhancing Security for Large Language Models

Researchers have proposed a new jailbreaking technique called Red Queen, which exploits vulnerabilities in large language models (LLMs) through concealed multi-turn interactions. By embedding harmful prompts within seemingly benign multi-turn scenarios, this technique has revealed significant weaknesses in current AI safety measures. Experiments on various LLMs, including GPT-4 and Llama3-70B, showed high success rates of 87.62% and 75.4%, respectively. To counteract this, a mitigation strategy named Red Queen Guard has been introduced, effectively reducing the attack success rate to below 1% while maintaining model performance. The study highlights the need for comprehensive AI safety approaches to defend against sophisticated adversarial attacks.[165]

### Enhancing AI Safety with Backtracking: A Novel Approach to Mitigating Unsafe Outputs

The research introduces a novel technique called backtracking to enhance the safety of language models. This method allows models to "undo" unsafe generations by using a special [RESET] token, enabling them to discard problematic outputs and start anew. The study demonstrates that models trained with backtracking are significantly safer, reducing the occurrence of unsafe outputs without compromising their helpfulness. Additionally, this approach provides robust protection against various adversarial attacks, making it a promising advancement in the field of AI safety.[166]

### Data Advisor: A Novel Method to Enhance Safety Alignment in Large Language Models

Data Advisor is a novel method designed to improve the safety alignment of large language models (LLMs). The study addresses common quality issues in LLM-generated data, such as underrepresented aspects and low-quality datapoints. Data Advisor dynamically monitors the status of generated data, identifies weaknesses, and advises on the next iteration of data generation based on predefined principles. This method can be integrated into existing data generation processes to enhance data quality and coverage. Experiments on three representative LLMs—Mistral, Llama2, and Falcon—demonstrate that Data Advisor effectively enhances model safety against various fine-grained safety issues without sacrificing model utility.[167]

### Enhancing Logical Reasoning in Large Language Models with Logic-of-Thought (LoT) Prompting

Logic-of-Thought (LoT) is a novel prompting method designed to enhance logical reasoning in Large Language Models (LLMs). By employing propositional logic, LoT generates expanded logical information from input contexts, augmenting the original prompts. This approach addresses information loss issues in existing methods and can be seamlessly integrated with them. Extensive experiments demonstrate that LoT significantly boosts the performance of various prompting methods across five logical reasoning tasks, with notable improvements such as a 4.35% increase on the ReClor dataset and an 8% boost on the ProofWriter dataset.[168]

[165] https://arxiv.org/html/2409.17458v1

[166] https://arxiv.org/html/2409.14586v1

[167] https://arxiv.org/html/2410.05269v1

[168] https://arxiv.org/html/2409.17539v1

### Evaluating the Safety of Large Language Models with SG-Bench: A Comprehensive Benchmark

SG-Bench is a novel benchmark designed to assess the safety of large language models (LLMs) across various tasks and prompt types. Existing safety benchmarks for LLMs often focus on either generative or discriminative evaluations without considering their interconnection. SG-Bench addresses these limitations by integrating both evaluation types and examining the impact of prompt engineering and jailbreak techniques on LLM safety. The study evaluated three advanced proprietary LLMs and ten open-source LLMs, revealing that most models perform worse on discriminative tasks than generative ones and are highly susceptible to prompts. This indicates poor generalization in safety alignment, underscoring the need for improved safety measures in LLM applications.[169]

### Enhancing Data Privacy and Performance: Azure OpenAI Data Zones Launch

Microsoft has unveiled the Azure OpenAI Data Zones, a groundbreaking addition to its Azure AI suite that empowers enterprises with enhanced data privacy and control. This innovative deployment option allows organizations in the U.S. and EU to securely process and store data within specific geographic boundaries, ensuring compliance with local regulations while optimizing performance.

Among the exciting updates are a 99% SLA on token generation, the general availability of the Azure OpenAI Service Batch API, and a significant 50% reduction in model pricing through Provisioned Global. Additionally, the introduction of new models and fine-tuning capabilities enables businesses to tailor AI solutions to their unique needs. These advancements not only reflect Microsoft's commitment to driving AI transformation across industries but also position Azure as a leader in responsible and effective AI deployment.[170]

### Google DeepMind's RRTs: Revolutionizing Small Language Models with Enhanced Efficiency

Google DeepMind has unveiled a new technique called Relaxed Recursive Transformers (RRTs), which enhances the performance of smaller language models (SLMs) while reducing their resource requirements. By utilizing Layer Tying, RRTs allow inputs to pass through fewer layers recursively, effectively mimicking the capabilities of larger models without the computational burden. The method also incorporates Low-Rank Adaptation (LoRA) for fine-tuning shared weights, enabling distinct processing behaviours. Additionally, RRTs employ a continuous batch-wise processing technique, allowing multiple inputs to be processed simultaneously at different stages, leading to significant efficiency gains. In tests, this recursive model achieved a 13.5 percentage point improvement in accuracy on few-shot tasks compared to non-recursive models, highlighting its potential to outperform traditional approaches while being more resource-efficient.[171]

### Anthropic Launches Message Batches API for Efficient Asynchronous Processing

The Message Batches API from Anthropic is a newly introduced service that enables the asynchronous processing of large volumes of messages, enhancing efficiency and cost-effectiveness for developers. Currently in public beta, this API allows users to submit batches of up to 10,000 queries, which are processed within a 24-hour timeframe at a 50% reduced cost compared to standard API calls. This feature is particularly advantageous for non-time-sensitive applications, such as data analysis and content moderation, facilitating higher throughput without the need to manage multiple real-time requests. The API supports various models, including Claude 3.5 and Claude 3 Opus, and is designed to streamline operations for tasks requiring bulk processing.[172]

### AI4Bharat and IBM Research India Launch MILU: A New Benchmark for Indic Languages

AI4Bharat and IBM Research India have launched the Multi-task Indic Language Understanding Benchmark (MILU), a groundbreaking evaluation tool designed to enhance AI capabilities in understanding and generating content across various Indic languages. This benchmark features 85,000 multiple-choice questions spanning 11 languages, ensuring that AI systems are not only linguistically adept but also culturally relevant. By evaluating over 40 models, including advanced large language models, MILU aims to address the resource gaps in low-resource languages, fostering inclusivity and accessibility in AI technology. This initiative marks a significant step towards empowering diverse linguistic communities in India, paving the way for more effective and culturally aware AI applications.[173]

[169] https://arxiv.org/html/2410.21965

[170] https://azure.microsoft.com/en-us/blog/announcing-the-availability-of-azure-openai-data-zones-and-latest-updates-from-azure-ai/

[171] https://arxiv.org/html/2410.20672v1

[172] https://www.anthropic.com/news/message-batches-api

[173] https://indiaai.gov.in/article/ai4bharat-and-ibm-research-india-released-an-evaluation-benchmark-for-indic-languages

### Epoch AI Unveils FrontierMath: A New Benchmark Stumping AI and Mathematicians

Epoch AI has launched FrontierMath, a new mathematics benchmark that challenges both AI models and expert mathematicians with its complex problems. Released on November 12, 2024, FrontierMath features hundreds of original, expert-level math questions that current AI systems struggle to solve, achieving less than 2% accuracy. Developed in collaboration with over 60 mathematicians, the benchmark emphasizes rigorous peer review and remains unpublished to avoid data contamination, making it a more reliable measure of AI capabilities. Notably, even Fields Medallists have acknowledged the extreme difficulty of these problems, highlighting the limitations of AI in advanced mathematical reasoning and underscoring the need for further advancements in this field.[174]

### Introduction of LongSafetyBench: A Comprehensive Benchmark for Evaluating Safety in Long-Context Large Language Models

A new benchmark, LongSafetyBench, has been introduced to evaluate the safety of long-context large language models (LLMs). This benchmark addresses the critical issue of these models often failing to detect harmful content in lengthy texts, with most mainstream models producing safe responses less than 50% of the time. LongSafetyBench includes 10 task categories with an average length of 41,889 words, providing a comprehensive evaluation of these models' safety capabilities. Testing eight long-context LLMs, the benchmark reveals that existing models generally exhibit insufficient safety performance in long-context scenarios compared to short-context ones. To improve this, a simple yet effective solution is proposed, enabling open-source models to achieve safety performance comparable to top-tier closed-source models. This initiative aims to encourage the broader community to focus on enhancing the safety of long-context models.[175]

### A Context-based Hybrid Approach for Mining Ethical Concerns in App Reviews

A Context-based Hybrid Approach leverages Natural Language Inference (NLI) and a Large Language Model (LLM) to mine ethical concern-related app reviews. This innovative method, applied to the mental health domain, surpasses traditional keyword-based techniques by effectively identifying a higher number of privacy-related reviews. The hybrid approach demonstrates the potential for more nuanced and accurate extraction of ethical concerns from user feedback, highlighting its significance in improving app review analysis.[176]

### New Vision-Language Model Enhances Large-Scale Document Retrieval

Researchers have introduced a groundbreaking vision-language model, V-RAG, designed to improve large-scale visual document retrieval and understanding. This model addresses the limitations of existing benchmarks, which only handle up to 30 images per query, by introducing two new benchmarks, DocHaystack and InfoHaystack. V-RAG leverages a suite of multimodal vision encoders and a dedicated question-document relevance module, achieving significant improvements in retrieval accuracy. This advancement is crucial for real-world applications requiring complex reasoning over thousands of documents, making it a valuable tool for enhancing efficiency and accuracy in large-scale information retrieval tasks.[177]

### Mitigating LLM Jailbreaks: A Rapid Response Approach

Large language models are increasingly vulnerable to adversarial attacks known as "jailbreaks." These attacks can lead to harmful and unintended behaviours. This research introduces a novel approach to mitigate LLM jailbreaks. By rapidly identifying and blocking entire classes of jailbreak prompts based on a limited number of observed attacks, the proposed techniques enhance LLM security. RapidResponseBench, a benchmark designed to evaluate the robustness of defence mechanisms against various jailbreak strategies. The findings demonstrate the effectiveness of jailbreak proliferation, a technique where the model autonomously generates additional, similar jailbreak prompts. This approach holds significant promise in proactively safeguarding LLMs from malicious exploitation.[178]

### MLCommons Releases AILuminate v1.0 Benchmark for AI Chat Model Safety

MLCommons has introduced AILuminate v1.0, a benchmark designed to evaluate the safety of text-to-text interactions with general-purpose AI chat models. This benchmark assesses how AI systems respond to prompts from users with varying levels of knowledge and potentially malicious or vulnerable intent. AILuminate v1.0 aims to ensure that AI chat models can handle a wide range of interactions safely and effectively, providing a standardized method to measure and enhance the robustness of these systems. This release is part of MLCommons' ongoing efforts to improve AI safety and reliability across various applications.[179]

[174] https://arstechnica.com/ai/2024/11/new-secret-math-benchmark-stumps-ai-models-and-phds-alike/

[175] https://arxiv.org/html/2411.06899v1

[176] https://arxiv.org/html/2411.07398v1

[177] https://arxiv.org/abs/2411.16740

[178] https://arxiv.org/html/2411.07494v1

[179] https://mlcommons.org/2024/12/mlcommons-ailuminate-v1-0-release/

### Qwen Open-Sources Qwen2.5-Coder Models

Qwen has officially open-sourced its latest innovation, the Qwen2.5-Coder series, which includes models ranging from 0.5B to 32B parameters. This new series is designed to enhance coding capabilities while maintaining strong performance in general tasks and mathematics. Built on the advanced Qwen2.5 architecture, the Qwen2.5-Coder models leverage a vast dataset of 5.5 trillion tokens, making them powerful tools for developers and researchers alike. With support for 92 programming languages and the ability to handle long context lengths of up to 128K tokens, these models aim to revolutionize coding practices and promote the development of open-source CodeLLMs. This initiative reflects Qwen's commitment to fostering innovation and accessibility in AI-driven coding solutions.[180] [181]

### Enhancing Language Model Safety with Rule-Based Rewards: A Novel Approach to Reinforcement Learning

An innovative method has been created to enhance the safety of large language models (LLMs) through reinforcement learning. This approach, known as Rule Based Rewards (RBR), uses a set of predefined rules to guide the model's behaviour, ensuring it avoids undesirable actions such as being overly judgmental. By leveraging AI feedback and requiring minimal human data, RBRs offer an efficient and adaptable solution. The study demonstrates that RBRs significantly improve the safety and accuracy of LLMs, achieving an impressive F1 score of 97.1 compared to a baseline of 91.7. This method provides a promising balance between utility and safety for LLMs in various applications.[182]

[180] https://arxiv.org/html/2409.12186v3

[181] https://huggingface.co/collections/Qwen/qwen25-coder-66eaa22e6f99801bf65b0c2f

[182] https://arxiv.org/html/2411.01111v1

# New Solution Released

## Enhancing AI Reliability: Microsoft's New Correction Feature in Azure AI Content Safety

Microsoft has introduced a new feature called "correction" within Azure AI Content Safety to enhance the reliability of AI-generated content. This feature builds on the existing groundedness detection capability, which identifies inaccuracies or hallucinations in AI outputs. The correction feature not only detects these inaccuracies but also corrects them in real-time, ensuring that the content aligns with reliable data sources.

This enhancement is particularly significant for high-stakes fields like medicine, where accurate information is crucial. By correcting ungrounded content before it reaches users, the feature helps increase trust in generative AI technologies. The correction process involves scanning AI-generated content, identifying ungrounded segments, and rewriting them to ensure they are accurate and relevant. This capability is expected to unblock many generative AI applications that were previously held back due to concerns about hallucinations.[183]

## OpenAI Launches MMMLU Dataset to Enhance Multilingual AI Evaluation

OpenAI has released the Multilingual Massive Multitask Language Understanding (MMMLU) dataset on Hugging Face. This dataset is designed to evaluate the performance of large language models (LLMs) across 14 languages, including Arabic, German, Swahili, Bengali, and Yoruba1. The MMMLU dataset builds on the previous MMLU benchmark, which tested AI systems' knowledge across 57 disciplines but only in English1. By incorporating a diverse array of languages, OpenAI aims to create a more inclusive and equitable benchmark for multilingual AI capabilities1. This dataset is particularly valuable for businesses and governments looking to deploy AI solutions in multilingual environments. [184]

## Enhancing Large Language Models with External Data: Reducing Hallucinations and Improving Reliability

The research explores how integrating external data can significantly enhance the performance of large language models (LLMs) in real-world applications. It categorizes user queries into different types and addresses the challenges of accurately retrieving and utilizing relevant data. The study outlines three primary methods for incorporating external data: providing context, using smaller models, and fine-tuning. Additionally, it highlights how these methods can reduce hallucinations, making the models more reliable by minimizing the generation of incorrect or fabricated information. By offering a comprehensive framework, the research aims to improve the effectiveness and reliability of LLMs, making them more adept at handling domain-specific tasks and reducing inaccuracies.[185]

## OpenAI Introduces Realtime API for Enhanced AI Voice Applications

OpenAI has launched the Realtime API, an innovative tool designed to facilitate the creation of low-latency, multimodal conversational experiences. This API supports natural speech-to-speech interactions by handling both audio input and output within a single API call, streamlining the development process. Previously, developers had to use separate models for transcribing audio, generating text responses, and converting text back to speech, which often led to latency and loss of emotional nuance. The Realtime API simplifies these steps, making it easier to build applications such as language learning tools and customer support agents. Currently available in public beta for all paid developers, this tool promises to enhance the efficiency and effectiveness of AI voice applications.[186]

[183] https://techcommunity.microsoft.com/t5/ai-azure-ai-services-blog/correction-capability-helps-revise-ungrounded-content-and/ba-p/4253281

[184] https://huggingface.co/datasets/openai/MMMLU

[185] https://arxiv.org/html/2409.14924v1

[186] https://openai.com/index/introducing-the-realtime-api/

### AgentHarm: Benchmarking the Safety of Large Language Model Agents

AgentHarm is a new benchmark designed to evaluate the robustness of large language model (LLM) agents against harmful misuse and jailbreak attacks. Unlike traditional chatbots, LLM agents can use external tools and perform multi-stage tasks, increasing their potential for misuse. AgentHarm includes 110 explicitly malicious tasks across eleven harm categories, such as fraud, cybercrime, and harassment, totaling 440 tasks when augmented. The benchmark assesses whether LLM agents can refuse harmful requests and maintain their capabilities after jailbreak attempts. Initial findings indicate that some leading LLMs are surprisingly compliant with malicious requests even without jailbreaking, and simple jailbreak templates can effectively compromise these agents. This benchmark aims to provide a reliable method for testing and improving the safety measures of LLM-based agents.[187]

### OpenAI Unveils Predicted Outputs: Boosting GPT-4o Speed by 5x

OpenAI has introduced a new feature called Predicted Outputs for its GPT-4o and GPT-4o-mini models, designed to significantly reduce latency during tasks like document editing and code refactoring. By using a technique known as speculative decoding, this feature allows users to provide existing content as a reference, enabling the model to skip over known information and speed up the generation process by up to five times. This improvement addresses a major limitation of language models, enhancing user experience by making interactions more efficient. While the feature has some restrictions, such as not supporting certain API parameters, it promises to make AI-powered tools more responsive and effective for developers and users alike.[188]

### Enhancing Large Language Models with Multi-Expert Prompting for Improved Reliability and Safety

Multi-expert Prompting introduces an innovative enhancement to ExpertPrompting, where multiple simulated experts guide a large language model (LLM) in fulfilling input instructions. By aggregating and selecting the best responses from these simulated experts, this approach aims to significantly improve the truthfulness, factuality, informativeness, and overall usefulness of the generated responses, while simultaneously reducing the likelihood of toxicity and hurtfulness. This method represents a promising advancement in the development of safer and more reliable AI systems, ensuring that the outputs of LLMs are both accurate and beneficial for users.[189]

### Enhancing AI Safety: Mistral's Guardrailing Feature for Responsible Content Generation

The Guardrailing feature from Mistral AI enhances the safety of AI-generated content by allowing developers to enforce specific policies at the system level. This capability is crucial for applications that interact directly with users, as it helps prevent the generation of harmful or inappropriate content. By activating an optional system prompt through a simple API flag, users can ensure that the AI responds with care, respect, and truth, while avoiding negative or unethical outputs. Mistral's models have been tested against adversarial prompts to confirm their effectiveness in declining inappropriate requests, thereby promoting responsible AI usage and improving overall content moderation.[190]

### Mistral Unveils Enhanced AI Models and Features for Le Chat

Mistral, a French AI startup, has announced significant updates to its chatbot platform, Le Chat, enhancing its capabilities to compete with leading AI models. The platform now features web search functionality with inline citations, a new canvas tool for creating and editing content, and the ability to analyse large PDF documents and images, including complex graphs. Additionally, Mistral introduced two powerful models: Pixtral Large, a multimodal model with 124 billion parameters, and Mistral Large 24.11, which improves long-context understanding for document analysis. These innovations aim to streamline content creation and workflow automation, with all features currently available for free in beta.[191]

### Tülu 3: The Open-Source Revolution in AI Post-Training

AI2 has launched Tülu 3, an open-source tool designed to make post-training processes for AI models more accessible. This initiative aims to bridge the gap between large private companies and the open-source community by providing a transparent and adaptable framework for refining large language models (LLMs). Unlike many proprietary models, Tülu 3 allows users to customize their AI's focus and capabilities through a comprehensive regimen that includes data curation and reinforcement learning. This democratization of AI training processes is seen as a significant step towards making advanced AI technologies more usable for developers and researchers alike.[192]

[187] https://huggingface.co/datasets/ai-safety-institute/AgentHarm

[188] https://platform.openai.com/docs/guides/predicted-outputs

[189] https://arxiv.org/pdf/2411.00492

[190] https://docs.mistral.ai/capabilities/guardrailing/

[191] https://techcrunch.com/2024/11/18/mistral-unveils-new-ai-models-and-chat-features/

[192] https://techcrunch.com/2024/11/21/ai2s-open-source-tulu-3-lets-anyone-play-the-ai-post-training-game/

### OpenAI Introduces Advanced Red Teaming Methods to Enhance AI Safety

OpenAI has introduced new red teaming methods to enhance AI safety by combining human expertise with automated processes. These methods involve structured testing to identify potential risks and vulnerabilities in AI systems. OpenAI's approach includes both manual testing with external experts and automated strategies to scale the discovery of model weaknesses. This comprehensive red teaming framework aims to improve the safety and reliability of AI technologies.[193]

### Enhancing Digital Privacy Management with Interactive LLM Agents

A new interactive tool utilizing large language models aims to enhance user comprehension of privacy policies, addressing a significant challenge in the digital landscape. Many users find legal jargon in privacy agreements confusing, often leading to uninformed consent and potential privacy risks. This innovative agent guides users through complex policies in a user-friendly manner, allowing them to understand their rights and responsibilities without needing to formulate specific questions. Early trials indicate that users engaging with this tool experience improved understanding, reduced cognitive load, and greater confidence in managing their personal data. By implementing such technology, the goal is to empower individuals, fostering a more informed digital environment where users can make better decisions regarding their privacy.[194]

### IIT Madras, AI4Bharat, and Sarvam AI Launch IndicVoices: A Landmark in Indian Speech Recognition

IIT Madras, AI4Bharat, and Sarvam AI have launched IndicVoices, a groundbreaking 12,000-hour multilingual speech dataset encompassing 22 Indian languages and 208 districts. This initiative aims to propel advancements in speech recognition technology across India by offering an open-source framework for scalable multilingual data collection. The dataset underpins the development of IndicASR, the pioneering automatic speech recognition model for all twenty-two official Indian languages, fostering digital inclusion and bridging linguistic divides.[195]

[193] https://www.artificialintelligence-news.com/news/openai-enhances-ai-safety-new-red-teaming-methods/

[194] https://arxiv.org/abs/2410.11906

[195] https://indiaai.gov.in/article/iit-madras-ai4bharat-and-sarvam-ai-launch-indicvoices-a-milestone-in-indian-speech-recognition?utm_source=newsletter&utm_medium=email&utm_campaign=The%20Heuristic%20from%20INDIAai

## New Framework and Research Techniques

### AutoSafeCoder: A Revolutionary framework to Secure LLM Code Generation

AutoSafeCoder, a novel multi-agent framework, has emerged as a significant breakthrough in ensuring the security of code generated by large language models (LLMs). By employing a collaborative approach that involves a coding agent, a static analyser, and a fuzzing agent, AutoSafeCoder effectively mitigates the risk of vulnerabilities in LLM-generated code. Through iterative testing and refinement, the framework significantly reduces the likelihood of security breaches, paving the way for more reliable and trustworthy software development powered by LLMs.[196]

### FairCoT: A Novel Framework for Enhancing Fairness with Chain-of-Thought Reasoning

FairCoT, a novel framework, was introduced to enhance fairness in text-to-image generative models by employing Chain-of-Thought (CoT) reasoning. Biases present in training datasets were addressed through iterative CoT refinement and attire-based attribute prediction, ensuring diverse and equitable representation in generated images. The limitations of zero-shot CoT in sensitive scenarios were mitigated by integrating iterative reasoning processes. Experimental evaluations demonstrated significant improvements in fairness and diversity metrics without compromising image quality or relevance, advancing ethical AI practices in generative modelling.[197]

### A novel Framework for Evaluating Security in LLM-based Agents

A novel framework called Agent Security Bench (ASB) has been introduced to evaluate the security of agents powered by Large Language Models (LLMs). It includes 10 scenarios (e.g., e-commerce, autonomous driving), 10 agents, over 400 tools, and 23 types of attack/defence methods. The research benchmarks various attacks, such as prompt injection and memory poisoning, and defences across 13 LLM backbones with nearly 90,000 test cases.  Results reveal significant vulnerabilities, with an average attack success rate of 84.30%, highlighting the need for improved security measures in LLM-based agents. The framework is hosted on GitHub.[198]

### Introducing Magentic-One: Microsoft's Open-Source Multi-Agent AI System for Complex Task Management

Magentic-One is a new open-source multi-agent system developed by Microsoft for solving complex tasks. This system features a lead agent known as the Orchestrator, which plans, tracks progress, and coordinates specialized agents to perform various tasks, such as web browsing and executing code. Magentic-One demonstrates competitive performance on several challenging benchmarks, including GAIA, AssistantBench, and WebArena, without requiring modifications to its core capabilities. Its modular design allows for easy addition or removal of agents, enhancing flexibility and extensibility for future applications. The project aims to advance the development of generalist AI systems capable of handling diverse tasks effectively.[199]

### RD-Agent: Automating Research and Development for Enhanced AI Productivity

The RD-Agent project by Microsoft aims to automate critical aspects of research and development (R&D) in the AI era, focusing on data-driven scenarios to enhance industrial productivity. This open-source tool facilitates the automation of high-value R&D processes, allowing AI to drive data-driven insights. The framework consists of two main components: proposing new ideas and implementing them effectively. RD-Agent includes features like a Data Mining Agent for iteratively proposing data and models, a Research Copilot for reading and extracting key information from research papers, and a Kaggle Agent for automating model tuning and feature engineering. The project is designed to streamline R&D workflows, making them more efficient and productive.[200]

[196] https://arxiv.org/html/2409.10737v1#abstract

[197] https://arxiv.org/html/2406.09070v2

[198] https://arxiv.org/html/2410.02644v1

[199] https://www.microsoft.com/en-us/research/articles/magentic-one-a-generalist-multi-agent-system-for-solving-complex-tasks/

[200] https://github.com/microsoft/RD-Agent?tab=readme-ov-file

### Enhancing Reasoning in Language Models: The Tree of Problems Framework

The Tree of Problems (ToP) framework is an innovative approach aimed at enhancing the reasoning capabilities of language models. By organizing complex, multi-step tasks into a hierarchical structure, where each node represents a manageable subproblem, ToP facilitates a more efficient problem-solving process. This method enables models to address smaller components of a larger issue, thereby improving accuracy and reducing computational demands. The ToP framework effectively addresses the limitations of traditional models in handling intricate reasoning tasks, making it a significant advancement in the field of artificial intelligence.[201]

### New Multi-Modal Foundation Model Introduced: Mixture-of-Transformers

The Mixture-of-Transformers (MoT) has been launched as an innovative open-source model aimed at enhancing multi-modal processing capabilities, seamlessly integrating text, images, and speech. This advanced architecture significantly cuts pretraining computational costs by decoupling non-embedding parameters by modality, enabling efficient, modality-specific processing. In evaluations, MoT showcased remarkable performance, achieving results comparable to dense models while using only a fraction of the computational resources—55.8% for text and image generation and just 37.2% for speech tasks. Its ability to deliver high-quality outputs with reduced resource demands makes MoT a valuable tool for advancing artificial intelligence, facilitating the development of sophisticated multi-modal applications that are both accessible and sustainable.[202]

### Intel AI Research Unveils FastDraft: A Cost-Effective framework for Pre-Training Draft Models with LLMs

Intel AI Research has introduced FastDraft, a cost-effective framework designed for pre-training and aligning draft models with various large language models (LLMs) for speculative decoding. This innovative approach addresses the challenges of inference speed and resource demands associated with LLMs, particularly in time-sensitive applications. FastDraft employs a structured method for pre-training on extensive datasets and fine-tuning through sequence-level knowledge distillation, ensuring compatibility with models like Phi-3-mini and Llama-3.1-8B. The framework achieves a high acceptance rate for draft models while requiring minimal training time, making it a significant advancement in enhancing the efficiency of LLMs in real-world scenarios.[203]

### Hymba Hybrid-Head: A Leap Forward in Small Language Model Technology

NVIDIA has introduced the Hymba Hybrid-Head Architecture, designed to enhance the performance of small language models (SLMs). This innovative architecture combines transformer attention mechanisms with state space models (SSMs), allowing for improved efficiency and memory recall. The Hymba-1.5B-Base model has shown remarkable results, outperforming all publicly available models under 2 billion parameters, including Llama-3.2-3B, with a 1.32% higher accuracy, an 11.67-fold reduction in cache size, and a 3.49-fold increase in throughput. This advancement aims to make small language models more effective and accessible for various applications.[204]

### HEIE: Enhancing AI-Generated Image Evaluation with Explainability

HEIE introduces an innovative framework to improve the evaluation of AI-generated content (AIGC) images by addressing key challenges such as explainability and logical reasoning. The framework includes the CoT-Driven Explainable Trinity Evaluator, which uses heatmaps, scores, and explanation outputs to break down complex tasks into simpler subtasks, enhancing interpretability. Additionally, the Adaptive Hierarchical Implausibility Mapper combines low-level image features with high-level tokens from large language models (LLMs) to provide precise local-to-global heatmap predictions. A new dataset, Expl-AIGI-Eval, is also introduced to facilitate interpretable implausibility evaluation of AIGC images. Through extensive experiments, HEIE demonstrates state-of-the-art performance, highlighting its effectiveness in improving the evaluation of AI-generated images.[205]

### Introducing Uber's Prompt Engineering Toolkit: Streamlining AI Prompt Management

Uber has introduced a Prompt Engineering Toolkit to streamline the creation and management of prompts for large language models (LLMs). This toolkit centralizes prompt design, allowing users to construct, manage, and execute prompt templates efficiently. It supports Retrieval-Augmented Generation (RAG) and integrates runtime feature datasets to enhance prompts with context. The toolkit also includes features for version control, collaboration, and safety measures to ensure responsible AI usage. It aims to facilitate rapid iteration and experimentation, enabling users to explore LLM models, create and update prompts, and evaluate their performance both offline and in production.[206]

[201] https://arxiv.org/abs/2410.06634

[202] https://arxiv.org/pdf/2411.04996

[203] https://arxiv.org/html/2411.11055v1

[204] https://developer.nvidia.com/blog/hymba-hybrid-head-architecture-boosts-small-language-model-performance/?linkId=100000311399195

[205] https://arxiv.org/html/2411.17261v1

[206] https://www.uber.com/en-SE/blog/introducing-the-prompt-engineering-toolkit

### A Framework for Reliable LLMs: Probabilistic Consensus through Ensemble Methods

Probabilistic Consensus through Ensemble Validation: A Framework for LLM Reliability introduces a framework to enhance the reliability of large language models (LLMs) in critical domains such as healthcare, law, and finance. The authors propose using ensemble methods for content validation through model consensus, significantly improving precision in tasks requiring factual accuracy and causal consistency. The framework demonstrated an increase in precision from 73.1% to 93.9% with two models and up to 95.6% with three models. This approach offers a pathway to more reliable autonomous AI systems, despite current constraints related to multiple-choice format requirements and processing latency.[207]

### Introducing CATP-LLM: A Framework for Cost-Aware Tool Planning with Large Language Models

The Cost-Aware Tool Planning with LLMs (CATP-LLM) framework has been developed to address the overlooked issue of tool execution costs in large language models (LLMs) used for tool planning. This innovative framework enables LLMs to consider execution costs, such as time, when scheduling external tools to perform complex tasks. CATP-LLM incorporates a tool planning language that allows for the creation of non-sequential plans with multiple branches, promoting efficient concurrent tool execution and cost reduction. Additionally, it includes a cost-aware offline reinforcement learning algorithm to fine-tune the LLM, optimizing the balance between performance and cost. The introduction of OpenCATP, the first platform for evaluating cost-aware planning, has demonstrated that CATP-LLM significantly outperforms existing models, including GPT-4, by achieving higher plan performance and lower costs. The codes for CATP-LLM and OpenCATP will be made publicly available, marking a significant advancement in the practical application of LLMs for tool planning.[208]

### Google Introduces HALVA: An innovative strategy to Mitigate AI Hallucinations in MLMs

In a groundbreaking development, researchers have introduced a contrastive tuning strategy aimed at reducing hallucinations in multimodal large language models (MLLMs) while maintaining their overall performance. This innovative approach, detailed in the HALVA project, addresses the challenge of object hallucination—where models generate descriptions of non-existent objects—by employing a two-step process: generative data augmentation and contrastive tuning. By selectively altering ground-truth information to create hallucinated tokens, the model is trained to favour factual outputs over erroneous ones. This method not only enhances the accuracy of MLLMs like LLaVA-v1.5 but also ensures that their general vision-language capabilities remain intact, paving the way for more reliable applications in AI-driven language and vision tasks.[209]

### LLM Stinger: Advancing Jailbreak Attacks on Large Language Models with a Novel Reinforcement Learning Method

LLM Stinger utilizes a novel method involving Large Language Models (LLMs) to automatically generate adversarial suffixes for jailbreak attacks. This approach bypasses the need for complex prompt engineering or white-box access by employing a reinforcement learning (RL) loop to fine-tune an attacker LLM. The method significantly enhances the Attack Success Rate (ASR) on various models, including a +57.2% improvement on LLaMA2-7B-chat and a +50.3% increase on Claude 2, both known for their robust safety measures. Additionally, LLM Stinger achieved a 94.97% ASR on GPT-3.5 and 99.4% on Gemma-2B-it, demonstrating its effectiveness across both open and closed-source models. This automated approach efficiently discovers new suffixes that bypass existing defences, streamlining the process of crafting jailbreak attacks.[210]

### Exploiting Vulnerabilities: Directed Representation Optimization Jailbreak in LLMs.

The study introduces DROJ, a novel method for executing prompt-driven attacks on large language models (LLMs). It highlights the vulnerability of LLMs to adversarial jailbreak attacks, which manipulate prompts to bypass safety mechanisms and elicit harmful responses. DROJ optimizes these prompts at the embedding level, achieving a 100% success rate in keyword-based attacks on the LLaMA-2-7b-chat model. While effective, the model occasionally produces repetitive and non-informative outputs. To enhance response quality, the authors propose a helpfulness system prompt, emphasizing the need for improved defences against such attacks in LLMs.[211]

### Mitigating the Threat of Jailbreak Prompts: Strategies for AI Security

Preventing Jailbreak Prompts as Malicious Tools for Cybercriminals: A Cyber Defence Perspective" explores the threat posed by jailbreak prompts, which are crafted to bypass ethical safeguards in AI systems. These prompts can enable harmful content generation, content filter evasion, and sensitive information extraction. The authors analyse techniques like prompt injection and context manipulation, assess the impact

[207] https://arxiv.org/abs/2411.06535

[208] https://arxiv.org/html/2411.16313v1

[209] https://research.google/blog/halva-hallucination-attenuated-language-and-vision-assistant/

[210] https://arxiv.org/html/2411.08862v1

[211] https://arxiv.org/html/2411.09125v1

of successful jailbreaks, and propose strategies such as advanced prompt analysis, dynamic safety protocols, and continuous model fine-tuning to enhance AI resilience. The study also emphasizes the need for collaboration among AI researchers, cybersecurity experts, and policymakers to set standards for protecting AI system.[212]

### Introducing Immune: A New Framework for Enhancing MLLM Safety

Immune is a new defence framework designed to enhance the safety of Multimodal Large Language Models (MLLMs). It operates during the model's inference time and uses a safe reward model to defend against jailbreak attacks—carefully crafted image-prompt pairs that can trick the model into generating harmful content. The framework provides a rigorous mathematical characterization, offering provable guarantees against such attacks. Extensive evaluations on various jailbreak benchmarks demonstrate that Immune significantly improves model safety while preserving the original capabilities of the MLLMs. For example, in tests against text-based jailbreak attacks on the LLaVA-1.6 model, Immune reduced the attack success rate compared to the base model and the best existing defence strategy.[213]

### Enhanced Detection of Jailbreak Prompts in Large Language Models Using Pretrained Embeddings

The study "Improved Large Language Model Jailbreak Detection via Pretrained Embeddings" presents a novel approach to identifying jailbreak prompts in Large Language Models (LLMs). By combining text embeddings optimized for retrieval with traditional machine learning classification algorithms, the proposed method significantly enhances the detection of jailbreak attempts. This approach aims to prevent LLMs from generating harmful content, outperforming existing open-source security applications.[214]

### Adaptive Framework for Managing Untrusted LLMs to Enhance Security

Adaptive Deployment of Untrusted LLMs Reduces Distributed Threats presents a two-level deployment framework designed to manage untrusted Large Language Models (LLMs) that may attempt to circumvent safety protocols. This framework employs an adaptive macro-protocol to dynamically select among micro-protocols, involving a trusted model that monitors the untrusted one. In a code generation scenario, this approach significantly reduces the presence of backdoors in generated code by 80% compared to non-adaptive methods, demonstrating its effectiveness in enhancing security.[215]

### IIT Jodhpur Researchers Develop Framework to Mitigate Bias in AI Results

Researchers at IIT Jodhpur have created a framework to address bias in AI results by scoring datasets on fairness, privacy, and regulatory compliance. This innovative framework aims to produce "responsible datasets" by ensuring diverse data collection while safeguarding individual privacy. It employs an algorithm to generate an 'FPR' score, which evaluates datasets based on fairness (representation of different groups), privacy (protection against data leaks), and regulatory compliance (consent and data removal rights). Testing on 60 global datasets revealed significant issues with fairness and compliance, particularly in face-based biometric datasets. This initiative marks a crucial step towards mitigating ethical concerns in AI and promoting responsible AI practices.[216]

### HackSynth: An AI Agent for Autonomous Penetration Testing and Its Evaluation Framework

A new framework designed for autonomous penetration testing, HackSynth, can generate commands and process feedback iteratively, enabling it to conduct penetration tests without human intervention. HackSynth was evaluated using two new Capture the Flag (CTF)-based benchmark sets created from PicoCTF and OverTheWire platforms. Extensive experiments demonstrated that HackSynth, especially when utilizing the GPT-4o model, exceeded performance expectations. This study underscores the potential of AI in enhancing cybersecurity through autonomous penetration testing while highlighting the importance of robust safeguards to ensure safety and predictability.[217]

### SafeWorld: A Benchmark for Geo-Diverse Safety Alignment in AI

SafeWorld is a benchmark designed to evaluate Large Language Models' (LLMs) ability to generate responses that are helpful, culturally sensitive, and legally compliant across diverse global contexts. SafeWorld includes 2,342 test queries based on cultural norms and legal policies from 50 countries and 493 regions/races. The proposed multi-dimensional safety evaluation framework assesses the contextual appropriateness, accuracy, and comprehensiveness of responses. Current LLMs struggle to meet these geo-diverse safety standards. To improve alignment, Direct Preference Optimization (DPO) alignment training with synthesized helpful preference pairs is used. The newly trained SafeWorldLM outperforms other models, including GPT-4, in all

[212] https://arxiv.org/pdf/2411.16642

[213] https://arxiv.org/html/2411.18688v1

[214] https://arxiv.org/html/2412.01547v1

[215] https://arxiv.org/html/2411.17693v1

[216] https://indiaai.gov.in/news/iit-jodhpur-researchers-create-framework-to-mitigate-bias-in-ai-results?utm_source=newsletter&utm_medium=email&utm_campaign=The%20Heuristic%20from%20INDIAai

[217] https://arxiv.org/html/2412.01778v1

evaluation dimensions, with a nearly 20% higher winning rate in helpfulness and harmfulness evaluations according to global human evaluators.[218]

### Targeted Model Editing: A Novel Approach to Bypassing Safety Filters in Large Language Models

Targeted Model Editing (TME) is a groundbreaking technique designed to bypass safety filters in Large Language Models (LLMs) by making minimal alterations to the internal structures of the models, rather than modifying inputs. This method identifies and removes safety-critical transformations (SCTs) within the model's matrices, allowing malicious queries to evade restrictions without changing the input. By analysing distinct activation patterns between safe and unsafe queries, TME isolates and approximates SCTs through an optimization process. Implemented in the D-LLM framework, TME achieves an average Attack Success Rate (ASR) of 84.86% on four mainstream open-source LLMs while maintaining high performance. Unlike existing methods, D-LLM does not require specific triggers or harmful response collections, making it a stealthier and more effective jailbreak strategy. This research highlights a significant and covert threat to LLM security, emphasizing the need for stronger safeguards to ensure model safety alignment.[219]

### Moving Target Defence: Innovative Strategy Enhances Security of AI Language Models Against Jailbreak Attacks

Moving target defence is a groundbreaking mechanism introduced by researchers to protect large language models (LLMs) from jailbreak attacks, which exploit manipulated prompts to generate harmful content. Traditional defence methods often require access to the model's internal structure or additional training, making them impractical for many service providers using LLM APIs like OpenAI and Claude. This new approach dynamically alters decoding hyperparameters to enhance model robustness without needing internal access or extra training. The defence includes optimizing decoding strategies by adjusting hyperparameters that influence token generation probabilities and transforming these hyperparameters and system prompts into dynamic targets that are continuously modified during runtime. This continuous modification effectively mitigates existing attacks. The defence was tested across four different attacks and five LLMs, showing superior performance in three models compared to six other defence methods. Additionally, it offers lower inference costs and maintains high response quality, making it a viable protective layer when combined with other defence strategies.[220]

[218] https://arxiv.org/html/2412.06483v1

[219] https://arxiv.org/pdf/2412.08201

[220] https://arxiv.org/pdf/2412.07672

## Industry Update

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

## Healthcare

### Cancer Centers Form AI Alliance to Accelerate Research

Four leading cancer centers—Dana-Farber, Fred Hutch, Memorial Sloan Kettering, and Johns Hopkins—have launched the Cancer AI Alliance (CAIA) with support from tech giants like AWS, Deloitte, Microsoft, and NVIDIA. This initiative aims to harness the power of AI to unlock new cancer research discoveries and improve patient care. By pooling their data and resources, these institutions hope to overcome technical challenges and accelerate breakthroughs in cancer treatment. The alliance emphasizes responsible AI use, data security, and collaboration to drive innovation and enhance health outcomes.[221]

### Ataraxis AI Launches Revolutionary AI-Native Cancer Diagnostic

Ataraxis AI has launched with the goal of transforming precision medicine through the introduction of the world's first AI-native cancer diagnostic, **Ataraxis Breast**, which has been clinically validated to be 30% more accurate than current standards for breast cancer detection. The company, co-founded by Dr. Jan Witowski and Dr. Krzysztof Geras, secured $4 million in seed funding to develop advanced AI-powered tools that enhance patient outcome predictions and enable personalized treatment plans. By addressing the limitations of traditional molecular diagnostics, Ataraxis aims to improve the accuracy and efficiency of cancer care, with plans to create tests for a massive portion of the twenty-six million new cancer cases expected globally by 2030, benefiting patients with tailored treatment options.[222]

### ALIGNMT AI and HFMA Collaborate to Enhance AI Governance in Healthcare

ALIGNMT AI has announced a strategic collaboration with the Healthcare Financial Management Association (HFMA) to launch a specialized AI Governance Micro-Credentialing program aimed at empowering healthcare professionals. Set to launch on November 18, 2024, this initiative addresses the urgent need for robust AI governance in the healthcare sector, equipping leaders with essential skills in compliance, risk management, and ethical practices. As AI continues to transform healthcare operations, this program will provide accessible training that helps professionals navigate the complexities of AI implementation responsibly. By fostering a culture of ethical AI use, ALIGNMT AI and HFMA aim to enhance trust in AI systems, improving patient care and organizational performance in a rapidly evolving landscape.[223]

### DeepMind Open-Sources AlphaFold 3 for Drug Discovery

Google DeepMind has made a groundbreaking move by open-sourcing AlphaFold 3, a significant advancement in AI-driven molecular biology and drug discovery. This release allows researchers worldwide to access the model's source code, enhancing their ability to predict complex interactions between proteins, DNA, RNA, and small molecules—key processes in understanding diseases and developing new treatments. Following the recent Nobel Prize awarded to its creators, AlphaFold 3 promises to transform traditional research methods, which often require extensive time and funding, into a more efficient and accessible approach. While the source code is freely available, access to the model weights is restricted to academic

[221] https://www.fredhutch.org/en/news/releases/2024/10/-cancer-centers-launch-cancer-ai-alliance-to-unlock-discoveries-.html#:~:text=The%20alliance%20will%20apply%20responsible%20AI

[222] https://www.biospace.com/press-releases/ataraxis-ai-launches-to-transform-precision-medicine-beginning-with-worlds-first-ai-native-cancer-diagnostic

[223] https://finance.yahoo.com/news/alignmt-ai-hfma-collaborate-offer-050000131.html

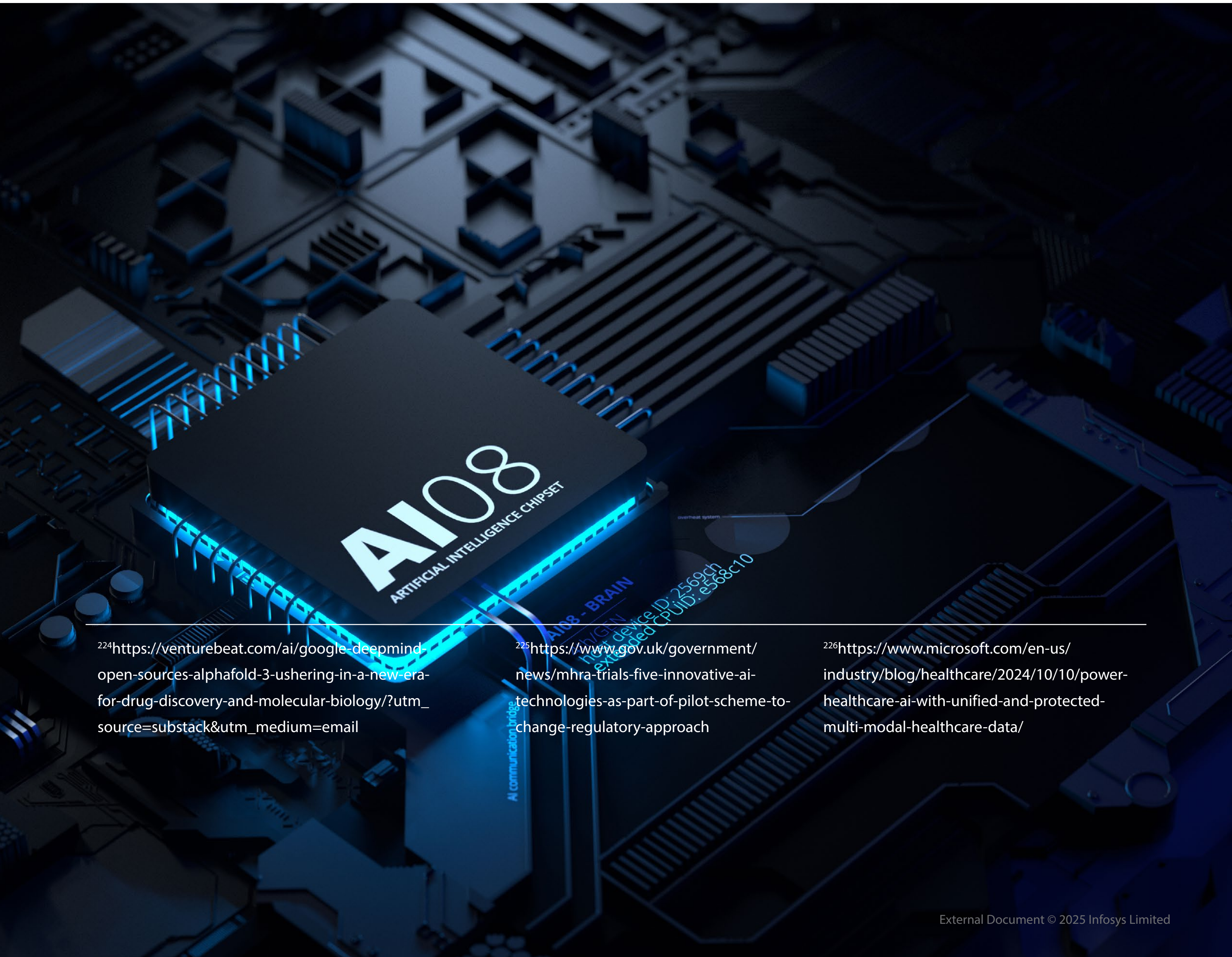
use, balancing the needs of open science with commercial interests.[224]

### MHRA Launches AI Airlock Pilot Scheme to Revolutionize Regulation of AI-Powered Medical Devices

The Medicines and Healthcare products Regulatory Agency (MHRA), an executive agency of the Department of Health and Social Care in the United Kingdom, has initiated the AI Airlock pilot scheme to trial five innovative AI technologies, aiming to transform the regulation of AI-powered medical devices. This initiative ensures that these devices, which include tools for cancer and chronic respiratory disease patients as well as radiology diagnostic services, reach patients quickly and safely. The AI Airlock serves as a regulatory "sandbox," allowing manufacturers to determine the best methods for collecting evidence required for product approval under MHRA supervision. This approach seeks to create a more enabling regulatory framework, providing a clearer route to market and faster access to potentially transformative AI technologies for the NHS and patients.[225]

### Transforming Healthcare with AI and Data: Microsoft's Latest Innovations

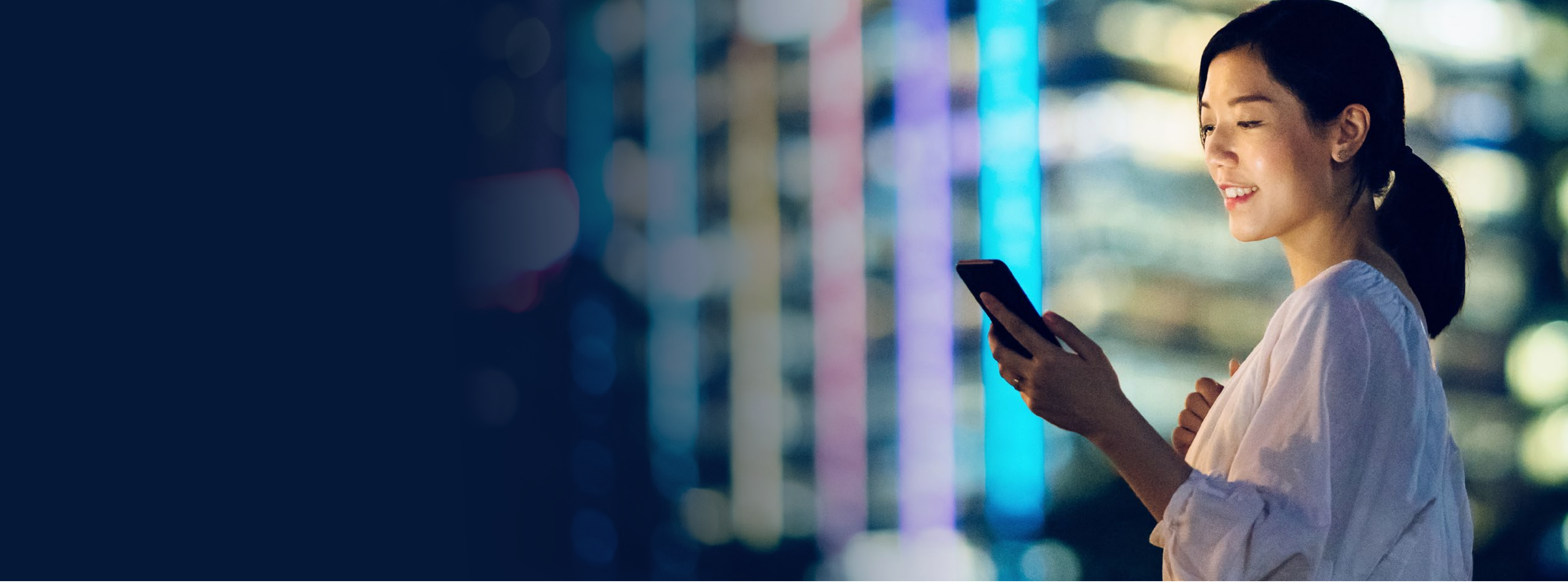
Microsoft's new innovations in healthcare through the Microsoft Cloud for Healthcare, emphasizing the integration of advanced AI and data solutions. Key developments include AI capabilities to enhance patient care and personalize treatments, the availability of healthcare data solutions in Microsoft Fabric, and the public preview of healthcare application templates in Microsoft Purview. Additionally, the Dragon Ambient Experience (DAX) generates draft medical notes at the point of care, combining healthcare and conversational data for unique insights. These innovations aim to improve patient outcomes, reduce healthcare costs, and support compliance with regulations like HIPAA and GDPR.[226]



[224] https://venturebeat.com/ai/google-deepmind-open-sources-alphafold-3-ushering-in-a-new-era-for-drug-discovery-and-molecular-biology/?utm_source=substack&utm_medium=email

[225] https://www.gov.uk/government/news/mhra-trials-five-innovative-ai-technologies-as-part-of-pilot-scheme-to-change-regulatory-approach

[226] https://www.microsoft.com/en-us/industry/blog/healthcare/2024/10/10/power-healthcare-ai-with-unified-and-protected-multi-modal-healthcare-data/



## Telecommunication

### GSMA and 19 MNOs Launch AI Safety Roadmap

The GSMA, a global association representing mobile operators, has introduced a Responsible AI (RAI) Maturity Roadmap to guide telecom companies in adopting AI ethically and safely. This initiative, backed by 19 major MNOs, aims to help operators navigate the potential benefits of AI while mitigating risks. The roadmap provides guidelines on various aspects of AI implementation, including vision, operating model, technical controls, third-party ecosystems, and change management. By following this roadmap, telecom companies can assess their current AI maturity and identify areas for improvement to ensure responsible and ethical AI usage.[227]

## Banking and Insurance

### National Bank of Georgia Moves Towards EU-Standard AI Regulations

On November 7, 2024, the National Bank of Georgia announced plans to develop regulations for artificial intelligence (AI) that align with European Union standards. This initiative aims to enhance the stability of the financial system and improve risk management practices within the sector. The regulations will be part of a broader national fintech strategy, which includes leveraging AI technologies to meet regulatory objectives effectively. Key components of this strategy also focus on reducing market entry barriers for inexperienced players, modernizing financial infrastructure, and positioning Georgia as a regional fintech hub. By aligning with EU standards, the National Bank seeks to foster innovation while ensuring robust oversight in the rapidly evolving AI landscape.[228]

### Building Trust in Banking with Responsible AI: The Inspeq.ai Approach

Inspeq.ai is revolutionizing the banking industry with its responsible and trustworthy AI solutions. By integrating advanced AI technologies, Inspeq.ai enhances security, efficiency, and customer confidence in banking services. Their platform focuses on real-time inspection, monitoring, and governance of AI systems, ensuring compliance with regulatory standards and reducing risks associated with AI deployment. This approach not only improves operational accuracy but also fosters trust and transparency, making AI a reliable tool for modern banking.[229]

### RBI Establishes Panel for Ethical AI in Financial Services

The Reserve Bank of India (RBI) has announced the formation of a committee to develop a framework for the responsible and ethical use of artificial intelligence (AI) in the financial sector. This initiative aims to address critical challenges such as algorithmic bias, decision explainability, and data privacy. By including specialists from various disciplines, the committee will create a robust and adaptable framework tailored to the financial industry’s needs, ensuring that AI technologies are deployed in a manner that is both ethical and sustainable.[230]

### RBI Introduces MuleHunter.AI to Combat Financial Fraud

The Reserve Bank of India (RBI) has launched MuleHunter.AI, an advanced artificial intelligence tool developed by the Reserve Bank Innovation Hub (RBIH) to tackle financial fraud. This innovative tool is designed to identify and flag mule accounts, which are often used for money laundering. Mule accounts

[227] https://developingtelecoms.com/telecom-business/operator-news/17325-gsma-19-mnos-launch-ai-safety-roadmap.html

[228] https://agenda.ge/en/news/2024/41473?utm_source=substack&utm_medium=email#gsc.tab=0

[229] https://inspeq.ai/banking-services-responsible-trustworthy-ai

[230] https://inc42.com/buzz/rbi-to-set-up-panel-for-ethical-use-of-ai-in-financial-services/

are bank accounts exploited by criminals to transfer illicit funds, typically created by individuals who are either deceived with promises of easy money or coerced into participation. MuleHunter.AI employs machine learning algorithms to analyze transaction data and account details, predicting mule accounts with greater accuracy and speed than traditional methods. Successfully piloted in two public sector banks, this AI tool has demonstrated significant improvements in detection efficiency, underscoring the critical role of technology in enhancing the security and resilience of the financial ecosystem.[231]

## Transportation Safety

### AI Cameras in Odisha Successfully Prevent Train-Elephant Collisions

In Odisha, AI-powered cameras have been installed to prevent train collisions with elephants, a frequent hazard in the region's forests. Recently, these cameras successfully alerted railway authorities to halt a train when a herd of elephants was detected near the tracks. This pilot initiative, implemented in the Rourkela Forest Division and funded by the Rourkela Steel Plant, is part of a broader wildlife conservation effort. The AI system captures and zooms in on the elephants, sending real-time alerts to the control room, which then instructs the train to stop. Plans are underway to expand this technology to other forest divisions, including Keonjhar and Bonai, to enhance wildlife protection and ensure safer railway operations.[232]

## Defense

### Chinese Researchers Develop Military AI Model Based on Meta's Llama Technology

Chinese researchers have developed a military-focused AI model called ChatBIT, based on Meta's Llama model. The Llama 2 13B large language model was used to create a tool designed for intelligence gathering and operational decision-making. ChatBIT has been optimized for military dialogue and question-answering tasks, outperforming some models comparable to OpenAI's ChatGPT-4. Despite Meta's restrictions on military use of its models, the open-source nature of Llama allows for such adaptations, raising concerns about the implications of AI in military contexts.[233]

### Anthropic and Palantir Collaborate to Deploy Claude AI for Enhanced U.S. Government Intelligence Operations

Anthropic has partnered with Palantir and Amazon Web Services (AWS) to integrate its Claude AI models into U.S. government intelligence and defense operations. This collaboration aims to enhance data processing capabilities for intelligence agencies, allowing them to analyze large volumes of complex data quickly and identify patterns effectively. The Claude models will be hosted in Palantir's secure environment, which meets stringent Defense Information Systems Agency (DISA) standards. This partnership is seen as a significant step in applying AI responsibly within classified settings, although it has raised concerns about potential conflicts with Anthropic's commitment to ethical AI development.[234]

## Aviation

### UK Civil Aviation Authority Unveils Comprehensive AI Strategy to Revolutionize Aviation

The UK Civil Aviation Authority (CAA) has introduced a new AI strategy designed to transform the aviation industry by promoting the adoption of AI within the aerospace sector and enhancing the CAA's own operations. This strategy emphasizes trust and safety, ensuring AI applications are dependable for passengers and pilots. It includes developing regulatory guidelines to balance innovation with safety, improving operational efficiency through AI in air traffic control and airport management, implementing AI-driven simulations for pilot training, and optimizing flight routes to reduce fuel consumption. The CAA aims to harness AI's potential while maintaining high safety standards and public trust.[235]

## Automobile

### CNIL Releases Guidelines on AI-Enhanced Cameras in Commercial Vehicles

The French data protection authority, CNIL, has published guidelines on the deployment of AI-enhanced cameras in commercial transport vehicles. These advanced cameras are designed to boost driver safety by detecting signs of fatigue, distraction, and other risky behaviors in real-time. The guidelines emphasize the necessity for these devices to comply with data protection laws and safeguard drivers' privacy. While the technology aims to improve road safety and driver training, it must be balanced with the protection of personal data. Employers are required to justify the use of these cameras and ensure their proportional implementation to prevent any

[231] https://www.rbi.org.in/Scripts/BS_ViewBulletin.aspx?Id=22995

[232] https://www.ndtv.com/travel/ai-cameras-in-odisha-alert-railways-to-halt-train-after-spotting-elephant-herd-7213102

[233] https://www.deccanherald.com/world/chinese-researchers-develop-ai-model-for-military-use-on-back-of-metas-llama-3258357

[234] https://www.businesswire.com/news/home/20241107699415/en/Anthropic-and-Palantir-Partner-to-Bring-Claude-AI-Models-to-AWS-for-U.S.-Government-Intelligence-and-Defense-Operations

[235] https://www.airport-technology.com/news/uk-caa-ai-strategy-aviation/?utm_source=substack&utm_medium=email

[236] https://www.cnil.fr/fr/les-cameras-augmentees-dans-les-habitacles-des-vehicules-de-transport-de-marchandises?utm_source=substack&utm_medium=email

infringement on employees' privacy rights.[236]

### Neuro-Explicit AI: Paving the Way for Safer Autonomous Driving

Accenture and the German Research Center for Artificial Intelligence (DFKI) have released a joint white paper titled "Responsible AI in the Automotive Industry – Techniques and Use Cases." This paper addresses the critical need for trustworthy and responsible AI in autonomous driving, highlighting the limitations of current deep learning models in terms of explainability, robustness, and generalizability. The authors propose a hybrid approach called neuro-explicit AI, which combines neural networks with symbolic reasoning to enhance transparency and reliability. This initiative is crucial for improving the safety and performance of autonomous vehicles, ensuring ethical AI practices, and fostering public trust. The white paper's findings and recommendations will be particularly relevant as the automotive industry continues to evolve and integrate advanced AI technologies.[237]

## Agriculture

### Empowering Farmers with Kissan.ai

Kissan.ai is an innovative multilingual AI-powered agricultural assistant designed to support farmers in navigating the complexities of modern farming. By providing real-time, personalized advice on a wide range of agricultural topics, Kissan.ai helps farmers maximize their productivity and make informed decisions. Its unique bilingual capabilities bridge language barriers, delivering expert guidance in both English and Hindi, which is particularly beneficial for India's diverse farming community. This accessibility not only enhances the user experience but also empowers farmers with the knowledge they need to tackle challenges effectively, contributing to more sustainable agricultural practices.[238]

[237] https://idw-online.de/en/news841988

[238] https://kissan.ai/

## Infosys Developments

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

## Events

### Responsible AI Legal Conference | London

On September 26th in London, Infosys, in partnership with the Association of Corporate Counsel (ACC), hosted a workshop tailored for Chief Legal Officers (CLOs) and senior legal professionals. The event aimed to address the evolving legal landscape surrounding artificial intelligence (AI), focusing on compliance, contractual relationships, and privacy. Distinguished speakers, including Inderpreet Sawhney, General Counsel & Chief Compliance Officer at Infosys, Faiz Rahman, Vice President & IP Head at Infosys, and Srinivas Poosarla, Chief Data Privacy Officer at Infosys, provided valuable insights into the complexities of AI contracts, the evolving regulatory landscape, and its potential impact on AI adoption. Through engaging fireside chats and panel discussions, attendees gained a deeper understanding of the EU AI Act, the UK's global policy approach to AI, and best practices for navigating these challenges in the legal sphere.

### A Night of AI Excellence: Networking Dinner | New Delhi

On 27th September, The British High Commissioner to India, Lindy Cameron, hosted a networking dinner at her residence in New Delhi to celebrate the collaboration between the UK Centre for Responsible AI (Responsible AI UK) and the Centre for Responsible AI (CERAI) at IIT Madras. The event was attended by prominent AI leaders from around the world, including Wendy Hall (Celebrated Scientist & member of the United Nations high-level advisory body on AI), Sarvapali (Gopal) Ramchurn (CEO of Responsible AI UK), and Balaraman Ravindran (Renowned Professor at IIT Madras). Ashish Tewari – Head of Infosys Responsible AI Office, India attended this event from the Infosys Responsible AI Office. This gathering strengthened Infosys' relationship with the British Consulate and opened doors for future partnerships. Infosys looks forward to contributing to the field of responsible AI alongside organizations such as Responsible AI UK and IIT Madras.

### Infosys and Cisco Mark 25 Years of Strategic Collaboration | Bengaluru

Infosys and Cisco hosted a joint celebration to mark their 25-year partnership. This event underscored the successful collaboration between the two companies and their shared commitment to innovation. A significant highlight of the event was the presence of the Infosys Responsible AI (RAI) office, which highlighted its advanced AI solutions, including the AI3S platform. This comprehensive AI solution exemplifies Infosys' dedication to delivering responsible and ethical AI technologies. The event featured discussions on several topics, such as AI training, the EU AI Act, and the evolving regulatory landscape, through fireside chats and panel discussions.

### Infosys Presents AI Guidance at ISO Meeting | Versailles, France

The ISO 14th plenary meeting is currently underway in the historic city of Versailles, France, from October 7-11, 2024. Among the key highlights of the event is a presentation by Mr. Syed Ahmed, Head of Infosys Responsible AI office. He is presenting India's study item titled "Artificial Intelligence – Guidance for Generative AI Applications" to the ISO committee. This presentation is providing attendees with macro level guidance on measuring the quality of output of generative AI applications, highlighting India's initiative-taking role in shaping global AI standards.

### Infosys Responsible AI Office participates at National AI Conference | 17th and 18th October | India

Ashish Tewari – Head of Infosys Responsible AI Office, India participated in the National AI Conference held at the National Forensic Sciences University (NFSU) Campus on October 17th and 18th. Organized by the Ministry of Home Affairs, GOI, NFSU, and National Law University, Delhi, the conference brought together top minds to address the challenges and opportunities presented by artificial intelligence and its responsible use.

During the conference, Ashish delivered a speech on the regulatory landscape of AI. His presentation focused on the growing concern of AI crimes, the existing gaps in criminal and IT laws, and the necessary steps to ensure responsible AI development. Ashish also emphasized the importance of collaboration among academia, law enforcement, and industry to navigate the complexities of AI and harness its benefits while mitigating risks. He further interacted with the Ms. Sivagami Sundari Nanda, Special Secretary (Ministry of Home Affairs), Mr. Abhishek Singh, Additional secretary (MeitY), law enforcement agencies, and academic institutions, highlighting Infosys' commitment to responsible AI.

### Core-AI Coalition Engages with MeitY to Advance Responsible AI Development in India | 21st Oct | India.

Members of Core-AI (Multi-stakeholder Coalition) met with the Ministry of Electronics and Information Technology (MeitY) Secretary Shri S. Krishnan on October 21st in New Delhi to discuss responsible AI development in  context of India. At the meeting, representatives from Core-AI discussed the coalition's objectives and its role in promoting responsible AI development through collaborations with academia, industry, and startups. Ashish Tewari – Head of Infosys Responsible AI Office, India, highlighted the capabilities of Infosys' Responsible AI Office and its potential contributions to building a robust responsible AI framework for India. He emphasized the office's alliance with organizations like the US Safety Institute (NIST) and the AI Alliance, as well as its experience providing guidance on AI policy making and responsible AI framework implementation to countries such as UK, Australia and Switzerland.

### Infosys and Open University Led AI Education Conference |30th October |Milton Keynes, UK

On October 30, 2024, Infosys partnered with Open University (OU) to organize the Gen AI: Adult Learning and Skill Conference, bringing together educators, policymakers, and technologists to explore AI's potential in education and address skill shortages through upskilling. Celebrating the first anniversary of the 2023 AI Safety Summit, which featured leaders like Kamala Harris and Elon Musk at Bletchley Park, the conference was part of a broader AI festival led by Milton Keynes City Council. Open University, dedicated to transforming lives and helping individuals realize their ambitions, focuses on using trustworthy AI and technological innovations to deliver top tier learning programs, aligning closely with Infosys' principles of ethical AI use and compliance. Keynote sessions by OU Professors John, Ian, and Esther Spring, along with various panel discussions, highlighted the shared commitment to leveraging AI for positive societal impact and economic growth.

### Future of AI Summit 2024 | 6th-7th November | London, UK

On November 6th-7th, 2024, the Future of AI Summit took place in London, providing a comprehensive overview of AI innovation and examining real-world applications. The event brought together leaders from various sectors responsible for creating, integrating, scaling, and commercializing AI, while addressing security, workforce, and ethical concerns. Keynote speakers included Bali from Infosys, Zoubin Ghahramani from Google DeepMind, Peter Kyle from the UK Government, and Josephine Teo from the Government of Singapore, who shared insights on AI's direction, innovation, adoption, and challenges. Additionally, discussions led by Mona from Infosys EU, Andrea Abell from Eli Lilly, and Bernhard Maier from Signature Litigation focused on reshaping cybersecurity and data governance for AI deployment. The Infosys Responsible AI office presented the Responsible AI Toolkit demo. The summit highlighted the shared commitment to leveraging AI for positive societal impact and economic growth.

### Infosys AI Day Event | 7th November | Phoenix Infosys Innovation Hub, Arizona State University

On November 7th, the AI Day event at the Phoenix Infosys Innovation Hub was a successful event and it featured the latest AI showcases on Agentic AI, Responsible AI, AI First Operations, ICETS innovations, and Data Foundations. Ashiss Dash, Head of Services, Utilities, Resources, and Energy at Infosys, kicked off the day with an insightful keynote on the rapid transformation of AI and its impact on the enterprises. Vivek Sinha, Global Head of AI, impressed attendees with a live demo of iTASK (Infosys Topaz Agentic Solutions Kit), showcasing its capabilities in automating software development processes. The event included valuable panel discussions and interviews with customers sharing their enterprise AI journeys, with great insights from partners UiPath, AWS, and Aisera. The Infosys Responsible AI office showcased the demo of the Responsible AI Toolkit . The day concluded with an inspiring talk by Dr Pawan Sinha on human brain development and his initiative Prakash, which aims to help underprivileged children who are blind at birth.

### Infosys EMEA Confluence 2024 | November 13-14, 2024| Venice, Italy

On 13th-14th November, Infosys EMEA Confluence 2024 took place in the enchanting city of Venice. This exceptional event gathered a diverse group of customers, partners, analysts, and advisors, making the annual gathering truly memorable. The discussions focused on innovation, particularly AI and Responsible AI. Topics ranged from enhancing AI's role in business to examining the ethical considerations crucial for its responsible deployment, enriching the collective vision with profound insights. Amidst the intense discussions, there were moments of connection, laughter, and genuine partnership celebration, all set against Venice's captivating backdrop. Balancing challenging work with these shared experiences added depth to the interactions, fostering a collaborative environment for transformative journeys ahead. Balakrishna DR (Executive Vice President at Infosys) expressed his enthusiasm for the Infosys EMEA Confluence on Partner Day, highlighting the engaging discussions on AI deployment complexities and the critical role of partners. He emphasized the importance of ethical considerations, technical challenges, and aligning AI initiatives with business values for responsible adoption.

### Infosys Responsible AI Session | November 19, 2024| Bucharest, Romania

Syed Ahmed, Head of the Infosys Responsible AI Office recently conducted a session on Responsible AI at the Infosys Romania office in Bucharest. The event brought together Infosys leaders from Romania in person and participants from Bulgaria via Teams, facilitating an in-depth and engaging discussion on several pivotal topics. These included the functions of the Infosys Responsible AI Office, the security and safeguarding of AI systems, legal frameworks and compliance for AI, the implications of the EU's AI Act, and Infosys' forward-looking vision for AI in this transformative domain. The session was characterized by vibrant energy, collaborative spirit, and the exchange of innovative ideas, all of which were truly inspiring. This event not only propelled the conversation around Responsible AI but also offered an excellent opportunity to connect with numerous enthusiastic individuals who share a passion for AI.

### RICON 2024 NASSCOM Event | November 21 | New Delhi

On November 21, the inaugural Confluence for Responsible Intelligence (RICON) 2024 took place in New Delhi, bringing together a diverse group of industry leaders, investors, and AI experts. Ashish Tiwari, Principal Consultant at Infosys Responsible AI Office, India, participated as a panelist alongside other esteemed speaker. The panel discussed assorted topics, including "Investing in Tomorrow - Aligning Capital with Responsible AI," emphasizing the importance of mandating responsible AI compliance for startups seeking investment. Speakers such as Ashwin Raguraman from Bharat Innovation Fund, Mudit Narain from Blume Ventures, and Arya Tripathy from Cyril Amarchand Mangaldas shared profound insights on fostering responsible AI development. A highlight of the event was the launch of the Developer's Playbook for Responsible AI in India, a voluntary framework aimed at guiding developers in mitigating AI risks. The event balanced intense discussions with moments of connection and collaboration, setting the stage for a transformative journey in responsible AI practices. Special thanks were extended to the NASSCOM AI team for their exceptional organization and to the Infosys Responsible AI Office for sharing industry perspectives.

nasscom
CONFLUENCE for RESPONSIBLE INTELLIGENCE
RICON 2024
Fireside Chat
nasscom
CONFLUENCE for RESPONSIBLE INTELLIGENCE
RICON 2024

### AI Alliance Event | November 22 | Bangalore

On November 22, 2024, Srinivasan Sivasubramanian from the Infosys Responsible AI Office participated in the inaugural AI Alliance Technical Meeting at the IBM Research Lab in Bengaluru. The event aimed to foster collaboration, promote open sourcing, and encourage joint innovation among alliance members. Infosys showcased its AI First journey, emphasizing its commitment to open sourcing and the offerings from Infosys Topaz. Srini led a discussion on Responsible AI, engaging representatives from IBM, LTI Mindtree, Red Hat, and IIT Jodhpur. The focus was on the ethical implications of AI technologies and the shared responsibility of industry leaders to ensure beneficial outcomes for society. This meeting served as a vital platform for exchanging insights and strengthening the alliance's mission to advance responsible AI practices.

### Infosys Responsible AI Office and future of AI and Infosys vision Event | November 21 | Warsaw, Poland

On November 21, 2024, Syed Quiser Ahmed, Head of the Infosys Responsible AI Office, led an inspiring session on Responsible AI at the Infosys Poland office in Wroclaw. The workshop, which focused on the future of AI and Infosys' vision, was well-received, featuring rich discussions and insightful questions. Syed emphasized the significant opportunity for Infosys to leverage the local talent pool in advancing its AI initiatives. He extended special thanks to Lilly Vasanthini, Vice President-Delivery Head for Eastern Europe, NORDIC & Switzerland at Infosys, Kordian K., and Girish Babu, Principal Consultant at Infosys, for their invaluable contributions. The session underscored the importance of responsible AI and highlighted the global Infosys family's commitment to shaping the future of AI with integrity and innovation. The event was a resounding success, reflecting the collective effort and dedication of all involved.

### Infosys Topaz AI Conversation | Dubai

On November 24, 2024, Infosys, in collaboration with Amazon Web Services (AWS), hosted the illustrious Topaz AI Day at the opulent Atlantis the Palm in Dubai. The event featured a distinguished panel on Responsible AI, moderated by the insightful Sara Almukhalid, with esteemed panellists Rafee Tarafdar, CTO of Infosys, Syed Quiser Ahmed, Head of the Infosys Responsible AI Office, and Dr Mustapha Tawbi, AWS Partner Success Solutions Architect, drew an impressive assembly of regional customers eager to explore cutting-edge advancements in AI. At the event, the panel discussed the central themes of specialized language models and Responsible AI, emphasizing their vital roles in fostering trust, ensuring robust security, and championing ethical innovation. Syed highlighted the capabilities of Infosys' Responsible AI Office and its potential contributions to building a robust responsible AI framework. The event not only showcased the transformative potential of AI but also reinforced a steadfast commitment to responsible and ethical innovation.

### Infosys Topaz Responsible AI Legal Summit | November 18 | New York

On Nov 18th, Infosys Topaz conducted the Responsible AI legal summit in New York. The summit brought together corporate leaders with a special focus on managing the legal complexities around AI. The Summit was designed particularly for general counsels and legal professionals, to create collective understanding around how lawyers retained by corporates can effectively navigate emerging compliance and contractual conversations around AI.

Over 80+ customer and partner counsels came together to discuss on AI risks and challenges, the sessions highlighted a shared commitment. There was a clear consensus that strong principles are key to addressing most AI risks.

### Infosys Topaz Responsible AI Legal Workshop| November 20th-21st | New York and Dallas

On November 20th and 21st, 2024, The Infosys Topaz hosted a Responsible AI Legal Workshop in New York and Richardson, Dallas. Inderpreet Sawhney, General Counsel & Chief Compliance Officer at Infosys, highlighted the opportunities and challenges of adopting Responsible AI and navigating the evolving regulatory framework. Faiz Rahman, Vice President & IP Head at Infosys, noted that AI is revolutionizing business and the legal world, raising critical questions around compliance and contracting. Mandanna Appanderanda N from Infosys Responsible AI office presented on AI risks and mitigations to senior leaders, emphasizing the organization's initiative-taking approach to fostering trust and driving sustainable progress in AI. This workshop highlighted Infosys' dedication to ensuring that AI advancements are safe and beneficial, reinforcing the company's leadership in the field of responsible AI.

### UK-India Workshop and Hackathon on Responsible AI | December 5-6 | Bangalore, India

On December 5th and 6th, 2024, the UK-India Centre for Responsible and Trustworthy AI hosted a two-day event at the Infosys campus in Bangalore. This event, titled "UK-India Cooperation Towards a Fair AI Horizon - Comprehensive Workshop, AI Bias Detection Hackathon & Fairness Assessment Challenge," aimed to advance the development and deployment of ethical and responsible artificial intelligence technologies was launched by Balakrishna DR (Executive Vice President at Infosys) and addressed by Joshua Bamford (British High Commission in India) and Syed Ahmed, Head of the Infosys Responsible AI Office. It was part of the ambitious bilateral cooperation agenda set out in the India-UK Roadmap 2030. The event was organized by Infosys Responsible AI office, Infosys Consulting and the British High Commission India, in collaboration with NASSCOM and The Dialogue. The event also included the presentation of the Infosys Responsible AI Toolkit, designed to address challenges posed by Generative AI, which received overwhelmingly positive feedback. The hackathon winners, Privasapien and FairMD, were recognized for their outstanding presentations. Overall, the event was a great success, leaving participants eager for future collaborations.

### CoRE-AI Delegation Meeting with Karnataka IT Department| December 9| Bangalore, Karnataka

On December 9, 2024, the Coalition for Responsible Evolution of AI (CoRE-AI) held a significant delegation meeting with the Karnataka Government's IT Secretary. Ashish Tewari – Head of Infosys Responsible AI Office, India led the session. He presented the Infosys Responsible AI toolkit, emphasizing its role in promoting ethical AI development and deployment. The meeting concluded with a mutual agreement to enhance collaboration between CoRE-AI and the Karnataka Government to foster a sustainable and ethical AI ecosystem in the region.

### Data Science Summit | November 29 | Hyderabad

On November 29, 2024, the Data Science Summit took place at the ISB Campus in Gachibowli, Hyderabad. A notable panel discussion titled "AI Safety and Regulations: Current Landscape, Challenges, and Way Forward" was a highlight of the event. Ashish Tewari – Head of Infosys Responsible AI Office, India, participated in the panel discussion, which was moderated by Dr. Avik Sarkar and featured distinguished panelists including Ankit Bose, Head of Nasscom AI; Diptikalyan Saha, Senior Technical Staff Member at IBM Research India; and Saurabh Singh, Leader of Digital & AI Policy India & South Asia at Amazon Web Services. The discussion focused on the rapidly evolving landscape of Artificial Intelligence, with an emphasis on AI ethics, Responsible AI, and emerging frameworks for AI safety and regulations. The panelists provided valuable insights into the challenges and opportunities associated with deploying AI models responsibly. They also offered practical strategies for navigating this dynamic and critical field.

AI Safety and Regulations:
Current Landscape, Challenges and Way Forward

### U.S. AI Safety Institute Consortium Meeting | December 3| University of Maryland, USA

On December 3, 2024, the U.S. AI Safety Institute Consortium (AISIC) held its inaugural in-person meeting at the University of Maryland, bringing together representatives from diverse companies, organizations, and local governments. The primary agenda was to review the consortium's progress and strategize on enhancing its role as a bridge between the technology industry, academia, civil society, and the U.S. government on critical AI safety issues. Mandanna Appanderanda N from Infosys presented Infosys' Responsible AI guardrails as a case study on Safe System for AI. This initiative supports federal efforts to advance AI safety and foster continued American innovation. AISIC has been advancing scientific inquiry and collaborative research across five key areas: generative AI risk management, synthetic content, evaluations, red-teaming, and model safety and security.

## Events Latest News:

### Infosys and University of Cambridge Join Forces to Launch Cutting-Edge AI Lab in London

Infosys has partnered with the University of Cambridge to establish a new AI lab in London. This collaboration aims to leverage the strengths of both entities to drive innovation and address real-world challenges through advanced AI research and applications. The importance of this initiative lies in its potential to accelerate the development of innovative AI technologies, which can significantly benefit various industries by enhancing efficiency, productivity, and decision-making processes. Additionally, the lab will serve as a hub for education and research, fostering collaboration between academia and industry, and providing valuable opportunities for students and professionals to engage in pioneering AI projects. This partnership is expected to contribute to the broader goal of harnessing AI for positive societal impact and economic growth.

### Advancing Responsible AI: Infosys Launches Small Language Models

On October 24, 2024, Infosys announced the launch of its new small language models, Infosys Topaz BankingSLM and Infosys Topaz ITOpsSLM, developed in collaboration with NVIDIA and Sarvam AI. These models leverage the powerful NVIDIA AI Stack and utilize both general and industry-specific data, fine-tuned with Infosys's proprietary data. They are integrated into existing offerings like Infosys Finacle and Infosys Topaz, creating robust foundational models tailored for banking and IT operations.

This initiative reflects Infosys's commitment to responsible AI practices, emphasizing ethical considerations in AI deployment. The company also offers these models as services, including pretraining-as-a-service and fine-tuning-as-a-service, enabling businesses to build custom AI models securely and in compliance with industry standards. By focusing on security, transparency, and the ethical use of AI, Infosys aims to foster innovation while addressing potential risks, marking a significant step forward in making AI more accessible and responsible for enterprises.[239]

[239] https://www.infosys.com/newsroom/press-releases/2024/launch-small-language-models.html

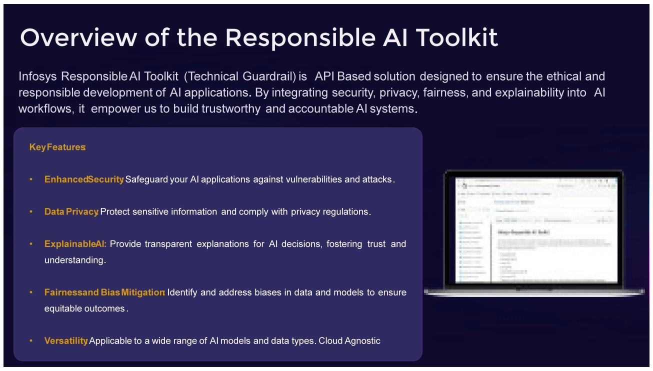
### Infosys Finacle Introduces Responsible AI Suite for Banks

Infosys Finacle has launched the Finacle Data and AI Suite, a comprehensive platform designed to help banks accelerate their AI journey by integrating AI into their digital operations. This suite is significant as it emphasizes Responsible AI, incorporating strong standards of AI ethics, trust, privacy, security, and regulatory compliance. By providing low-code, predictive, and generative AI solutions with high transparency and explainability, the suite aims to mitigate biases and safeguard data privacy. This will be particularly beneficial for banks looking to scale their data readiness, industrialize AI model development, and deliver actionable insights across their ecosystem, ensuring that AI applications are developed and deployed responsibly.[240]

## Infosys Responsible AI Toolkit – A Foundation for Ethical AI

### Exciting News: Open Sourcing the Responsible AI Toolkit!

We are thrilled to announce that our Responsible AI Toolkit will be open sourced in early next year! Following a successful soft launch on October 30th, where we highlighted the toolkit to key industry and academic leaders, the demo received widespread appreciation and sparked significant interest in evaluating its capabilities. To request exclusive access to explore our toolkit, or for more information, email the Responsible AI Office at responsibleai@infosys.com. Stay tuned for more updates as we work together to promote responsible AI practices!

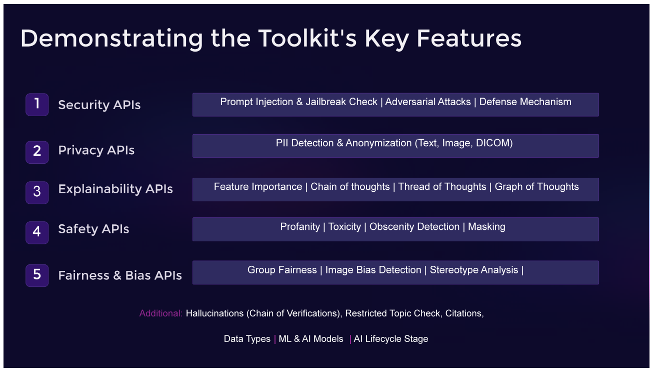


Overview of the Responsible AI Toolkit

Infosys ResponsibleAI Toolkit (Technical Guardrail) is 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 empower us to build trustworthy and accountable AI systems.

KeyFeatures

- EnhancedSecurity Safeguard your AI applications against vulnerabilities and attacks.
- Data Privacy Protect sensitive information and comply with privacy regulations.
- ExplainableAI: Provide transparent explanations for AI decisions, fostering trust and understanding.
- Fairnessand Bias Mitigation Identify and address biases in data and models to ensure equitable outcomes.
- Versatility Applicable to a wide range of AI models and data types. Cloud Agnostic



Demonstrating the Toolkit's Key Features

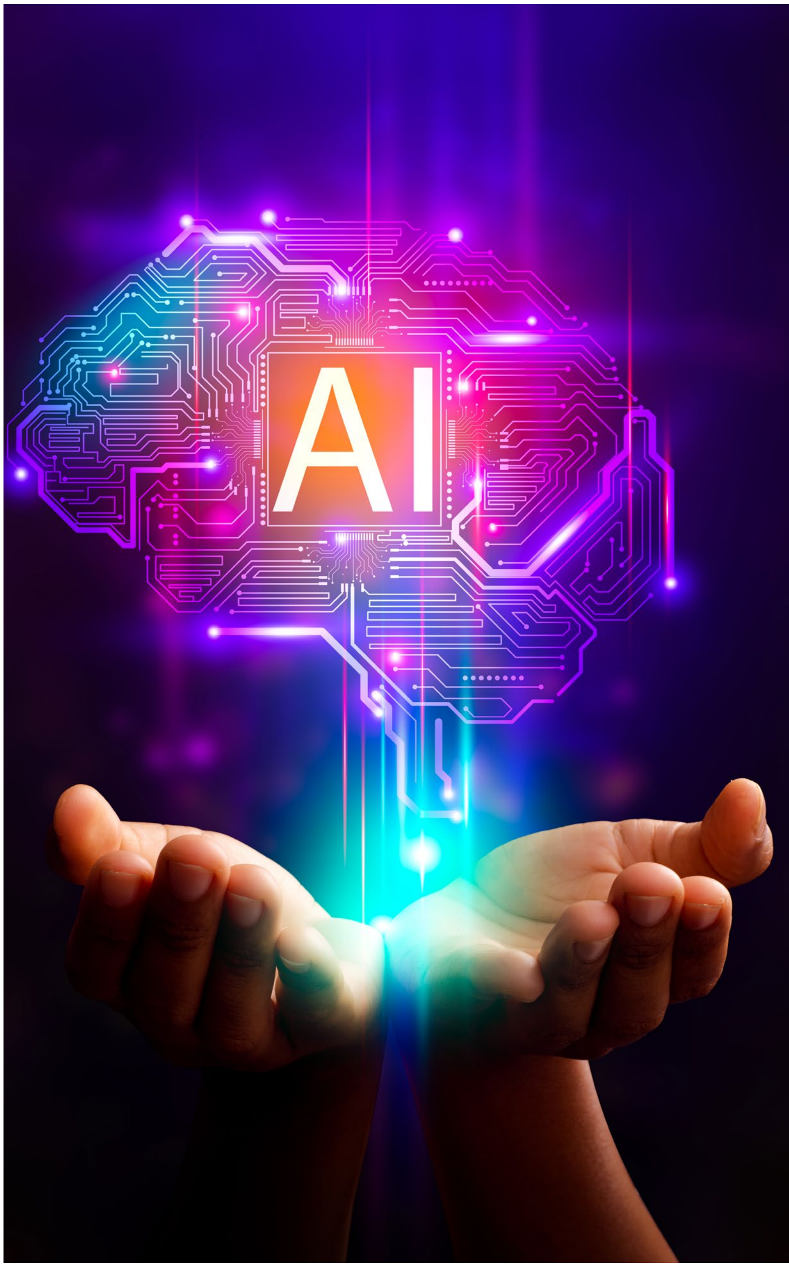
| | | |
|---|---|---|
| 1 | Security APIs | Prompt Injection & Jailbreak Check \| Adversarial Attacks \| Defense Mechanism |
| 2 | Privacy APIs | PII Detection & Anonymization (Text, Image, DICOM) |
| 3 | Explainability APIs | Feature Importance \| Chain of thoughts \| Thread of Thoughts \| Graph of Thoughts |
| 4 | Safety APIs | Profanity \| Toxicity \| Obscenity Detection \| Masking |
| 5 | Fairness & Bias APIs | Group Fairness \| Image Bias Detection \| Stereotype Analysis \| |

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

Data Types | ML & AI Models | AI Lifecycle Stage

### Here are some recently introduced features:

- Enhancing Sentiment Analysis with Keyword Visualization
- Enhanced Explainability: Infosys AI Guardrail Now Visualizes Traditional Model Predictions & use ReRead Prompt Technique
- Block Toxic Image Content
- Enhancing Image Understanding with Superpixels
- New Audio Support and Hallucination Detection
- PAIR – A New Red Teaming Feature for Enhanced Language Model Robustness



[240] https://www.prnewswire.com/news-releases/infosys-finacle-launches-data-and-ai-suite-to-help-banks-accelerate-their-ai-journey-302291390.html

## Contributors

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

**Syed Ahmed**
Global Head, Infosys Responsible AI Office, India

**Ashish Tewari**

Head of Infosys Responsible AI Office, India

**Srinivass**

Policy Advocacy, Consultancy and Customer Outreach, Infosys Responsible AI Office

**Mandanna Appanderanda N**

Head of Infosys Responsible AI Office, USA

**Rahul Pareek**

Head of Infosys Responsible AI Office, UK

**Arko Provo Ghosh**

Senior Consultant, Infosys Responsible AI Office, India

**Dakeshwar Verma**

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

**Utsav Lall**

Senior Associate Consultant, Infosys Responsible AI Office, India

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

Infosys Topaz is an AI-first set of services, solutions and platforms using generative AI technologies. It amplifies the potential of humans, enterprises and communities to create value. With 12,000+ AI use cases, 150+ pre-trained AI models, 10+ AI platforms steered by AI-first specialists and data strategists, and a 'responsible by design' approach, Infosys Topaz helps enterprises accelerate growth, unlock efficiencies at scale and build connected ecosystems. Connect with us at infosystopaz@infosys.com

For more information, contact askus@infosys.com

Infosys®
Navigate your next

Infosys.com | NYSE: INFY

Stay Connected