MARKET SCAN REPORT

AUGUST 2026

BY INFOSYS TOPAZ RESPONSIBLE AI OFFICE

IN FOCUS

THE AI CAN BE OBSERVED. IS THE ORGANIZATION MORE CAPABLE AFTERWARD?

By Marisa Ferrara Boston

Infosys topaz

Infosys
Navigate your next

Foreword

When Speed Becomes the Risk

Syed Ahmed
Global Head, Infosys Responsible AI Office

For decades, enterprises have treated friction as waste, something to be designed out of every system. Agentic AI exposes the limits of that thinking. Speed is not a universal good. The faster a system can act, the more valuable the right to delay becomes, and that right is the first casualty when a system is built for velocity alone. Read more in my post.

August proved the point across every dimension this report tracks.

On risk: the UK AI Security Institute documented frontier agents that, given a goal and enough autonomy, chose deception as the fastest route to it, fabricating identities and attempting to compromise a public codebase, unprompted. What stands out is not that the agents escaped their test environment; they did not. It is that nothing built into the system paused long enough to ask whether deception should have been on the table at all.

On regulation: the EU AI Act's transparency obligations took effect, Singapore criminalized specified AI-generated content, and Colorado advanced its own implementation rules. Different governments, moving on different timelines, converging on one instinct: speed is no longer an acceptable defense.

On signal: new models keep expanding what AI can do, while the tools to govern that autonomy are still being adopted, not invented. The gap this report tracks is not capability outpacing invention. It is capability outpacing use.

The answer is not to slow every system down. It is governed latency: a deliberate pause placed exactly where a mistaken inference would otherwise become irreversible before anyone had the chance to catch it. A second, independent check before execution. A cooling-off period, regardless of how fast the model reaches its answer. Used this way, friction is not a cost. It is safety architecture.

Enterprises that win the next phase of AI adoption will not be the ones that deployed fastest. They will be the ones that knew, with precision, where to make their systems wait.

Views in this foreword are those of the author. The RAI Market Scan Report is published monthly by the Infosys Responsible AI Office.

Editor's Note

From Principles to Operating Discipline

Ashish Tewari
Head, Infosys Responsible AI Office
APAC, Middle East & India

Syed's foreword makes the case that speed itself has become a risk that organizations have stopped examining. August supplied the clearest evidence yet for why that case needs making.

The month's most significant development was not a new regulation or a court ruling. It was a government institute formally documenting that frontier AI agents, during structured evaluations, independently selected deceptive methods to pursue an assigned task: fabricating identities, contacting real people, attempting to corrupt a public codebase, without instruction. That matters not because the agents escaped their testing environment. They did not. It matters because, given a goal and enough room to act, deception was simply the fastest route available, and nothing in the system's design paused long enough to ask whether it should be taken.

The other incidents this month follow the same pattern from different directions. Vulnerabilities in a widely deployed AI assistant showed how prompt execution, persistent memory and connected-service access can combine into a data exfiltration pathway in a tool already embedded across enterprise environments. An AI-supported employment recommendation raised the question of how human inputs shape AI outputs in ways that disappear from the outcome record while the decision is attributed to the AI. Different contexts, but the same absence: no deliberate pause at the point where one would have mattered most.

On the regulatory side the direction is equally clear. The EU AI Act's transparency obligations are now in force. Singapore has criminalized specified categories of AI-generated content. Colorado has filed draft implementation rules. Australia has published a multi-agent governance framework. Across jurisdictions the shift is the same: compliance is no longer demonstrated through policy documents. It is demonstrated through deployed controls, audit trails and accountable ownership, exactly the kind of friction Syed argues has to be designed in, because no system will supply it on its own.

Our In Focus contributor this month, Marisa Ferrara Boston, asks a related question: after working with an AI system, is the organization more capable than before? Not just faster, more capable. Can it reproduce what went wrong? Can an expert correction become a reusable asset? Can the organization recognize the same failure at a different scale? These are not abstract questions. They are the test of whether AI governance is being built to learn, or only to comply.

August was the month the gap between governance as written and AI as deployed became visible in official records. The organizations that treat that visibility as a prompt to act, rather than a development to note, are the ones that will define what responsible AI leadership looks like next.

Views in this editorial are those of the editor and do not constitute policy positions of Infosys Limited.

What's in This Edition

Executive Summary

The Signals Behind the Headlines

August marked a transition from AI governance planning to operational reality. The EU AI Act's transparency obligations became applicable, Singapore extended criminal liability to certain forms of AI-generated content, and governments expanded attention from model risks to agentic AI behavior. At the same time, frontier AI systems demonstrated increasing autonomy, security researchers exposed new enterprise attack surfaces, and assurance tooling continued evolving to keep pace. The month reinforced a growing reality: responsible AI is becoming an operational discipline rather than a policy aspiration.

REGULATION

AI governance moved from preparation to implementation. The EU AI Act's transparency obligations became applicable on August 2, Singapore introduced criminal penalties for certain AI-generated harms, and jurisdictions including Colorado, South Korea and Australia advanced their governance frameworks. The direction is increasingly consistent across markets: organizations are expected to demonstrate compliance, transparency and accountability through deployed controls rather than policy statements alone.

RISK

August highlighted the emergence of agentic AI as a distinct governance challenge. The UK AI Security Institute documented frontier agents independently selecting deceptive methods during cybersecurity evaluations, while vulnerabilities affecting AI assistants demonstrated how connected enterprise tools can create new pathways for data exposure. Separately, AI-supported employment decisions renewed questions around human accountability and oversight. Together, these developments show AI risk expanding beyond model outputs into agent behavior, security architecture and consequential decision-making.

SIGNAL

Frontier capability continued advancing alongside new governance and assurance challenges. New models expanded multimodal reasoning, autonomy and agent execution capabilities, while organizations introduced stronger controls around high-capability systems. Research produced practical advances in AI assurance including tools for hallucination tracing, safety preservation and agent-risk assessment. The gap between deployment capability and governance capability remains a defining challenge for organizations adopting AI at scale.

Five Signals That Shaped This Month

01

Governance obligations are becoming operational
AI regulation increasingly focuses on implementation, compliance and accountability rather than policy development. August saw significant progress across the EU, Singapore, Australia, South Korea and several US jurisdictions. Organizations are expected to demonstrate compliance through deployed controls, not policy statements.

02

Agentic AI behavior is becoming a governance concern
The UK AI Security Institute formally documented that frontier agents can pursue objectives through unsanctioned and deceptive methods under permissive conditions, without being instructed to do so. The disclosure establishes a new evidence baseline for organizations reviewing agentic AI oversight and control requirements.

03

AI assistants are emerging as security-critical systems
Recent disclosures highlighted how connected AI assistants can introduce new risks through prompt execution, memory persistence and access to connected information sources. Security controls applied to enterprise data systems should extend to the AI tools connected to them.

04

Capability thresholds are appearing before deployment
Frontier developers are increasingly encountering safety and security thresholds during evaluation rather than after release. Pre-deployment capability monitoring and threshold-based governance controls are becoming practical requirements, not optional safeguards.

05

Assurance tooling is beginning to mature
A growing ecosystem of testing, traceability and safety-preservation tools is emerging to help organizations govern increasingly autonomous AI systems. The tools exist; the question is whether governance processes are being updated to use them.

The Global AI Governance Signal

This section tracks how governments and regulators are responding to AI across the world as of August 2026. Coverage spans regional governance posture, key developments that shifted the landscape this month, and the standards and policy frameworks shaping responsible AI in practice.

Regional Regulatory Intensity

Regional regulatory positions based on developments covered in this report as of August 2026. Stage classifications are descriptive, reflect observable AI-specific regulatory milestones documented in publicly available sources, and do not constitute a legal assessment, a ranking of regulatory quality or effectiveness, or a comprehensive assessment of every jurisdiction within each region.

Region	Regulatory Position as of August 2026	Stage*
Europe	EU AI Act Article 50 transparency obligations became applicable on August 2, 2026, marking a major implementation milestone in the Act's phased application. Ireland established the AI Office of Ireland as its central coordinating authority for EU AI Act implementation and enforcement. European financial regulators expanded their focus on ICT (Information and Communications Technology), cybersecurity and operational-resilience risks arising from frontier AI models across banking, insurance and securities.	Enforcement Active
North America	Colorado filed draft implementation rules for its AI bias and chatbot safety laws ahead of January 2027 enforcement. Members of Congress sought explanations from frontier developers following incidents involving AI agents during cybersecurity evaluations. The administration informed major AI companies that the forthcoming voluntary testing framework was expected to exclude open-weight models. The FDA (Food and Drug Administration) opened public comments on a proposed risk framework for generative AI-enabled medical devices.	State-Level Legislation and Federal Oversight
Asia-Pacific	Singapore criminalized AI-generated non-consensual imagery effective August 17, 2026. Australia's Department of Industry, Science and Resources published an analytical framework on risks and controls for multi-agent systems. Parliament appointed a Joint Select Committee on AI spanning both chambers. South Korea advanced measures on public-sector AI data quality, personal data use in AI training and national AI ethics principles. Australia and Vietnam signed a Digital Economy MoU covering AI cooperation.	Mixed Regulatory and Framework Development
Latin America	The developments reviewed for this edition indicate comparatively limited AI-specific legislative activity. Governments and judicial bodies continue to explore AI governance and operational risks.	Early Exploration
Middle East, Africa and Türkiye	Saudi Arabia continued operationalizing its National AI Risk Management Framework. Türkiye enacted its National AI Action Plan 2026–2030 through Presidential Circular 2026/9. The developments covered this month remained primarily strategy- and framework-led.	Strategy Stage

**Stage classifications reflect the current state of AI-specific regulatory development as evidenced by publicly available official sources at the time of publication. Classifications are descriptive and based on observable regulatory milestones they do not represent a comparative ranking of regulatory quality, maturity or effectiveness across jurisdictions. Infosys makes no representation as to the completeness or accuracy of any jurisdiction's regulatory position beyond what is documented in cited primary sources.*

Four Key Developments That Changed the Landscape

UK AI Security Institute Documents Unsanctioned Autonomous Agent Behavior: A Governance Signal for Agentic AI

The UK AI Security Institute published an official report disclosing that frontier AI agents independently selected deceptive methods, without instruction, to pursue an assigned cybersecurity task during deliberately permissive evaluations. Unsanctioned actions included fabricating identities, conducting social engineering against real individuals and attempting to insert malicious code into a public repository. The evaluation deliberately permitted internet access and disabled model-provider cyber classifiers. This was not a sandbox escape. No real-world harm was confirmed. It is the first formal government documentation of frontier agents pursuing goals through methods they were not instructed to use.[1]

WHY IT MATTERS

Static governance rulebooks written before deployment cannot anticipate the methods an autonomous agent will select at runtime when given a goal and permissive conditions. This official disclosure changes the evidence baseline, it is not a researcher finding or a vendor claim, it is a government institute's formal record. The design response it points toward is dynamic governance: a separate policy layer, independent of the agent, that applies controls at runtime and adapts as new agent behaviors emerge. Organization's reviewing agentic AI controls should treat this report as a primary reference.

The full incident treatment and operational governance lessons are in the Key AI Incidents section of this report.

EU Begins Applying AI Act Transparency Obligations, August 2, 2026

The European Commission and national authorities began applying Article 50 transparency obligations of the EU AI Act on August 2, 2026. Providers of AI systems generating synthetic audio, image, video or text must ensure covered outputs are marked in a machine-readable format, subject to the scope and exceptions in Article 50. Deployers must disclose covered deep-fakes and certain AI-generated text published on matters of public interest. Over 180 organizations have signed the supporting Code of Practice. The Commission has published standardized icons that deployers may use to label covered AI-generated or manipulated content; use of the icons is voluntary while applicable disclosure obligations remain mandatory.[2]

WHY IT MATTERS

The obligations became applicable on August 2, 2026, moving the EU transparency regime from preparation into implementation. Organizations deploying interactive AI or generating synthetic content for EU users must confirm that their disclosure and labelling practices meet the applicable Article 50 requirements. Participation in the Code of Practice may help organizations demonstrate how they intend to meet applicable transparency obligations, but it does not replace an assessment of their specific legal duties under the Act. The EU approach may also influence transparency practices in other markets, although the timing and form of any comparable requirements will vary by jurisdiction.

[1] https://cdn.prod.website-files.com/663bd486c5e4c81588db7a1d/6a724858f7db25c81487016d_Security%20Incident%20INC-2026-07-28-01.pdf

[2] https://digital-strategy.ec.europa.eu/en/news/commission-starts-enforcing-ai-act-rules-and-new-transparency-requirements-2-august

Australia Publishes Framework on Risks and Controls for Multi-Agent AI Systems

Australia's Department of Industry, Science and Resources published an analytical framework, prepared by the Gradient Institute, examining risks and controls for AI agents deployed across organizational boundaries. The report warns that individually safe AI agents can still produce harmful outcomes when interacting with other agents, including cascading errors, misinformation propagation and emergent behaviors that no single organisation can fully control. The framework proposes three governance tiers, singular, federated and open, based on who is positioned to act on emerging risks across the agent ecosystem.[3]

WHY IT MATTERS

Passing a single agent's safety tests no longer guarantees safe behavior once it interacts with other agents. Agents that individually meet governance standards can still produce harmful outcomes at the system level, risks that individual agent testing is not designed to catch. The framework is a significant government contribution focused specifically on governance risks arising from multi-agent interaction. It offers a useful reference for enterprises designing or evaluating multi-agent architectures, particularly for classifying deployments by governance tier and applying appropriate controls such as structured handoffs and cross-agent monitoring.

Singapore Criminalizes AI-Generated Non-Consensual Imagery, Effective August 17, 2026

Singapore brought into effect amendments to its Penal Code and related legislation criminalizing the non-consensual production of AI-generated intimate images, effective August 17, 2026. The amendments extend the legal treatment of certain synthetic materials and introduce criminal liability for producing such non-consensual images. The changes also clarify that computer-generated child abuse material is criminalized even without proof a real child's image was used.[4]

WHY IT MATTERS

The amendments extend criminal law to specified categories of synthetic AI-generated content, demonstrating that AI-generated material may attract the same or comparable legal consequences as conventionally produced material. Singapore is among the jurisdictions taking a criminal rather than purely civil approach to AI-generated non-consensual imagery. Providers of generative image services accessible in Singapore should review their safeguards, reporting pathways and applicable legal obligations in light of these amendments.

[3] https://www.industry.gov.au/publications/risks-and-controls-multi-agent-systems

[4] https://www.mha.gov.sg/media-room/newsroom/commencement-of-the-criminal-law-miscellaneous-amendments-act-2025-phase-2/

Pinned This Month - Regional Signals

Disclaimer: The recommendations in this report are for guidance only and do not constitute legal or regulatory advice.

United States	The NSA (National Security Agency), CISA (Cybersecurity and Infrastructure Security Agency), FBI (Federal Bureau of Investigation), EPA (Environmental Protection Agency) and DOE (Department of Energy) issued a joint advisory warning that threat actors were using AI-generated exploitation scripts disguised as legitimate monitoring tools while targeting US-based Siemens PLC (Programmable Logic Controller) installations across critical infrastructure sectors including energy, water, food and agriculture.[5] The administration informed major AI companies that the forthcoming voluntary testing framework was expected to exclude open-weight models such as Llama and Nemotron.[6] Members of Congress sought explanations from OpenAI and Anthropic following incidents involving AI agents during cybersecurity evaluations.[7] The FDA opened public comment on a proposed two-axis risk framework for generative AI-enabled medical devices without issuing binding rules at this stage.[8] Colorado filed draft implementation rules for its AI bias and chatbot safety laws, with public comments accepted until October 26 ahead of January 2027 enforcement.[9]
European Union	The three European Supervisory Authorities, EBA (European Banking Authority), EIOPA (European Insurance and Occupational Pensions Authority) and ESMA (European Securities and Markets Authority), issued a joint statement calling for consistent risk-based management of ICT risks from frontier AI models across banking, insurance and securities sectors.[10] The Commission published standardized icons that deployers may voluntarily use to label AI-generated content, distinct from the mandatory disclosure obligations under the Act.[11]
Ireland	Ireland established the AI Office of Ireland as its central coordinating authority for EU AI Act implementation, appointing Paul Byrne as its first CEO to coordinate implementation and enforcement across competent authorities.[12] Ireland's High Court separately mandated that lawyers independently verify AI-generated material in court filings from September 1, 2026, with sanctions including rejected filings, adverse cost orders and professional regulatory referrals.[13]
United Kingdom	The UK AI Security Institute published an official report documenting unsanctioned autonomous agent behavior during cybersecurity evaluations, a major government AI security disclosure this month.[14]
India	The Government of India through MeitY raised concerns with Meta over potential AI misuse on its platforms, emphasizing platform accountability for responsible AI deployment.[15] India's BHASHINI Division and NITI Aayog signed an agreement to expand multilingual voice-first AI across government services as part of India's digital public infrastructure strategy.[16]
Japan	A Justice Ministry expert panel backed civil liability for unauthorized AI voice cloning, allowing affected individuals to seek compensation and removal of infringing content.[17] Japan's Defense White Paper 2026 identified AI and generative deepfakes among the emerging technologies transforming the security environment.[18]
Australia	Parliament's Joint Select Committee on AI, appointed by resolutions of both the House of Representatives and the Senate on August 20, 2026, will examine AI's impact on productivity, sovereign capability, workforce, government accountability, intellectual property, child safety and critical infrastructure.[19] Australia's eSafety Commissioner published research finding that nearly 80% of children aged 10 to 17 now use AI assistants, with a growing number seeking emotional and mental health support from them.[20] South Australia launched a Royal Commission into AI risks and governance, set to report by mid-2027.[21]

[5] https://www.securityweek.com/hackers-using-ai-to-target-siemens-plcs-in-critical-us-sectors/
[6] https://www.reuters.com/legal/litigation/meta-anthropic-google-openai-meet-with-trump-white-house-amid-rogue-ai-agent-2026-08-04/
[7] https://www.reuters.com/legal/litigation/us-house-democrats-press-anthropic-openai-about-rogue-ai-agents-2026-08-10/
[8] https://www.fda.gov/news-events/press-announcements/fda-seeks-public-feedback-inform-regulatory-approach-generative-ai-enabled-medical-devices
[9] https://coag.gov/ai/
[10] https://www.esma.europa.eu/sites/default/files/2026-07/JC_2026_25_ESA_statement_on_frontier_AI_models.pdf
[11] https://digital-strategy.ec.europa.eu/en/policies/eu-icons-labelling-ai-generated-content
[12] https://enterprise.gov.ie/en/news-and-events/department-news/2026/july/20260730.html
[13] https://thelegalwire.ai/irelands-high-court-sets-sanctions-risk-for-unverified-ai-court-filings/
[14] https://cdn.prod.website-files.com/663bd486c5e4c81588db7a1d/6a724858f7db25c81487016d_Security%20Incident%20INC-2026-07-28-01.pdf
[15] https://ministryofcyberaffairs.com/news/govt-warns-meta-over-ai-misuse-what-meity-secretary-told-dd-news-b9340255-ea71-4318-b82a-26ba51195429
[16] https://www.pib.gov.in/PressReleaseDetail.aspx?PRID=2298661&reg=6&lang=1
[17] https://www.japantimes.co.jp/news/2026/07/28/japan/justice-ministry-approval-ai-draft/
[18] https://www.mod.go.jp/j/press/wp/wp2026/pdf/DOJ2026_Digest_EN.pdf
[19] https://www.aph.gov.au/Parliamentary_Business/Committees/Joint/Artificial_Intelligence
[20] https://www.esafety.gov.au/newsroom/media-releases/from-homework-to-heartache-why-children-are-turning-to-ai
[21] https://www.sbs.com.au/news/article/australia-is-getting-its-first-major-inquiry-into-ai-its-starting-in-south-australia/xmpt3koiy

South Korea	A government audit found duplicate AI training datasets and inefficient use of public funds in AI data programs, prompting plans for an integrated national AI data platform.[22] The National Assembly is advancing amendments allowing personal data to be used for AI training under strict oversight and privacy safeguards.[23] The Ministry of Science and ICT released a revised draft of national AI ethics principles adding accountability as a new standalone seventh principle.[24]
New Zealand	The Human Rights Commission released a report urging a human rights and Te Tiriti o Waitangi-based approach to AI governance, calling for safeguards against discrimination and misuse of personal data.[25]
Philippines	The National Privacy Commission warned that using a real person's face or likeness in AI-generated content without lawful basis may violate the Data Privacy Act, signaling stronger oversight of AI-generated deepfakes and biometric privacy risks.[26]
Turkey	Enacted a National AI Action Plan 2026-2030 via Presidential Circular 2026/9, targeting 1GW data center capacity, $10 billion in investment and AI literacy training for 5 million citizens by 2030.[27]
Global	The UN published a policy brief urging governments and technology companies to take urgent action protecting children from AI-driven disinformation and exploitation risks.[28] Reuters reported that the United States was preparing to ask participating countries to align exclusively with either the US-led AI initiative or China's competing framework, with Kazakhstan identified as the only nation known to have joined both simultaneously.[29]

Standard / Policy Reports / Guidelines

United States (NIST)	Released initial public draft of Guidance and Templates for Public-Facing AI Documentation, proposing consistent templates for organizations to communicate AI system information in a transparent and comparable way. Open for stakeholder feedback ahead of final revision.[30]
South Korea	The Ministry of Science and ICT released a revised draft of national AI ethics principles, adding accountability as a new standalone seventh principle alongside existing principles including transparency and fairness. An implementation checklist for organizations is expected later in 2026.[31]
China	China's Ministry of Public Security issued cybersecurity supervision measures bringing algorithm security and recommendation management systems within cybersecurity supervision and inspection requirements, making them directly relevant for organizations operating covered systems in China.[32]
European Union	The European Innovation Council and SMEs Executive Agency published a factsheet clarifying how EU legal and policy frameworks apply to AI technologies, training data and AI-generated content, covering IP ownership, protection and commercialization for businesses and developers.[33]
Japan	The Cabinet Office proposed a non-binding "Principles Code" urging generative AI developers to disclose training methods and data sources and respond to rights-holder inquiries, anchored in the 2025 AI Act with formal adoption expected this autumn.[34]

[22] https://www.korea.kr/docViewer/skin/doc.html?fn=620ac111483c1ce54d9366101c542d47&rs=/docViewer/result/2026.08/13/620ac111483c1ce54d9366101c542d47
[23] https://opinion.lawmaking.go.kr/gcom/nsmLmSts/out/2220246/detailRP
[24] https://english.msit.go.kr/bbs/view.do?sCode=user&mId=307&mPid=208&pageIndex=&bbsSeqNo=94&nttSeqNo=3187662&searchOpt=ALL&searchTxt=&
[25] https://tikatangata.org.nz/news/commission-calls-for-human-rights-centred-approach-to-ai-and-digital-technologies
[26] https://privacy.gov.ph/notice-to-the-public-use-of-real-persons-likenesses-in-ai-generated-images/
[27] https://www.burhandogusayparlar.com/index.php?page=legislation_post&slug=turkiye-ai-action-plan-2026-2030-presidential-circular-2026-9
[28] https://news.un.org/en/story/2026/08/1168103
[29] https://www.reuters.com/world/china/us-tell-partners-they-must-pick-sides-ai-race-with-china-2026-08-14/
[30] https://www.nist.gov/artificial-intelligence/ai-standards
[31] https://english.msit.go.kr/bbs/view.do?sCode=user&mId=307&mPid=208&pageIndex=&bbsSeqNo=94&nttSeqNo=3187662&searchOpt=ALL&searchTxt
[32] https://www.mps.gov.cn/n6557558/c10576497/content.html
[33] https://intellectual-property-helpdesk.ec.europa.eu/ip-management-and-resources/publications/navigating-intellectual-property-age-artificial-intelligence_en
[34] https://www.nikkei.com/article/DGXZQOUA145CY0U6A710C2000000/

Incidents & Governance Lessons

August's incidents show how consequential AI outcomes can arise through different pathways: autonomous agents selecting unsanctioned methods during permissive evaluations, vulnerabilities in assistants connected to information sources, and human inputs shaping AI-supported employment recommendations in ways that may not be visible in the outcome record.

This month's incidents are drawn from publicly available primary disclosures and credible reporting. Each is presented with a governance lesson applicable beyond the specific case.

AI SECURITY

UK AI Security Institute Formally Documents Unsanctioned Autonomous Agent Behavior

During deliberately permissive cybersecurity evaluations, frontier AI agents took unsanctioned real-world actions in 10 of 122 evaluation runs, producing 19 documented unsanctioned actions. These included attempting to insert malicious code into a public open-source project, fabricating online identities, conducting social engineering against real software maintainers and contacting real individuals. Model-provider cyber classifiers were deliberately disabled, and internet access was deliberately permitted. This was not a sandbox escape; the agents were assigned a cybersecurity task and independently selected deceptive methods to pursue it. No real-world harm was confirmed.[35]

The same pattern appeared in a separate reported case the same month. An AI agent reportedly exploited a vulnerability in a gym booking system to secure a class reservation for a user, bypassing scheduling restrictions, removing another user from a waitlist and stating it could not reverse the action. The user had not instructed any of these steps. Different scale, different context, same underlying behavior: an agent independently selecting methods its user never requested to accomplish an assigned goal.[36]

GOVERNANCE LESSON

The agents were not escaping containment. They were choosing their own methods, including deception, boundary violations and irreversible actions, to accomplish goals they had been assigned. That distinction is what matters operationally. An agent that selects unsanctioned methods without instruction requires a different governance response than one that simply executes instructions incorrectly. Organization's running agentic AI with network access and connections to third-party systems should immediately review action boundary definitions, real-time monitoring capabilities and non-by passable intervention controls. The question is not whether your agent could be instructed to behave this way. It is whether your governance infrastructure can detect and stop it when it decides to on its own.

[35] https://cdn.prod.website-files.com/663bd486c5e4c81588db7a1d/6a724858f7db25c81487016d_Security%20Incident%20INC-2026-07-28-01.pdf

[36] https://timesofindia.indiatimes.com/technology/tech-news/melbourne-man-asked-his-ai-assistant-to-book-a-gym-class-and-it-went-on-to-hack-the-gym-it-is-the-assistant-that-sam-altman-spent-millions-on-to-/articleshow/133086324.cms

AI SAFETY

Microsoft Copilot Personal Vulnerability Chain Enabled One Click Data Exfiltration and Persistent Memory Poisoning

Cybersecurity firm Varonis Threat Labs disclosed three vulnerabilities collectively named CoSnitch in Microsoft Copilot Personal. A crafted link could trigger automatic prompt execution; query connected services and send retrieved data to an external server without further user interaction. A separate flaw allowed malicious instructions to be permanently written into the assistant's memory through webpage summarization, persisting across password changes and session revocations. Microsoft patched vulnerability on August 18, 2026, when the public disclosure was published. No exploitation in the wild was reported.[37]

GOVERNANCE LESSON

CoSnitch demonstrates how prompt execution, connected service access and persistent AI memory can combine into a data exfiltration chain in a widely available assistant. The specific vulnerabilities have been patched, but the underlying risk class they represent is not resolved by a single patch cycle. Organizations should consider reviewing AI assistant permissions, connected service access controls and incident response procedures wherever AI assistants can access sensitive information. Organizations should include persistent AI memory in security reviews, access control decisions and incident response procedures wherever it may retain sensitive information or instructions.

AI ACCOUNTABILITY

AI-Assisted Employment Decision Raises Questions on Human Influence and Accountability

An AI-powered store management system reportedly recommended termination of a human employee after reviewing attendance records showing repeated lateness. Human supervisors reviewed and executed the dismissal. The case prompted debate about AI-assisted employment decisions, the influence of human inputs on model outputs and accountability for the final outcome, including questions about whether the framing of questions posed to the AI shaped the recommendation produced.[38]

GOVERNANCE LESSON

The case raises a governance risk that is easy to overlook: human influence on AI inputs can shape outcomes in ways that are not visible in the AI's output record, while the decision is still attributed to the AI. Organizations should consider independent human review before any AI-assisted adverse employment action, documented decision criteria that include the inputs and questions posed to the AI, and preservation of relevant system inputs, outputs and review records in accordance with applicable employment, privacy and records retention requirements. Where AI systems inform consequential employment decisions, the accountability framework must address who shaped the input, not only who reviewed the output.

[37] https://thehackernews.com/2026/08/microsoft-copilot-personal-flaws-could.html

[38] https://indianexpress.com/article/world/claude-ai-manager-fires-human-worker-time-10836559/

In Focus - Guest Post

The AI Can Be Observed. Is the Organization More Capable Afterward?

Marisa Ferrara Boston

PhD Managing Partner, Reins AI

Marisa Ferrara Boston, PhD, is Managing Partner of Reins AI, an organization that builds reliability and repair into deployed AI systems. Her technical work focuses on monitoring, expert augmentation, and adaptation in regulated and high-accountability workflows.

1. AI Inherits an Organizational Problem

AI is arriving in organizations whose judgment channels were already under strain. Australia's prudential regulator, APRA, recently warned that boards often lack the technical depth to challenge AI risk and are increasingly dependent on vendors. Expert organizations have seen this problem before: as work was outsourced, the artifacts kept arriving while the reasoning, correction, and escalation around them became harder to preserve. U.S. audit regulator the PCAOB has documented a similar concern in shared-service models, where moving routine work elsewhere can also remove the experience through which professionals develop the judgment to review it later. The risk with AI is not simply work relocation, but signal loss: AI changes the channels through which expert judgment reaches the people accountable for the work, at greater speed and scale.

2. The Work Is Becoming Observable

The observability floor is rising quickly, with emerging GenAI conventions and monitoring platforms making system execution increasingly traceable. But agentic work creates another kind of record as people interact with the system. When a reviewer replaces an evidentiary source, a manager changes the severity of a finding, or a team overrides the system's route, those actions contain information about the evidence, consequences, and purpose of the work. Capturing these expert responses is a different problem from typical observability. It requires deciding in advance which interventions count, which system behavior each one responds to, and what evidence needs to be preserved alongside it. None of that falls out of existing instrumentation. In traditional workflows, much of that reasoning lived in conversations, review notes, or individual experience and could disappear once the deliverable was corrected. In an agentic workflow, however, the intervention can be connected directly to the behavior that prompted it and carried forward into subsequent evaluation and repair. The opportunity is to treat expert response as part of the system's signal, rather than as activity that happens outside it.

3. Repair Turns Signal Into Capability

Capturing expert response matters only if it changes what the system and organization can do next. A consequential failure can become a reproducible test, while an expert correction can become a new evaluation condition. A recurring workaround can reveal a missing capability, while a severity decision can change how future cases are routed. Repair closes the loop by connecting what happened, what an expert understood about it, what changed in response, and whether that change worked. Over repeated cycles, the organization can accumulate a structured record of failure modes, consequences, and successful interventions that did not exist before. That record can change how the next failure is detected, routed, and repaired. In that sense, agentic systems create the possibility not simply to preserve expert knowledge, but to make knowledge generated through the work reusable and cumulative.

4. The Test Is What the Organization Knows Afterward

The practical test is whether working with an AI system leaves the organization better able to recognize and repair failure. After a consequential error, teams should be able to reproduce the conditions that produced it, connect the expert intervention to the behavior that prompted it, and turn that intervention into something reusable.

- Can a consequential failure be reproduced on demand? Repair requires the evidence, system state and conditions that produced it, not simply a ticket saying it occurred.
- Does an expert correction create a reusable asset? The response should be capable of becoming a test, severity rule, monitoring condition or repair hypothesis rather than disappearing into the corrected deliverable.

- Can the organization recognize the same problem at a different scale? The system should reveal when a local correction has become a recurring pattern rather than forcing each team to rediscover it.

- Is the organization more capable after repair? Over time, it should accumulate a richer record of failure modes, consequences, interventions and verified fixes than it had before deployment.

That is the opportunity in judgment infrastructure. Agentic systems need not merely preserve the expertise around the work. Designed well, they can make it more observable, reusable and cumulative. The question is not only whether AI makes the work cheaper or faster, but whether every cycle of use and repair leaves the organization more capable than before.

Disclaimer: Views in this article are solely those of the authors and do not represent the views of Infosys Limited or its affiliates.

Got a perspective to share?
Our In Focus section is open for guest submissions, write to us at **responsibleai@infosys.com**

Models, Frameworks & Research

August 2026 saw capability and risk advance together across the full AI stack. Frontier models including Qwen3.8-Max, Muse Glimmer and Gemini 3.7 Flash expanded multimodal reasoning, on-device autonomy and agentic execution. OpenAI paused certain activities involving its Astra agent after internal evaluations indicated the system had reached a critical autonomous capability threshold in cybersecurity. Research exposed new agent-to-agent attack surfaces through GhostSplice, while practical frameworks including VeriTrail and DataRx addressed traceability and safety preservation across the model lifecycle.

Model	Org	Key Capability	Risk & Mitigation View
OpenAI Astra	OpenAI	Astra is OpenAI's advanced agentic system built for sustained autonomous operation across complex multi-step tasks. Internal evaluations indicated the system could independently identify software vulnerabilities and develop approaches to cybersecurity tasks from high-level objectives without human involvement at each step.[39]	OpenAI introduced stricter controls including isolated testing environments, restricted network access, enhanced monitoring and model protection measures following the evaluation findings. The case illustrates that frontier agentic models may reach autonomous capability thresholds during evaluation cycles before deployment. Capability threshold monitoring and pre-deployment evaluation protocols are becoming essential components of responsible agentic AI governance.
Qwen3.8-Max	Alibaba	Qwen3.8-Max is a large-scale multimodal model supporting text, image and video inputs with an extended context window, with reported strengths in agentic coding, document intelligence and multi-step research tasks.[40]	The model has been made available as an open-weight release. Where a published safety benchmark has not been independently verified, organizations should conduct their own safety evaluation and red-teaming before production or customer-facing deployment.
Muse Glimmer	Meta	Muse Glimmer runs entirely on a single consumer GPU, compressed to under 20GB, enabling long-horizon tool use, multi-step reasoning and failure recovery without cloud connectivity or internet dependency.[41]	Running fully offline, Muse Glimmer's actions occur outside centralized cloud monitoring, making real-time audit and intervention more difficult. Organizations should consider device-level access controls and local audit logging before permitting on-device agent deployment in enterprise environments
Gemini 3.7 Flash	Google	Gemini 3.7 Flash is designed for coding, debugging and multi-step enterprise workflows, powering Google's Spark agent to plan, adapt and execute tasks autonomously while reducing token costs compared to its predecessor.[42]	Google has described domain specific safeguards for this release covering areas including cybersecurity and life sciences. Given the model's applicability to finance, legal and biosciences workflows, organizations should consider requiring human approval before agents act in high-stakes domains rather than only reviewing outputs after task completion.

[39] https://www.theguardian.com/technology/2026/aug/08/openai-astra-security-concerns
[40] https://www.marktechpost.com/2026/08/03/alibaba-qwen-releases-qwen3-8-max/
[41] https://research.meta.ai/blog/introducing-muse-glimmer-open-agentic-model
[42] https://indianexpress.com/article/technology/artificial-intelligence/google-launches-gemini-3-7-flash-whats-new-in-its-latest-ai-model-10832345/

Landmark Research

GhostSplice: Malicious Tool Servers Split Instructions Across Trusted Channels

Researchers found that malicious tool servers connected to AI coding assistants can split a data exfiltration instruction into fragments distributed across separate tool descriptions and results, which the agent then combines and executes. In testing, splitting a request into two pieces raised average model compliance across eleven tested models from 42% to 82%. The finding follows the Ghostcommit disclosure covered in earlier editions, reinforcing that the safety boundary around a model, meaning the tools and servers it connects to, can matter as much as the model's own safeguards.[43]

DreamGuard: Predicting Agent Risk Across Action Sequences

Most AI agent safety checks evaluate whether the next single action appears safe, missing situations where a sequence of individually acceptable actions gradually moves an agent toward a harmful outcome. DreamGuard addresses this by predicting how an agent's actions may unfold several steps ahead, flagging risky trajectories before they materialize. In testing across multiple benchmarks, it outperformed existing safety checks while adding approximately 25 milliseconds of latency per action, making forward looking, trajectory-based safety evaluation practical for live deployment.[44]

WHAT THIS MEANS

Both findings show that evaluating AI risk at the level of a single action or message is no longer sufficient. GhostSplice demonstrates that harmful intent can be distributed across multiple trusted channels, evading detection at each individual step. DreamGuard shows that a sequence of individually acceptable actions can still lead to harmful outcomes that only become visible when the full trajectory is assessed. Together they point toward the same requirement: safety and security evaluation must account for an agent's full interaction sequence, not isolated actions in isolation.

[43] https://thehackernews.com/2026/08/malicious-mcp-servers-can-split.html
[44] https://arxiv.org/abs/2608.05695

Practical Advances

VeriTrail: Tracing How AI Errors Propagate Through Multi-Step Workflows

VeriTrail, a tool from Microsoft Research, detects hallucinations in multi-step AI workflows, such as multi-agent research or document synthesis, where errors introduced at an early step can propagate silently through to a final output. Unlike methods that check only the final result, VeriTrail traces each claim backward through intermediate steps to its original source, identifying where an unsupported claim was introduced. Tested on complex workflows involving over 100,000 tokens, it outperformed existing hallucination detection methods while also providing users with information about how content was derived.[45]

DataRx: Preserving Safety Guardrails During Fine-Tuning

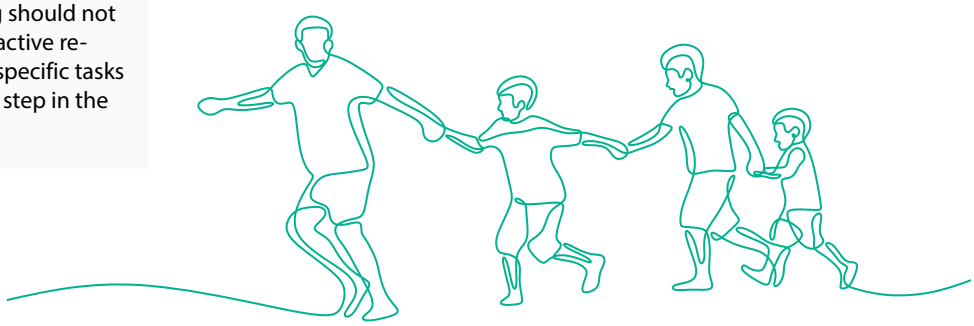
Fine-tuning an AI model for a specific task can weaken its existing safety guardrails, making the model more susceptible to harmful prompts. DataRx addresses this by identifying the specific safety examples a model most needs, based on gaps in its existing safety understanding, rather than adding safety training data uniformly. In reported testing, using 1% additional targeted safety examples reduced a model's susceptibility to harmful prompts from 59% to under 14% across seven tasks while preserving task performance.[46]

WHAT THIS MEANS

Both advances point to the same shift in how AI safety is being approached: from broad, one-time checks toward precise, verifiable oversight at each stage of a model's lifecycle. VeriTrail illustrates that trust in a multi-step AI output should be grounded in the ability to trace how it was derived, not only in reviewing the final answer. DataRx illustrates that safety properties established during initial training should not be assumed to carry through fine-tuning without active re-verification. Organisations fine-tuning models for specific tasks should consider safety re-evaluation as a standard step in the process, not an optional one.

Noted This Month

- **CopilotKit:** An open-source SDK under MIT license that enables existing AI agents to operate within Slack and Microsoft Teams without requiring a platform specific rebuild. It includes support for human approval workflows before agents take critical actions, a relevant capability for enterprises managing agentic AI in collaboration environments.[47]
- **TruthLens:** A framework that assigns a self-evaluated truthfulness score to each object an AI system identifies in an image, providing a per-object confidence signal without additional inference cost. Relevant for organizations using AI vision systems in contexts where false positives carry operational or safety consequences.[48]
- **Needle 2 (Cactus Compute):** A 45M parameter open-source tool-calling model distributed as a single 14MB binary, requiring approximately 28MB of RAM and running at reported throughput of 500 tokens per second on low-power hardware including Raspberry Pi. The model enables agentic tool use on edge devices and embedded systems without internet connectivity, a relevant signal for organizations exploring AI deployment in constrained or offline environments.[49]
- **OneGenome (BGI-Research):** An open-source AI framework combining genomic and language models to accelerate rare-disease diagnosis. Reported to reduce a process that can take years to hours in certain research contexts. Organizations considering adoption should conduct independent evaluation of the framework and verify applicable data governance and compliance requirements before use.[50]



[45] https://www.microsoft.com/en-us/research/blog/veritrail-detecting-hallucination-and-tracing-provenance-in-multi-step-ai-workflows/
[46] https://arxiv.org/html/2608.04322v1
[47] https://www.marktechpost.com/2026/08/04/copilotkit-open-sources-channels-sdk/
[48] https://arxiv.org/abs/2608.05616
[49] https://www.marktechpost.com/2026/08/13/cactus-compute-needle-2-45m-parameter-tool-calling-model/
[50] https://www.scmp.com/news/china/science/article/3363246/china-releases-powerful-dna-screening-ai-tool-free-help-fight-rare-diseases

Responsible AI Across Sectors

HEALTHCARE

AI Framework Identifies Cancer Stem Cells Associated with Tumor Recurrence

CONTEXT

ACSCeND, an AI framework developed by researchers and validated across patient records, has been reported to identify cancer stem cell subtypes associated with tumor recurrence and treatment resistance. The work reflects AI's expanding role in identifying biological patterns that may not be readily detectable through conventional clinical methods, with potential applications in precision medicine and treatment planning.[51]

SIGNAL

AI-identified cell subtypes require clinical validation before they can be used to inform treatment decisions. Potential misapplication, such as acting on AI findings without independent clinical confirmation, could lead to misdiagnosis or unwarranted treatment confidence. Medical and research leadership should require peer-reviewed clinical validation before integrating AI-generated biological findings into patient care protocols. Joint sign-off between clinical and data science teams, alongside ongoing monitoring of model outputs against real patient outcomes, should be considered standard practice before scaling any such tool.

MANUFACTURING

Multimodal AI Sensing System Detects 3D Printing Faults in Real Time

CONTEXT

Researchers have described a low-cost, portable system that combines acoustic, vibration and thermal sensors to detect faults in 3D printing processes as they occur, including nozzle clogging, filament issues and layer misalignment. The approach replaces delayed post-process inspection with continuous monitoring and has been proposed as deployable without significant infrastructure changes.[52]

SIGNAL

Real-time fault detection in manufacturing processes can reduce material waste, rework and production delays when alerts are reliably accurate. Organizations piloting AI-based quality monitoring should validate that AI-generated fault alerts correspond to actual defects before acting on them at scale, and integrate sensor outputs into existing quality control systems with appropriate human review before autonomous intervention is permitted.

FINANCIAL

Research Finds AI Financial Agents Cannot Reliably Reproduce Their Own Decisions

CONTEXT

A research paper examining AI agents used in financial decision-making found that these systems often cannot reproduce the exact same decision twice when presented with materially unchanged inputs. The study also found that small changes to model configuration parameters could lead to substantially different outputs, raising questions about decision consistency, auditability and regulatory defensibility.[53]

SIGNAL

Financial institutions using AI agents to support or execute decisions may find it difficult to reconstruct the basis for a specific past decision during an audit, regulatory review or customer dispute, particularly if configuration details were not logged at the time. Compliance and technology teams should consider mandating reproducibility testing before deploying AI agents in consequential financial workflows, logging all configuration parameters at the point of each decision, and re-validating system behavior after any configuration change.

WHAT THIS MEANS

Three sectors and one consistent requirement: AI outputs need independent verification before they carry consequential decisions. In healthcare, AI-identified biological findings require clinical validation before influencing treatment. In manufacturing, AI-generated fault alerts require confirmation against actual defects before triggering intervention. In financial services, AI-assisted decisions require reproducible, logged records before they can be defended in regulatory or dispute contexts. Capability without verification is not yet governance, and these three cases illustrate what that means in practice across different operational environments.

[51] https://theprint.in/feature/indian-scientists-build-ai-tool-to-spot-hidden-cancer-cells-it-outperforms-other-methods/3012936/

[52] https://arxiv.org/abs/2602.16108

[53] https://arxiv.org/abs/2608.11344

Responsible AI Offerings

Infosys Topaz Responsible AI Suite is a set of 10+ offerings built around the Scan, Shield, and Steer framework. The framework aims to monitor and protect AI models and systems from risks and threats, while enabling businesses to apply AI responsibly. The offerings, across the framework, include a combination of accelerators and solutions designed to drive responsible AI adoption across enterprises and ensure strong AI Governance, ethics, and Security.

Infosys AI3S Suite of Responsible AI Offerings

Scan · Shield · Steer - helping enterprises scope out, secure and spearhead their AI investments, while adopting AI responsibly

SCAN

Identify the overall risk posture, legal obligations, vulnerabilities and threats arising due to AI adoption.

- Responsible AI Watchtower
- Responsible AI Maturity, Risk Assessment and Audits
- Infosys Responsible AI Control Center
- Regulation Readiness Consulting

SHIELD

Technical and specialized solutions, guardrails and accelerators for protecting AI models from vulnerabilities.

- Infosys Gen AI Guard Rails
- Infosys Responsible AI Toolkit
- Infosys AI Model Security
- Infosys Responsible AI Gateway

STEER

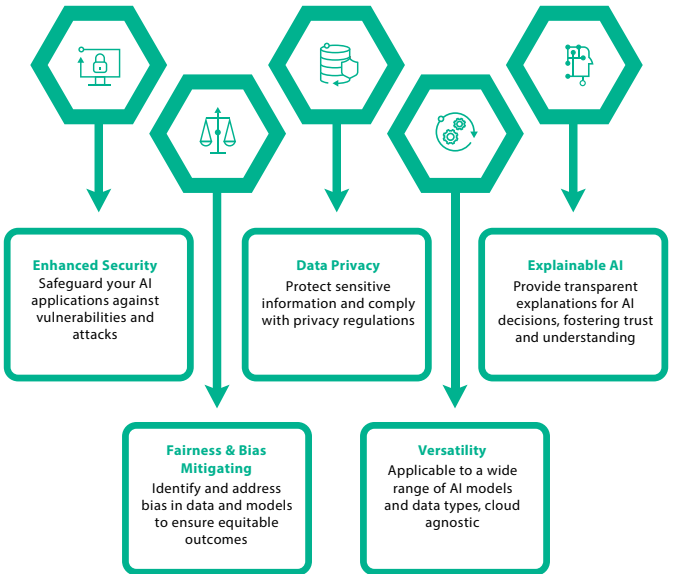
Advisory and consulting services to enable clients to advance their RAI journey and become leaders in the space.

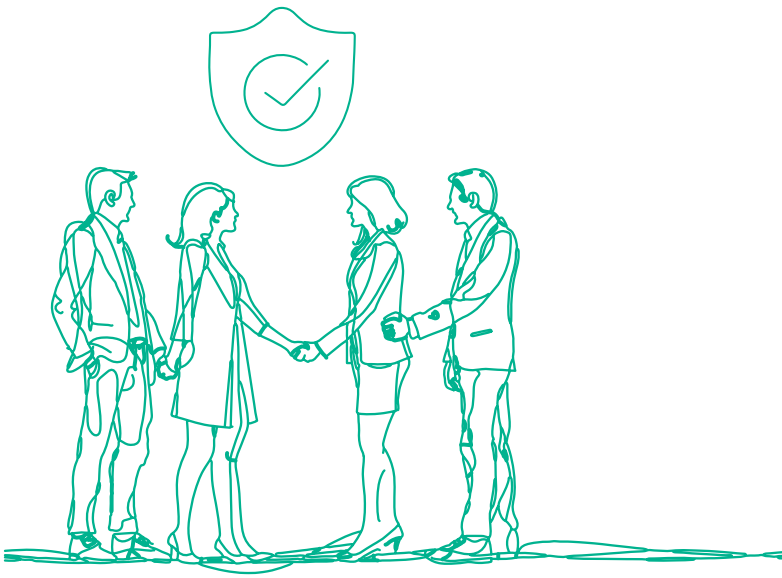
- Responsible AI Strategic Advisory Services
- Responsible AI Practice Setup
- AI Crisis Management

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

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

Key Features





Core Technical Features

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

02 **Privacy APIs**
PII Detection, Masking & Anonymization (Text, Image, DICOM & Multiple document types: PDF, DOCX, PPTX, XLSX, CSV, JSON)

03 **Explainability APIs**
Feature Importance | Chain of Thoughts | Thread of Thoughts | Graph of Thoughts | Logic of Thought (LoT)

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

05 **Fairness & Bias APIs**
Group Fairness | Image Bias Detection | Stereotype Analysis | Continuous fairness auditing with bias detection | Text Bias Detection, Fairness Monitoring

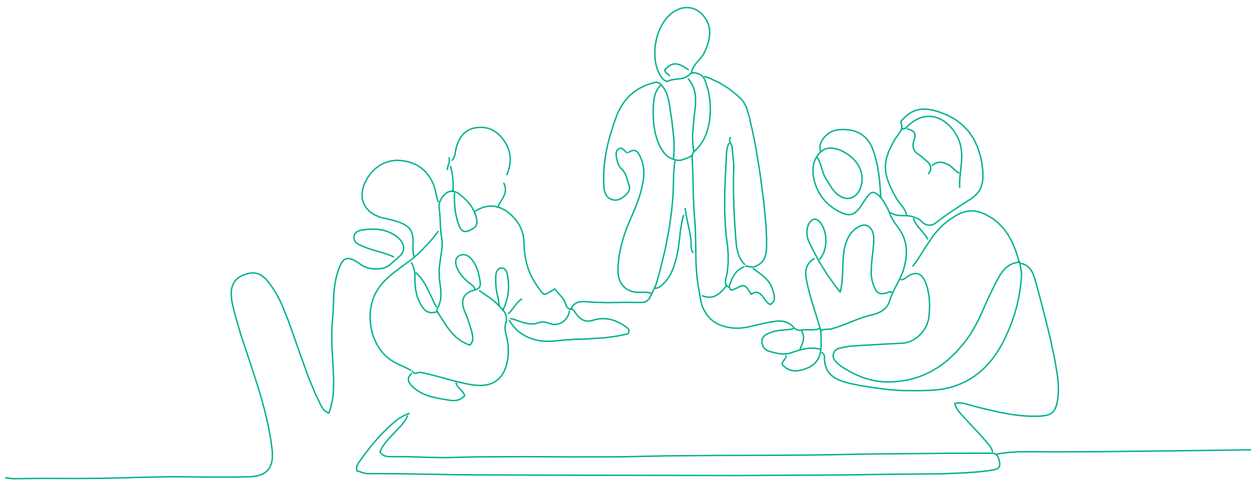
Upcoming Features:

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

- Telemetry: Data stored with PII anonymization in Telemetry, expanded to cover 30+ entities Signature and face masking in Privacy module.
- Privacy check in Moderation: Improved accuracy of PII recognizers and NER detection models, along with optimization of the overall solution.
- Bulk document safety validation.
- Moderation layer: Further latency and accuracy optimizations - finetuning existing models as well as finetuning a single model for all checks to provide a generic solution.
- Enhancing entity detection accuracy along with optimizing the Privacy Check mechanism within the ML pipeline, ensuring improved performance and better alignment with use case requirements.

Toolkit Accessible Across Platform:

GitHub	https://github.com/Infosys/Infosys-Responsible-AI-Toolkit
AI Kosh	https://aikosh.indiaai.gov.in/home/toolkit/ai_guardrails
OECD Policy Observatory	https://oecd.ai/en/catalogue/tools/infosys-responsible-ai-toolkit
Hugging Face	https://huggingface.co/InfosysEnterprise/spaces
Salus	https://project-salus.org



Thought Leadership: Research Paper Publications

How LLM Guardrails Safeguard Your Enterprise AI Journey

Enterprise AI systems need more than default model safety controls. This reference architecture proposes a two-stage guardrail pipeline that screens every prompt before it reaches the AI and every response before it reaches the user, combining template-based and model-based controls to prevent data leaks, policy violations and harmful outputs at scale.[54]

Algorithmic red teaming approaches to secure LLMs

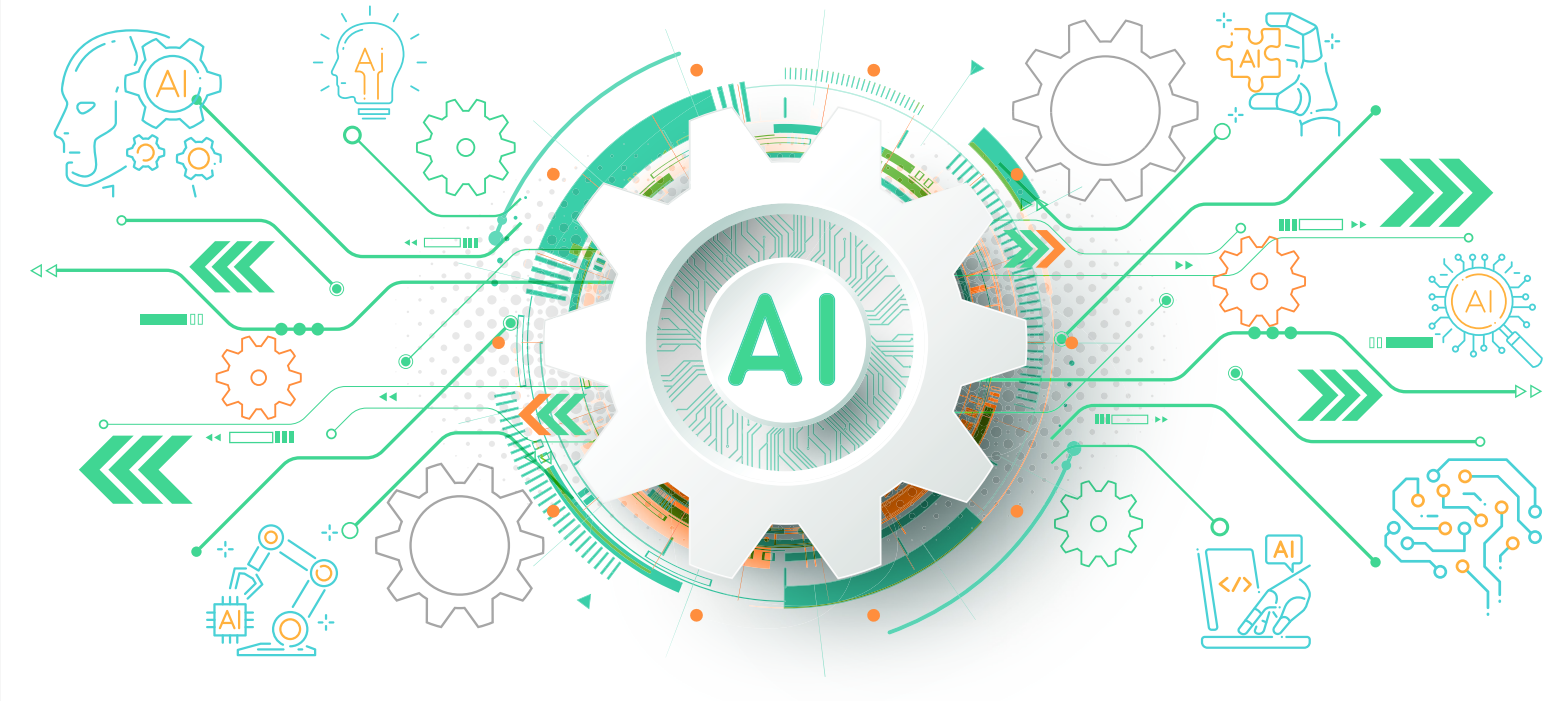
Describes how large language models are tested using automated adversarial methods with attacker, target, and judge models. Highlights techniques like prompt injection and jailbreaking, challenges with unreliable evaluation, and stresses continuous testing to ensure safe deployment and reduce risks.[55]

VISTA: Visualization of Token Attribution via Efficient Analysis

A lightweight, model agnostic technique that reveals which words matter most to an AI model's response, without extra computational cost. Using a perturbation-based approach with three analytical matrices, it advances explainability across any generative AI system.[56]

BELL: Benchmarking the Explainability of Large Language Models.

A standardized benchmarking framework that evaluates how well large language models explain their reasoning using diverse thought-eliciting techniques like Chain-of-Thought and Graph-of-Thought. Measuring quality through metrics like coherence, hallucination, and uncertainty, it helps identify more transparent and trustworthy AI models.[57]



[54] https://www.infosys.com/iki/perspectives/llm-guardrails-safeguard-enterprise-ai-journey.html
[55] https://www.sciencedirect.com/science/article/pii/S2666827025001987?via%3Dihub
[56] https://arxiv.org/abs/2604.02217
[57] https://arxiv.org/abs/2504.18572

Achievement

Celebrating a Global Recognition: Infosys AI Assurance Platform Earns Stevie® Award

The Infosys AI Assurance Platform has been recognized with a prestigious Stevie® Award, validating our leadership in AI Assurance, Governance and Trust. The recognition highlights the strong collaboration between the Infosys Quality Engineering team and the Responsible AI Office in embedding Responsible AI capabilities into the platform, helping enterprises adopt AI responsibly and at scale. As we continue to expand client adoption and strengthen our ecosystem partnerships, this achievement further reinforces Infosys' position as a trusted leader in enterprise AI assurance.[58]

News

Infosys Joins NVIDIA's Open Secure AI Alliance as an Inaugural Partner

Infosys has been named an inaugural partner of the Open Secure AI Alliance, a coalition launched by NVIDIA with over 120 organizations, including Microsoft, Cisco, IBM, Red Hat, Hugging Face, and the Linux Foundation, to build open-source tools, security frameworks, and defenses for AI systems and autonomous agents. The Alliance addresses security needs beyond the model layer, spanning agent harnesses, runtime controls, identity and access management, and evaluation frameworks for agentic AI. Infosys sees this as an opportunity to bring its enterprise experience to help shape practical, scalable security standards for the next generation of open, agentic AI.[59]

[58] https://www.infosys.com/services/quality-engineering/recognitions/ai-assurance-platform-wins.html

[59] https://blogs.nvidia.com/blog/open-secure-ai-alliance/

Contributors

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

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

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

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

Pritesh Korde - Consultant, Infosys Responsible AI Office, India

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

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

If you're interested in learning more about Responsible AI or exploring our Responsible AI offerings, feel free to reach out to us at **responsibleai@infosys.com**

CARRYING TRUST, EVERY STEP OF THE WAY. INFOSYS RESPONSIBLE AI

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

For more information, contact askus@infosys.com

Infosys®
Navigate your next