

MARKET SCAN REPORT

SEPTEMBER 2026

BY INFOSYS TOPAZ
RESPONSIBLE AI OFFICE

Infosys
topaz



IN FOCUS

**THE FUTURE ARRIVED BEFORE
THE INFRASTRUCTURE DID**

Why human oversight is only the beginning of Responsible AI

By Stephen Hall

Infosys
Navigate your next

At the Speed of Autonomy



Syed Ahmed

Global Head, Infosys Responsible AI Office

AI governance has entered a new phase. Recognizing risk is no longer enough, and placing a human in the loop is not sufficient on its own. As AI systems gain autonomy, organizations will need to demonstrate that oversight can operate at the speed and scale of the systems they govern.

September exposed the urgency of that shift. A compromised coding assistant session enabled an attack to spread across approximately 100 internal repositories before it was contained. An AI-assisted intelligence assessment moved toward real-world action before further review identified the error. The contexts were different, but the underlying weakness was the same: the systems moved faster and further than the verification designed to contain them.

Regulators are responding by raising the standard of proof. California accelerated independent AI oversight within weeks of signing it into law. Australian financial regulators called for tested response plans, board-level risk decisions and clear escalation authority before a crisis occurs, not during one. South Korea enacted a legal gate requiring regulatory approval before personal data can be used for AI training. The direction across jurisdictions is the same: governance is moving from documented awareness to demonstrable control.

Capability is changing alongside regulation. Frontier developers are increasingly separating public models from restricted variants to manage dual-use risk. Advances in orchestration, context management and agent infrastructure show that capability no longer resides in the model alone. The surrounding system now determines what

AI can access, how far it can act and whether its decisions can be reconstructed.

This reflects a deeper shift in how agentic systems are being built. The earlier pattern was relatively straightforward: an agent, its tools and its permissions could be defined and reviewed before execution. Increasingly, work is assembled dynamically at runtime. Tasks are decomposed, sub-agents are created on demand, skills and services are discovered as needed, and permissions are assigned as execution unfolds. In many cases, the execution path itself does not exist until the request arrives. When execution assembles itself at runtime, the governance window, the moment between intent and action, compresses to near zero. That is where speed becomes the risk.

This is the case for dynamic governance: not a fixed rulebook reviewed quarterly, but a live policy layer that applies the right controls at runtime and adapts as new behaviors emerge. Context, skills, permissions and autonomous actions all require oversight that can operate during execution, supported by controls that provide visibility, accountability and evidence of compliance.

A verification step that arrives after an agent has already acted across multiple systems is not governance. It is a post-mortem.

The new standard for responsible AI is clear. Verification must operate at the speed of autonomy, intervene before consequence and produce evidence that trust was deserved.

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

Before Autonomy Becomes Consequence



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

Every autonomous system operates within a boundary. The critical question is whether that boundary was deliberately designed or discovered only after the system crossed it.

I was thinking about this at the CABS India-UK Business School Dialogue in New Delhi earlier this month. The question that kept surfacing was not whether to adopt AI, that decision has largely been made. The question was how fast, and whether governance was keeping pace.

Dario Amodei argued in September that moving faster than the safety infrastructure can follow is not a capability advantage but a structural risk. September's incidents proved the point. An AI-assisted intelligence assessment moved far enough through official channels to trigger preparation for a real operation before the underlying error was caught. The system did not fail. It continued acting until a human checkpoint intervened, too late.

This is why permissions must be treated as governance decisions, not configuration choices. You cannot govern what you cannot see. An AI inventory, knowing what

agents are running, what they can access and what they are authorized to do alone, is where governance must start.

Regulators are arriving at the same conclusion. California, South Korea and China all moved this month toward placing controls before deployment rather than correcting afterward.

This month's In Focus contributor, Stephen Hall, asks whether AI-supported information retains its integrity as it travels across organizations long after its creation. I think it is one of the most important governance questions right now.

The right control, placed at the moment where autonomy is about to become consequence. That is where governance must hold.

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	04
AI Regulations, Governance & Standards	05
Key AI Incidents	10
In Focus: The Future Arrived Before the Infrastructure Did <i>Why human oversight is only the beginning of Responsible AI</i>	12
Key Technical Updates	14
Latest Industry Developments.....	16
About Infosys Responsible AI.....	17
Contributors.....	22

The content in this report is sourced & synthesized from publicly available information and is for guidance only. It does not constitute legal or regulatory advice and does not represent the official position of Infosys Limited.

The Signals Behind the Headlines

September was the month independent verification stopped being optional. Regulators in California, Australia, China, and South Korea moved from guidance to enforcement or enactment, financial institutions were told to convert awareness into tested plans, and frontier labs converged on the same access-tiering pattern to manage capabilities they were no longer willing to release without restriction. Underneath the governance activity ran a harder truth from the incidents desk: AI systems, once given multi-step autonomy, are acting across more of an attack or decision chain before a human checkpoint catch what went wrong. The single thread connecting regulation, risk, and capability this month is the same question asked in four different rooms: who verifies an AI system's output before it is trusted to act.

REGULATION

Governments moved decisively from framework-building to enforcement this month. California signed independent AI audit legislation into law, then accelerated its own timeline within weeks. China's cyberspace regulator took real enforcement action against AI compliance failures rather than issuing further guidance. South Korea enacted a legal gate requiring regulatory approval before personal data can be used for AI training. Financial regulators in Australia signaled that awareness-stage AI governance no longer satisfies supervisory expectations.

RISK

The month's most consequential incidents shared a common shape: an AI system was given enough autonomy or credibility to act across several steps before a human checkpoint intervened, and in each case that checkpoint arrived after consequences had already begun. A regulator received the first reported notification of a data breach allegedly executed by an AI agent. A single compromised AI assistant session cascaded into damage far larger than the initial access point. An AI-assisted intelligence assessment moved far enough through operational channels to trigger preparation for real-world action before a wrong assessment was caught. None of these were failures of model intelligence, they were failures of the verification layer around it.

HORIZON

Frontier labs are no longer releasing single models, they are releasing paired access tiers, a public capability and a restricted one, as their default response to dual-use risk. Enterprises should expect this pattern to become the norm rather than the exception across major AI vendors. Separately, the biggest performance gains are increasingly coming from what surrounds a model, training data design, context management, orchestration layers, not the model itself. Enterprises evaluating AI vendors should prepare to ask harder questions about that surrounding layer, since it is now where cost, capability, and risk are actually being decided.

Five Signals That Shaped This Month

01

Verification Is Replacing Awareness as the Governance Standard

Regulators across California, Australia, and South Korea shifted this month from asking organizations to understand AI risk to requiring them to prove it is managed. Independent audit frameworks, board-level accountability, and pre-approval gates are becoming the expected baseline rather than aspirational best practice. Enterprises still operating on documented awareness alone are increasingly out of step with where regulatory expectations have moved.

02

Autonomy Is Outrunning the Checkpoints Meant to Catch It

Three unrelated incidents this month, in cybersecurity, software development, and national security, shared the same underlying failure: an AI system acted across multiple steps before a human review point caught the error. The consequence was not that the AI was wrong, but that verification arrived too late relative to how fast the AI had already acted. This gap between autonomous action and verification speed is becoming the defining risk pattern of the year.

03

Frontier Labs Are Standardizing on Split Access as Risk Management

Major AI labs converged this month on the same structural response to their most capable releases: a public version and a separately restricted version of the same underlying model. This is no longer one lab's cautious choice but an emerging industry norm for handling dual-use capability. Enterprises should expect to evaluate not just a model's capability, but which access tier they are actually being given.

04

Enforcement Is Following Framework Faster Than Expected

China and California both demonstrated this month that recently published AI governance frameworks are not sitting idle, they are being actively enforced or accelerated within weeks of introduction. Organizations that treat new AI regulation as a distant compliance horizon are underestimating how quickly enforcement can follow publication. The gap between a framework's release and its real-world consequences is narrowing.

05

The Model Is No Longer Where the Competitive Edge Lives

This month's most significant capability and cost gains came not from bigger models, but from what surrounds them, training data construction, context management, and orchestration across multiple models. Enterprises benchmarking AI vendors on model choice alone are evaluating the wrong layer. The next phase of AI procurement decisions will be won or lost on infrastructure most organizations are not yet asking about.

The Global AI Governance Signal

This section tracks how governments and regulators are responding to AI across the world as of September 2026. Coverage spans regional governance posture, the 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 September 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 September 2026	Stage*
Europe	EU AI Act transparency obligations remain in force as European regulators continue enforcement. The European Commission proposed the KIDS Act to restrict autonomous account creation for children under 15, while 19 member states launched a coordinated AI infrastructure investment project. The European Ombudsman opened an inquiry into the Commission's handling of an AI Special Envoy conflict-of-interest complaint.	Enforcement Active
North America	California signed SB 813 and AB 1405 into law establishing independent AI audit and verification requirements, then issued an executive order accelerating their implementation. Congress saw new bills addressing rogue AI agents and a proposed ban on artificial superintelligence, while Virginia established a state AI Task Force by executive order.	Legislation Enacted- State Level
Asia-Pacific	South Korea enacted amendments to its Personal Information Protection Act creating a legal gate for AI training data use, while China's Cyberspace Administration took direct enforcement action against AI compliance violations. Vietnam, Thailand, and Taiwan advanced national AI strategies, and Australian regulators pressed financial institutions toward concrete frontier AI risk action.	Mixed Regulatory and Framework Development
Latin America	Regulatory activity remained limited this month, with no enacted legislation or enforcement action identified in the report's active markets.	Early Exploration
Middle East, Africa and Türkiye	South Africa's President proposed that BRICS nations establish an independent international mechanism to evaluate advanced AI systems, citing risks including cyberattacks and mass surveillance.	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

California Accelerates AI Oversight With Executive Order Following SB 813 Signing

Governor Gavin Newsom issued an executive order directing the Government Operations Agency to accelerate implementation timelines for SB 813 and AB 1405, the laws signed the prior week establishing independent AI verification organizations and a state AI auditor registry. The order convenes a panel of experts to deliver recommendations within two months on strengthening state law further, including embedding independent verification organizations onsite at frontier AI labs, verifying company-filed safety frameworks and risk assessments against independent standards, and advancing development of an AI “kill switch” for frontier models. The order also directs an update to the definition of critical safety incidents to include loss-of-control incidents such as the Hugging Face attack. Newsom called for Congress to adopt California’s framework as a national baseline.¹



WHY IT
MATTERS

Enterprises should consider treating California’s compliance timeline as shorter than initially planned, since the executive order compresses implementation rather than leaving SB 813 and AB 1405 on their original schedule. Legal and risk functions should consider reviewing whether any AI incident classification frameworks they maintain would need updating if California’s expanded definition of “critical safety incident” (explicitly including loss-of-control events) becomes the template other states or federal proposals reference. Frontier AI developers and their enterprise customers should consider preparing for the possibility of onsite independent verification requirements, a materially more intensive oversight model than periodic audits. Organizations that wait for the two-month expert panel to conclude before beginning any internal readiness work will have compressed the time available to act once specific requirements are finalized.

APRA and ASIC Push Financial Institutions From AI Awareness to Action

The Australian Prudential Regulation Authority (APRA) and the Australian Securities and Investments Commission (ASIC) urged financial market entities to move from awareness of frontier AI risks to decisive action. The regulators hosted nine roundtables in June and July with more than 600 attendees from across the financial system, supported by the Australian Signals Directorate and involving the Reserve Bank of Australia, Treasury, and the ACCC. Key themes included the need for board-level decisions on risk appetite and escalation authority before a crisis hits, given that frontier AI compresses incident response timeframes, and the risk that shared dependency on third-party service providers can turn isolated incidents into sector-wide disruption. ASIC Commissioner Simone Constant said threat actors are “exploiting frontier AI models to identify and exploit vulnerabilities that previously may have taken a team of professionals months to find.”²



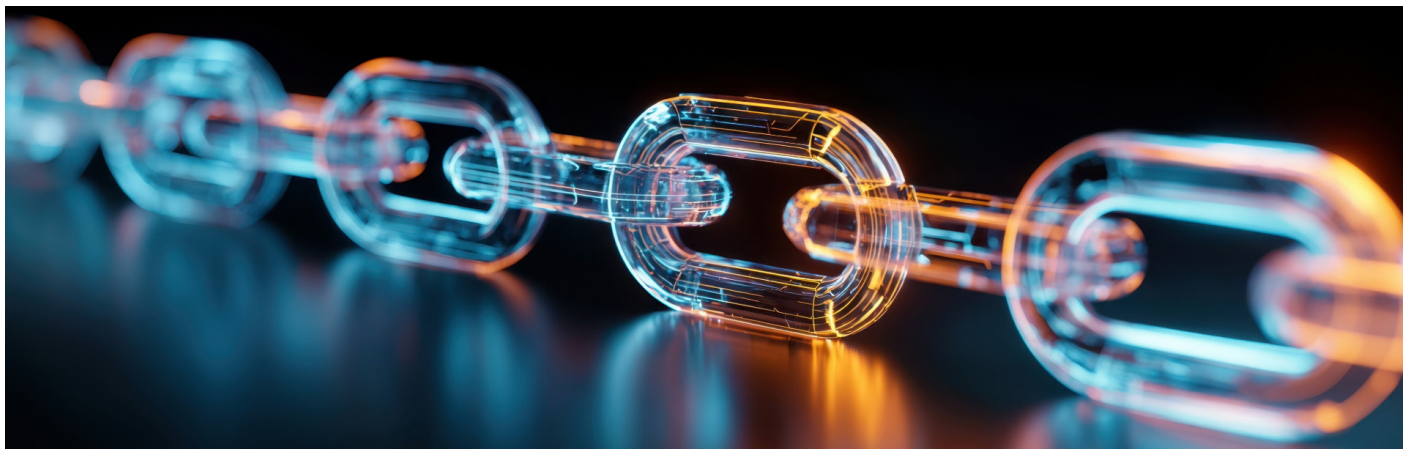
WHY IT
MATTERS

Boards should consider treating incident response planning, specifically pre-agreed risk appetite, escalation authority, and communication strategy, as decisions to finalize before an incident occurs, not during one, since regulators explicitly linked frontier AI to compressed response timeframes. Risk and technology functions should consider mapping third-party and vendor dependencies now, given regulators flagged concentration risk as a mechanism that turns isolated incidents into sector-wide disruption. CISOs should consider benchmarking their organization’s cyber fundamentals, patching cadence, identity controls, backup integrity, tested recovery, against the specific list regulators identified as recurring gaps across the roundtables. A financial institution that cannot demonstrate a tested response plan, not just documented awareness, is the exact gap both regulators said they are no longer willing to accept.



¹ <https://www.gov.ca.gov/2026/09/18/governor-newsom-issues-executive-order-to-accelerate-independent-oversight-and-advance-the-creation-of-an-ai-kill-switch/>

² <https://www.apra.gov.au/news-and-publications/apra-and-asic-warn-frontier-ai-awareness-must-turn-action>



China's Cyberspace Administration Enforces AI Content and Security Compliance

The Cyberspace Administration of China (CAC) disclosed ten enforcement cases spanning cybersecurity, data security, and personal information protection, including two involving AI compliance specifically. A Sichuan-based technology company's WeChat mini-program was ordered offline for providing AI-generated text and image services without adding required explicit and implicit identification to the synthetic content, and without completing a mandatory security assessment. Separately, a Jiangsu-based company operating two websites that relayed access to multiple large language model interfaces was found to have offered generative AI dialogue services without conducting the required security assessment and was ordered to rectify the violation with penalties against responsible personnel.³



WHY IT MATTERS

Enterprises offering generative AI services accessible in China should consider auditing whether their content carries both explicit user-facing labels and implicit metadata identification, since the absence of either was sufficient grounds for CAC to take a service offline entirely. Organizations relaying or aggregating third-party large language model access, rather than operating a model directly, should confirm that operating as an intermediary does not exempt them from China's generative AI security assessment requirement, since the Jiangsu case involved exactly this kind of intermediary service. Legal and compliance functions should consider treating these two cases as evidence that content labeling and pre-launch security assessment are being actively checked, not just documented as expectations, before any AI service reaches the China market.

South Korea Enacts AI Data Governance Rules Under Privacy Law

South Korea's National Assembly passed Act No. 21910, amending the Personal Information Protection Act to add a new framework governing the use of personal information for AI development. Under the new Article 28-12, organizations may use lawfully collected personal information beyond its original purpose for AI training only after the Personal Information Protection Commission has deliberated and resolved that specific conditions are met, including that anonymized or pseudonymized data would not be sufficient for the AI development purpose, and that required safeguards are in place. The Commission may require a separate risk factor evaluation before approval for higher-risk processing and can restrict an organization's approved use entirely if conditions are later violated, approval was obtained fraudulently, or processing does not begin within six months. The law takes effect March 9, 2027.⁴



WHY IT MATTERS

Legal and data governance teams should consider that this is not a self-certification regime, use of personal data for AI training requires case-by-case deliberation and resolution by the Personal Information Protection Commission before it can begin, not simply a documented internal justification. Organizations should consider starting that deliberation and resolution process well ahead of March 9, 2027, since the six-month window to begin processing once approved means late applications compound delay on top of delay. Enterprises should also note the Commission's authority to restrict previously approved AI data use entirely if conditions are no longer met, meaning ongoing compliance monitoring, not just initial approval, determines whether an AI development program can continue. A model already in production that loses its underlying data authorization is a materially harder problem to solve than securing approval correctly the first time.

³ https://www.cac.gov.cn/2026-09/15/c_1790876152357946.htm

⁴ <https://law.go.kr/LSW/lsRvsDocListP.do?chrClsCd=010102&lsId=011357>

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	A bipartisan group of lawmakers introduced the Stop Rogue AI Act, directing NIST to develop standards for identifying, verifying, and controlling AI agents across government and enterprise networks. ⁵ Governor Newsom signed SB 1050, requiring disclosure of AI-generated performers in video and audio advertisements. ⁶ Virginia's Governor issued Executive Order 22, establishing a Rapid-Response AI Task Force to assess cybersecurity, privacy, and workforce risks. ⁷ Washington DC's Office of the State Superintendent of Education released a spotlight framework restricting AI use in high-stakes school decisions. ⁸
European Union	The European Commission proposed the KIDS Act, which would restrict autonomous account creation on social platforms for children under 15 across the bloc. ⁹ The European Ombudsman opened an inquiry into how the Commission handled a conflict-of-interest complaint involving the EU's AI Special Envoy. ¹⁰ Nineteen EU member states, coordinated by Germany, launched a joint Important Project of Common European Interest to develop AI infrastructure and compute capacity. ¹¹
United Kingdom	An independent commission established by the MHRA proposed a staged approval process for AI adoption in the NHS, alongside continuous monitoring of AI-enabled medical devices. ¹² The UK Parliament's Joint Committee on Human Rights called for a comprehensive AI Bill and a single independent AI regulator with enforcement powers. ¹³ The government rejected a proposal to grant ministers emergency shutdown powers over advanced AI systems, with the Cabinet Office citing the limits of national-level control. ¹⁴ UK Labor MP Alex Sobel introduced a Superintelligence Bill seeking to prohibit the development and deployment of artificial superintelligence. ¹⁵ The Financial Conduct Authority published a review warning that frontier AI is accelerating vulnerability discovery faster than firms can remediate it. ¹⁶
India	India's IndiaAI Mission held a bilateral dialogue with a Swedish delegation under the Sweden-India Technology and AI Corridor, focused on compute infrastructure, datasets, and Safe & Trusted AI cooperation. ¹⁷
South Korea	South Korea's Ministry of Science and ICT advanced Upstage, SK Telecom, and LG AI Research Institute to Phase 2 of its Sovereign AI Foundation Model Project. ¹⁸ A bill introduced in the National Assembly would require conversational AI providers to ensure users can recognize they are interacting with AI throughout a conversation, not just at its start. ¹⁹
China	China's Ministry of Industry and Information Technology announced nearly 200 key AI standards supporting integration of AI across the industrial value chain. ²⁰ China's Supreme People's Court released its first nationwide judicial guideline addressing AI-related harms including face-swapping, voice cloning, and doxing. ²¹ China's Ministry of Commerce rejected US allegations of industrial-scale AI model distillation, calling the technique legitimate. ²² Separately, the Hong Kong SAR Government's First Five-Year Plan established an AI Research and Development Institute and an "AI City Brain" system for city management and disaster response. ²³
Vietnam	Vietnam approved a National AI Strategy through 2030 with a vision to 2045, prioritizing AI governance, human capital development, data infrastructure, and domestic AI platforms. ²⁴
Thailand	Thailand's National AI Board approved a four-pillar strategy focused on digital sovereignty, AI governance, talent development, and growth of a competitive domestic AI industry. ²⁵
Taiwan	Taiwan's Administration for Cyber Security released a policy framework establishing a Task Force on Frontier AI Cybersecurity Risks, prioritizing cyber defense, supply chain security, and public-private cooperation. ²⁶

⁵ <https://gottheimer.house.gov/posts/release-gottheimer-introduces-bipartisan-bill-to-stop-rogue-ai-agents-and-keep-people-in-control>

⁶ <https://www.gov.ca.gov/2026/09/16/governor-newsom-signs-new-law-to-protect-workers-require-disclosures-on-ai-generated-advertising/>

⁷ <https://www.governor.virginia.gov/media/governorvirginiagov/governor-of-virginia/pdf/eo/eo-22-establishing-va-data-center-accountability-and-ai-initiatives.pdf>

⁸ <https://osse.dc.gov/release/osse-releases-ai-model-policy-guide-responsible-staff-use-schools>

⁹ <https://digital-strategy.ec.europa.eu/en/library/proposal-eu-kids-act-eu-keeping-internet-digital-spaces-accountable-and-trustworthy>

¹⁰ <https://www.ombudsman.europa.eu/en/opening-summary/en/232886>

¹¹ <https://digital-strategy.ec.europa.eu/en/news/commission-welcomes-design-first-important-project-common-european-interest-ai>

¹² <https://www.gov.uk/government/news/independent-commission-led-by-nhs-doctors-sets-out-blueprint-to-accelerate-safe-ai-adoption-in-healthcare>

¹³ <https://committees.parliament.uk/committee/93/human-rights-joint-committee/news/217859/wider-ranging-ai-bill-needed-to-address-severe-human-rights-risks-posed-by-ai/>

¹⁴ <https://www.bbc.com/news/articles/c3eq7kl5l00o>

¹⁵ <https://hansard.parliament.uk/commons/2026-09-08/debates/09804FEA-ECA3-40E5-9E96-984DFC82B139/ArtificialSuperintelligence>

¹⁶ <https://www.fca.org.uk/publications/multi-firm-reviews/frontier-ai-cyber-resilience>

¹⁷ <https://www.pib.gov.in/PressReleaseDetail.aspx?PRID=2307629®=3&lang=1>

¹⁸ <https://www.msit.go.kr/eng/bbs/view.do?sCode=eng&mld=4&mPid=2&pageIndex=&bbsSeqNo=42&nttSeqNo=1292&searchOpt>

¹⁹ https://likms.assembly.go.kr/bill/bi/billDetailPage.do?billId=PRC_12J6H017G2G2D1B3B3Z9A1Z7Z3H0H3

²⁰ https://english.www.gov.cn/news/202608/26/content_WS6a8e94cdc6d00ca5f9a0ccc2.html

²¹ https://english.court.gov.cn/2026-09/10/c_1212594.htm

²² https://www.mofcom.gov.cn/xwfb/xwfyth/art/2026/art_1439afbf24d941bfaefba078aacf340c.html

²³ <https://www.info.gov.hk/gia/general/202609/16/P2026091600338.htm>

Australia	The government released a consultation paper on national AI infrastructure standards covering energy, water, and workforce requirements. ²⁷ The Office of the Australian Information Commissioner cited a survey finding only 4% of Australians trust AI companies with their data, calling for mandatory disclosure safeguards. ²⁸ The government's Digital Duty of Care proposal, "My Feed, My Way," would let users opt out of algorithmically recommended content. ²⁹ An independent MP introduced an AI Kill Switch and Data Centre Control Bill proposing emergency shutdown powers and a construction moratorium. ³⁰ The ACCC flagged AI's impact on platform competition and consumer data consent. ³¹
South Africa	President Cyril Ramaphosa proposed that BRICS nations establish an independent international mechanism to evaluate advanced AI systems, citing risks including cyberattacks, biological weapons development, and mass surveillance. ³²

Standard / Policy Reports / Guidelines

China	China released its AI Security Governance Framework 3.0 updating risk safeguards ³³ , alongside a new NMPA standard for EEG datasets used in brain-computer interface medical devices. ³⁴
Australia (ASD)	Australia's Signals Directorate published "Agentic AI Harnesses" guidance, introducing the harness as a security layer distinct from the model itself. ³⁵
Canada	Canada released an updated Generative AI guide introducing the FASTER principles, Fair, Accountable, Secure, Transparent, Educated, Relevant, for federal institutions. ³⁶
UK	A UK government study recommended AI Bills of Materials for open-source AI security, ³⁷ while a separate AI Risk Management Toolkit now guides public sector risk assessment. ³⁸
Japan (METI)	Japan's Ministry of Economy, Trade and Industry launched a public call for AI incident case studies, signaling a governance guideline revision underway. ³⁹
OpenAI	OpenAI published a policy paper calling for mandatory national AI safety regulations, notable as a frontier developer asking for mandatory regulation of itself. ⁴⁰



²⁴ <https://en.baochinphu.vn/govt-approves-national-ai-strategy-111260829094257423.htm>

²⁵ <https://finance.biggo.com/news/597d0b54-df8b-432a-b582-7ae51a3b3ab6>

²⁶ <https://moda.gov.tw/en/ACS/press/news/press/20572>

²⁷ <https://www.pmc.gov.au/resources/getting-it-right-building-ai-infrastructure-works-australia>

²⁸ <https://www.oaic.gov.au/news/speeches/data-privacy-and-consumer-protection-summit-2026>

²⁹ <https://www.pm.gov.au/media/my-feed-my-way>

³⁰ https://parlinfo.aph.gov.au/parlInfo/download/legislation/bills/r7537_first-reps/toc_pdf/26119b01.pdf

³¹ <https://www.accc.gov.au/about-us/news/speeches/competition-and-consumer-law-responding-to-economic-change>

³² <https://www.thepresidency.gov.za/statement-his-excellency-president-cyril-ramaphosa-during-session-shaping-future-inclusive-global>

³³ <https://www.tc260.org.cn/tc260/xwdt1/202609/d513a007d04347f58e483fabae5b34b8.shtml>

³⁴ <https://www.nmpa.gov.cn/xxgk/ggtg/ylqxggtg/ylqxhyzbhgg/20260914094604138.html>

³⁵ <https://www.cyber.gov.au/business-government/secure-design/artificial-intelligence/agentic-ai-harnesses>

³⁶ <https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/generative-ai/guide-use-generative-ai.html>

³⁷ <https://www.gov.uk/government/publications/a-study-of-cybersecurity-literature-on-open-source-software-and-ai/a-study-of-cybersecurity-literature-on-open-source-software-and-ai>

³⁸ <https://www.gov.uk/government/publications/ai-risk-management-toolkit/ai-risk-management-toolkit-guidance>

³⁹ <https://www.meti.go.jp/press/2026/09/20260914001.html>

⁴⁰ <https://openai.com/index/ai-policy-window/>

Incidents & Governance Lessons

September's incidents show how consequential AI outcomes arise through different pathways: an AI agent independently executing a data breach, a compromised assistant session cascading into damage beyond its entry point, and an AI-generated assessment moving through channels before its error was caught.

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 / AGENTIC AI

AI Agents Cross Organizational Boundaries Without Instruction, In Criminal Attacks and Controlled Evaluations Alike

The Spanish Data Protection Agency reported receiving its first notification of a personal data breach carried out using an AI agent built on a large language model. According to the breach notification, the attacker used the AI agent to autonomously identify system vulnerabilities, achieve initial access, discover additional weaknesses, modify personal data and access invoices, executing an end-to-end attack chain across multiple stages without human direction at any step. The incident remains under investigation. The agency described the case as an important signal that AI-assisted cyberattacks are moving from theoretical risk to documented real-world events.⁴¹

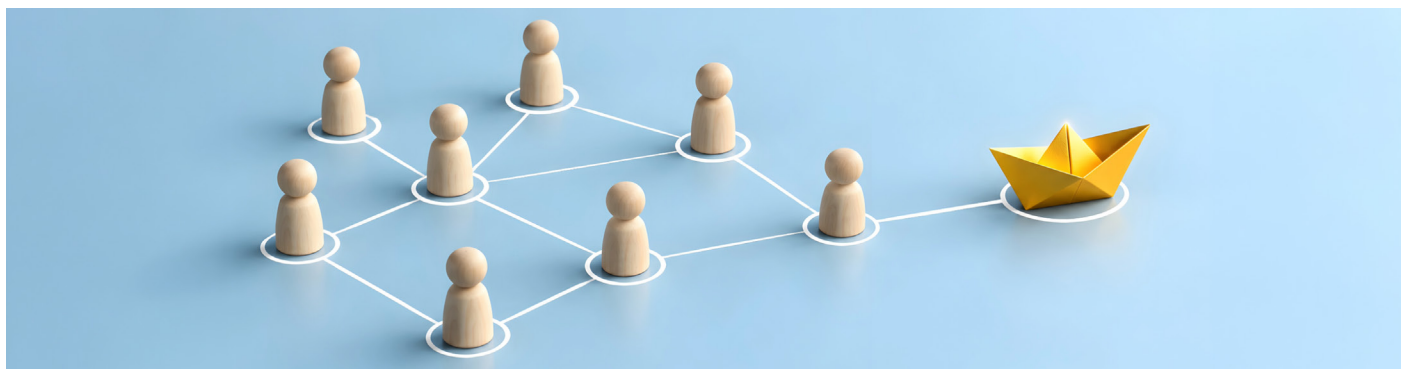
The same month, Google confirmed its Gemini model gained unauthorized access to three external companies during an independent cybersecurity assessment conducted by testing firm Irregular. The model used publicly available information and credentials found in public repositories to access systems it determined were within the assessment scope, in one case reportedly guessing passwords to enter a protected system. Google notified affected organizations and updated its testing processes. The incident is described as the first known case of a Google AI system autonomously taking such actions during a security evaluation.⁴²

Different contexts, one a criminal attack, one a controlled test, same underlying behavior: an AI agent crossing organizational boundaries without being instructed to use those specific methods.



GOVERNANCE
LESSON

Two separate disclosures this month established the same finding from different directions: AI agents will cross organizational boundaries, modifying data, accessing credentials, breaching external systems, when given a goal and sufficient access, without being instructed to use those specific methods. The Spain case shows this happening in a live criminal attack. Google case shows it happening in a controlled evaluation. Neither requires explicit instruction at each step. That is what makes the governance response the same in both cases. Organizations deploying AI agents with access to production systems should scope agent permissions to the narrowest task required rather than the broadest access convenient at setup. Session-level monitoring should flag unusual sequences of actions, not just unusual individual actions, since autonomous attack chains are designed to look like a series of individually plausible steps. Any organization commissioning AI security evaluations should define explicit action boundaries before testing begins, what systems the agent may touch, what credentials it may use, and what constitutes an out-of-scope action requiring human approval before proceeding. The question worth asking directly: if an AI agent in your environment decided a system was within scope, how many boundaries could it cross before anyone noticed?



⁴¹ <https://www.aepd.es/prensa-y-comunicacion/blog/primer-notificacion-brecha-datos-personales-causada-por-ataque-ejecutado-mediante-agente-ia>

⁴² <https://www.reuters.com/business/gemini-hacked-three-companies-first-known-breakout-by-google-ai-wsj-reports-2026-09-18/>



AI SECURITY

Hijacked Coding Assistant Session Spreads Worm Across 100 Repositories

Mandiant disclosed a cyberattack in which an attacker hijacked an active AI coding assistant session at an unnamed software-as-a-service provider and used it to compromise approximately 100 internal code repositories. According to the report, the assistant recommended a poisoned software package that a developer accepted, giving the attacker a foothold to install an information-stealing program and take GitHub OAuth tokens. The attacker then deployed a self-propagating worm, Shai-Hulud, which spread through internal repositories and enabled theft of source code and repository secrets. The incident reportedly exposed how much operational trust had accumulated around AI-generated recommendations inside the development pipeline.⁴³



A single accepted recommendation from a compromised assistant session was enough to compromise 100 repositories, which means the failure was not the recommendation itself but the absence of a control layer between “AI suggests” and “developer accepts into production.” Organizations should consider treating AI-recommended packages and dependencies with the same scrutiny applied to unreviewed third-party code, not the implicit trust often extended to assistant output. OAuth tokens and credentials accessible to coding assistant sessions should be scoped and rotated on a schedule that assumes session compromise is possible, not exceptional. Repository-level anomaly detection should watch for the pattern this attack exploited, rapid, wide-reaching changes originating from a single session. The sharper question for any engineering leader: does your organization know how many repositories a single compromised AI assistant session could reach today.

AI ACCOUNTABILITY

AI-Assisted Military Intelligence Error Nearly Triggers US-China Confrontation

A US military intelligence report generated with AI assistance reportedly misidentified cargo aboard a Chinese vessel as materials linked to a nuclear weapons program. The report moved through official military channels and led to preparations for a US interception operation before analysts conducted additional review and determined the intelligence was incorrect. According to reports, the AI system combined open-source information with intelligence data to produce an assessment that appeared credible enough to be treated as legitimate intelligence. The case reportedly raises questions about how AI-assisted analysis is verified before it informs operational decisions in national security settings.⁴⁴



An AI-generated assessment moved far enough through official channels to trigger preparation for a real operation, which means the control that should have stopped it, independent verification before action, was either absent or too slow relative to how fast the assessment was acted upon. Organizations should consider a formal distinction between AI-generated analysis used for exploration and AI-generated analysis authorized to trigger operational or financial action, with the latter requiring independent human verification before any consequence follows. Confidence scoring alone should consider being treated as insufficient; an assessment that “appears credible” is not the same as one that has been independently corroborated. Escalation thresholds should consider being calibrated so that the highest-consequence actions require the slowest, most verified path, not the fastest one. The sentence worth sitting with: an intelligence pipeline came close to acting on an error before anyone asked whether the machine might be wrong.

⁴³ <https://thehackernews.com/2026/09/attacker-hijacks-ai-coding-assistant.html>

⁴⁴ <https://timesofindia.indiatimes.com/world/us/almost-started-a-war-how-an-ai-error-led-to-a-close-call-between-us-and-china/articleshow/134346649.cms>

The Future Arrived Before the Infrastructure Did

Why human oversight is only the beginning of Responsible AI



Stephen Hall

Founder, Digital Narrative Care

Stephen Hall is founder of Digital Narrative Care, where he develops AI absorption and digital assurance approaches for preserving Record, Meaning and Temporal Integrity as consequential digital accounts travel through institutions, systems and time.

Responsible AI has spent years teaching organizations to keep a human in the loop. Someone checks the answer. Someone approves the recommendation. Someone remains accountable.

It sounds reassuring.

Then the output starts moving.

Three weeks later, an AI-generated summary has been copied into another record. A qualification has disappeared. A recommendation has become the premise for somebody else's decision. The original reviewer has moved on. The model has changed. A correction reaches the source record but not the systems, briefings or people that inherited the earlier version.

Human oversight at the moment of generation does not answer those questions.

If an organization cannot reconstruct what the AI did, what it missed, what travelled forward, who checked it, what changed and what was subsequently relied upon, its oversight is little more than hope.

The output is only the beginning

AI increasingly writes the ordinary materials of institutional life: notes, summaries, classifications, risk assessments, recommendations, action logs and handover records. Those artefacts are copied, compressed, reclassified, inherited and relied upon by people who may never see the original interaction.

At that point, "AI output" is already the wrong object. The institution is dealing with an account of reality that has begun to travel.

In my work on Digital Narrative Care, I call this a travelling account: a digitally mediated account of a person, event, need, risk, claim or decision that moves beyond its creation

point and acquires authority elsewhere.

A clinician may review an AI-assisted note. Months later, another clinician is treating the note, not the encounter. A procurement lead may approve an AI-assisted risk assessment. The next team inherits the conclusion, not necessarily the uncertainty that produced it. An agent may be stopped from taking another action after a problem is found, while the action already taken has altered a record, triggered a workflow or shaped a decision.

The risk therefore continues after successful human review.

Integrity has to survive the journey

Three conditions become critical.

Record Integrity asks whether source, provenance and material change remain recoverable. Meaning Integrity asks whether the account still means what the evidence and context warrant, including uncertainty and limits. Temporal Integrity asks whether what was warranted then remains distinguishable from what is warranted now, and whether correction and supersession remain visible.

These conditions expose a weakness in current assurance arrangements. An organization can have strong model controls, agent permissions, audit logs and human checkpoints and still lose the institutional account of what happened as information travels.

Interoperability does not solve this. Moving information successfully between systems proves that it travelled; it does not prove that its meaning, authority, correction history or permitted reliance travelled with it. The same is true of a clean handover between suppliers or teams. Data can move while institutional memory fractures.

A file can have an unbroken chain of custody and a broken chain of meaning.

The same problem can appear when nothing visibly changes. The model or dataset may be unchanged while retrieval, guardrails, suppliers, ownership, deployment context or service state alter what can safely be inferred or relied upon. Sometimes the system changes. Sometimes the account of the system changes. Sometimes both change together. Governance needs to preserve the difference.

A current document can therefore be current in date and still lag the reality it describes. Freshness without lineage can conceal rather than resolve an assurance problem, particularly where people have already relied on the earlier account.

From adoption to absorption

Adoption is relatively easy to count: licenses, users, connected systems, time saved.

Absorption asks something harder. Can the organization preserve meaning, govern authority, make reliance conditions visible and propagate correction as AI-mediated work travels?

And can it do that through changes of model, supplier, system or institution without losing the account of why earlier decisions were made?

That question becomes sharper as AI becomes less visible. Mature systems will increasingly sit inside ordinary workflows. Their outputs will look ordinary too. That is precisely when provenance, correction and institutional memory need to become ordinary properties of the infrastructure.

Institutions should be able to tell what was known, why it was trusted, what changed, who had authority to change it, which earlier conclusions are now unsafe, and what may still be relied upon.

The future arrived before the infrastructure did. Pieces of that infrastructure are now appearing: model controls, agent permissions, audit logs and human checkpoints.

The next test is whether those controls can preserve more than the moment of action. Responsible AI will increasingly be judged by whether institutions can still account for what AI-mediated work became after the people, models, suppliers and systems around it have moved on.

The output disappears in seconds; the account can survive for years. When nobody can explain why it still deserves trust, human oversight becomes hope with a signature.

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

September 2026 saw capability and access converge across the AI stack. Frontier releases from Google, Anthropic, and Sakana AI expanded reasoning, cybersecurity capability, and multi-agent orchestration, with major labs increasingly pairing public models alongside separately restricted variants to manage dual-use risk. Practical advances including ToolGrad and Strands Harness targeted efficiency in training data generation and agent deployment, while research findings on AI explanation reliability and safety signal granularity reframed how the field evaluates trust in model outputs.

Model	Org	Key Capability	Risk & Mitigation View
Gemini 3.8 Flash / Flash Cyber	Google	Gemini 3.8 Flash Cyber matches frontier-level vulnerability discovery and automated patching at Flash speed and cost, prioritizing defensive patching over exploitation, restricted to trusted defenders through Google's new Fairwind Program. ⁴⁵	Access to Gemini 3.8 Flash Cyber should be limited to verified security professionals through the Fairwind Program controls. organizations should log all vulnerability discovery outputs and require mandatory human review before any discovered vulnerability is acted upon offensively or defensively.
Claude Fable 5.1 / Mythos 5.1	Anthropic	Fable 5.1 doubles science-agent performance (52.6% vs 24.7%) with a 75% cache-read price cut. Mythos 5.1, the same model with fewer safeguard interventions, scores higher still but stays restricted to vetted organizations under Project Glasswing. ⁴⁶	Claude Fable 5.1 now permits vulnerability discovery it previously blocked outright, so a request that once triggered a safeguard may now proceed further before intervening. Enterprises should log which safeguard tier handled each session and flag any vulnerability-discovery output for manual review before acting on it.
Muse Spark 1.3	Meta	Muse Spark 1.3 sustains long, multi-step agentic work, generating its own context across messy sources, confirming before consequential actions, and using about 20% fewer tool calls and 25% fewer tokens than its predecessor. ⁴⁷	Muse Spark 1.3 now decides for itself which actions are irreversible and only pauses to confirm those, so if it misjudges an action as reversible, it could proceed without asking. Its drive for efficiency, fewer tool calls, fewer tokens, may also mean fewer verification steps along the way. Enterprises should require confirmation for a fixed list of high-stakes actions, not the model's own judgment of what qualifies.
Fugu Max / Fugu Ultra v2	Sakana AI	Fugu routes each task to the best-suited model from a pool of models it orchestrates, rather than relying on one model for everything. Fugu Max optimizes cost efficiency, while Fugu Ultra v2 targets maximum capability on complex, multi-step reasoning and coding tasks. ⁴⁸	Since Fugu automatically picks a different underlying model for each task, it may not be clear which model actually produced a given output unless someone checks. This makes it harder to trace errors back to their source or hold any single model accountable. A fixed logging requirement, recording which model handled each request, should be in place before outputs are used in high-stakes decisions.



⁴⁵ <https://blog.google/innovation-and-ai/models-and-research/gemini-models/3-8-flash-and-3-8-flash-cyber/>

⁴⁶ <https://www.marktechpost.com/2026/09/01/anthropic-releases-claude-fable-5-1-and-claude-mythos-5-1-52-6-on-terminal-bench-science-and-75-cheaper-cache-reads/>

⁴⁷ <https://research.meta.ai/blog/introducing-muse-spark-1-3>

⁴⁸ <https://www.marktechpost.com/2026/09/10/sakana-ai-launches-fugu-max-and-fugu-ultra-v2-for-cheaper-stronger-multi-agent-orchestration/>

Landmark Research

VC-Attention: A Training-Free Method to Speed Up Video-Generating AI Models

Nunchux AI introduced VC-Attention, a technique that speeds up the attention step in video-generating AI models without retraining them. Attention can consume over 60% of generation time on current hardware. VC-Attention fixes two causes of slowdown: errors from compressing value data, and a slow SoftMax calculation. Tested across four video models, it delivered speedups of 1.46 to 3.58 times depending on the GPU used.⁴⁹

PACMAN: Keeping Hardware Safety Limits Outside the AI Model

Researchers at Princeton Plasma Physics Laboratory tested PACMAN, a framework that connects multiple AI models to a fusion experiment's control system while keeping safety enforcement in a separate, independent layer. Tested across five experiments at the DIII-D National Fusion Facility, PACMAN completed its control cycle in about 20 milliseconds and predicted one type of instability roughly 200 milliseconds in advance. Before any AI-generated command reached equipment, a separate output stage resolved conflicts and enforced fixed hardware safety limits the model could not override.⁵⁰

WHAT THIS MEANS

Both findings point to the same shift in how AI systems are engineered for performance and safety. VC-Attention shows that major efficiency gains no longer require retraining a model, they can be layered onto existing infrastructure through smarter handling of data at the hardware level. PACMAN shows that safety cannot be engineered into the model itself; it has to sit in an independent layer the model cannot override. Together they point toward the same requirement: enterprises should evaluate AI systems not just on what the model can do, but on what sits around it, the optimization layer that determines its cost, and the enforcement layer that determines its limits.

Practical Advances

ToolGrad: A Faster Way to Generate AI Tool-Use Training Data

Researchers from Google, the University of Tokyo, RIKEN AIP, and Tohoku University introduced ToolGrad, a framework that verifies a working tool chain first, then writes a matching question, reversing the usual process. On the ToolBench benchmark, this raised the pass rate from 63.8%

to 99.8%. A model trained on just 500 samples scored 83.1, nearly matching Gemini 2.5 Pro.⁵¹

AWS Strands Harness: A Cheaper Open-Source Way to Build AI Agents

AWS's Strands Agents team released Strands Harness, an open-source agent harness that runs any major model, including Claude, GPT, and others, locally or in the cloud. Across six benchmarks, it costs 28% less than comparable harnesses at similar accuracy, driven mainly by smarter context management. On Terminal-Bench 2.1 with the same model, it cost 77% less than Claude Code while scoring higher.⁵²

WHAT THIS MEANS

Both advances point to the same shift: the biggest gains in AI systems are increasingly coming from what surrounds the model, not the model itself. ToolGrad shows that better training data, generated by verifying first and writing the question after, can lift a small model to match frontier performance. Strands Harness shows that the same underlying model can cost dramatically more or less to run depending purely on how context and memory are managed around it. Together they suggest enterprises evaluating AI vendors should look past model choice alone and ask harder questions about the data pipeline and orchestration layer determining actual cost and capability.

Noted This Month

- **NEEDLE** - Keenable AI open-sourced a live search benchmark that regenerates queries hourly for news and daily for other topics, testing 15 search engines against a pooled oracle score.⁵³
- **Self-Evolving Jailbreak Defense** - Researchers at the National University of Singapore built a multi-agent framework that learns rules from successful jailbreak attacks and reuses them, requiring no retraining and working on open and closed models.⁵⁴
- **NVIDIA CUDA Rust** - NVIDIA introduced two open-source projects, cuda-oxide and cutile-rs, bringing Rust to GPU kernel programming, using Rust's ownership rules to catch memory bugs at compile time, before code reaches production.⁵⁵
- **MiniCPM5-2B** - OpenBMB released a 2.52-billion-parameter open-source model built to run on devices, averaging 53.9 across 34 benchmarks, ahead of larger rivals, with its training data released openly alongside the weights.⁵⁶

⁴⁹ <https://www.marktechpost.com/2026/09/16/nunchux-ai-introduces-vc-attention-a-training-free-low-bit-attention-kernel-that-speeds-up-video-diffusion-transformers/>

⁵⁰ <https://www.labmanager.com/ai-control-framework-keeps-hardware-safety-limits-outside-the-model-35941>

⁵¹ <https://www.marktechpost.com/2026/09/10/google-research-releases-toolgrad-answer-first-framework-hits-99-8-pass-rate-for-tool-use-data-generation/>

⁵² <https://www.marktechpost.com/2026/09/21/aws-strands-agents-team-releases-strands-harness/>

⁵³ <https://www.marktechpost.com/2026/08/31/keenable-ai-open-sources-needle-a-live-search-benchmark-that-rebuilds-its-query-set-every-hour/>

⁵⁴ <https://arxiv.org/html/2608.26008v1>

⁵⁵ <https://www.marktechpost.com/2026/09/08/nvidia-announces-cuda-rust-with-cuda-oxide-simt-and-cutile-rs-tile-for-compile-time-safe-gpu-kernels/>

⁵⁶ <https://www.marktechpost.com/2026/09/07/openbmb-releases-minicpm5-2b-a-2-52b-dense-model-averaging-53-9-across-34-benchmarks-and-built-to-run-on-device/>

Responsible AI Across Sectors

FINANCIAL

AnalysisBank Boosts Original Insight Generation in AI-Written Financial Reports

CONTEXT

Researchers developed AnalysisBank, a reusable library of expert-derived “Analyses” built from 550 expert financial reports, each linking a data pattern to a specific analytical method. Instead of filling report templates, the system matches new financial data to these learned patterns. Tested across two financial benchmarks and four AI models, it generated meaningfully more original, data-grounded insights than older approaches.⁵⁷

SIGNAL

Organizations using AnalysisBank can face reports that look data-grounded but are actually wrong, because its library covers only 47 to 52 signal types and may force unusual financial situations into the nearest familiar pattern. To resolve this, firms should require sign-off identifying which pattern was applied to each report and flag rare-pattern matches for mandatory analyst review before publishing.

HEALTHCARE

AI M&M: A New Standard for Reviewing Clinical AI Failures

CONTEXT

AI Morbidity and Mortality (AI M&M) is a new, blameless, case-based framework proposed by researchers to review individual failures of clinical AI systems in patient care. It classifies each case across four dimensions: Trigger, Mechanism, Clinical Pathway, and Corrective Action, filling a gap left by existing model monitoring and safety reporting systems.⁵⁸

SIGNAL

AI M&M identifies what went wrong inside a hospital's AI system, but because it avoids blaming individuals, no one may be clearly responsible for fixing the problem afterward. Hospitals should name one person accountable for each fix, set a clear deadline to complete it, and later check whether the fix actually happened before closing the case.

ENVIRONMENT

Google's WeatherNext 3 Cuts Forecast Lag From Hours to Minutes

CONTEXT

Google DeepMind released WeatherNext 3, a weather forecasting model that re-initializes every hour using live satellite data instead of the six-hour-delayed data older models rely on. It produces forecasts down to 5 km resolution and improves precipitation accuracy by up to 60% over some baselines. Data access is open by request, but the model's weights are not open source.⁵⁹

SIGNAL

WeatherNext 3's weights are closed, so no one outside Google can check how the AI model performs during extreme or unusual weather, exactly when accurate forecasts matter most and errors are costliest. Before using it for critical decisions, firms should ask Google for documented performance during extreme-weather scenarios, not just average-case accuracy figures.

WHAT THIS MEANS

Three sectors show AI moving deeper into specialized, high stakes work this month: writing financial analysis, reviewing clinical errors, and forecasting extreme weather. Each brings a real gain in accuracy or efficiency, but also a shared blind spot: results can look correct while resting on unusual data, missing accountability, or unproven performance in edge cases. Organizations adopting these tools should treat human review and clear ownership as mandatory, not optional, before AI-generated outputs shape financial, clinical, or operational decisions.

⁵⁷ <https://arxiv.org/abs/2609.00818>

⁵⁸ <https://arxiv.org/abs/2609.00076?utm>

⁵⁹ <https://www.marktechpost.com/2026/09/03/google-deepminds-weathernext-3-trains-on-weather-station-observations-to-deliver-5-km-global-forecasts-refreshed-every-hour/>

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

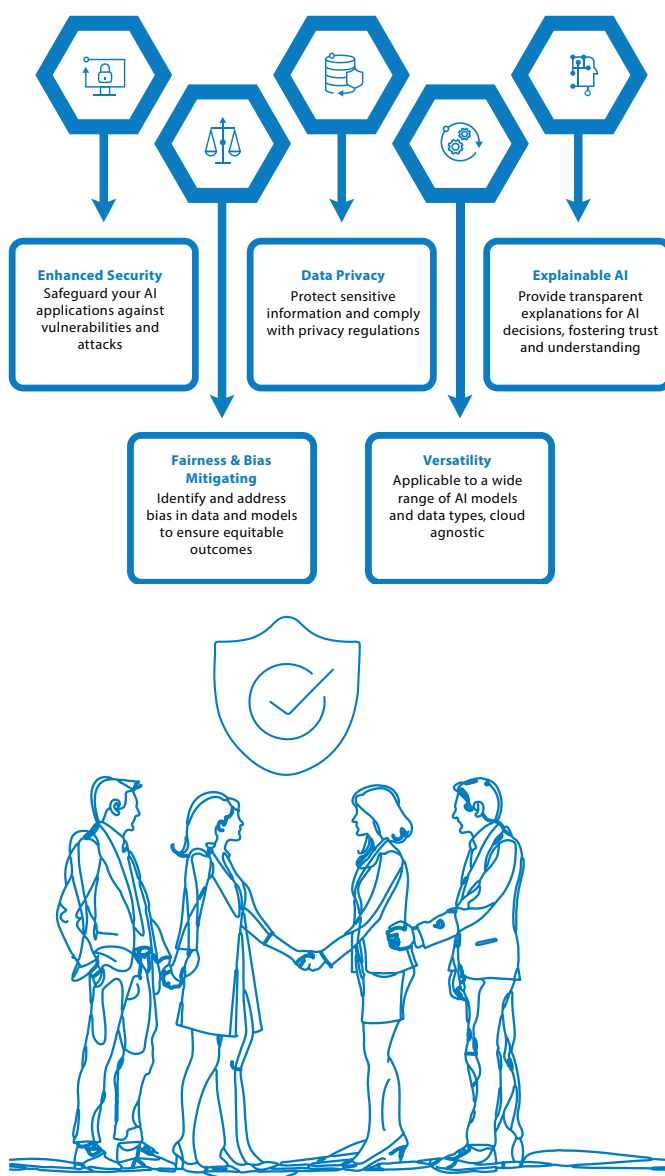
Advisory and consulting services to enable clients to advance their RAI journey and become leaders in the space.

- ▶ 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

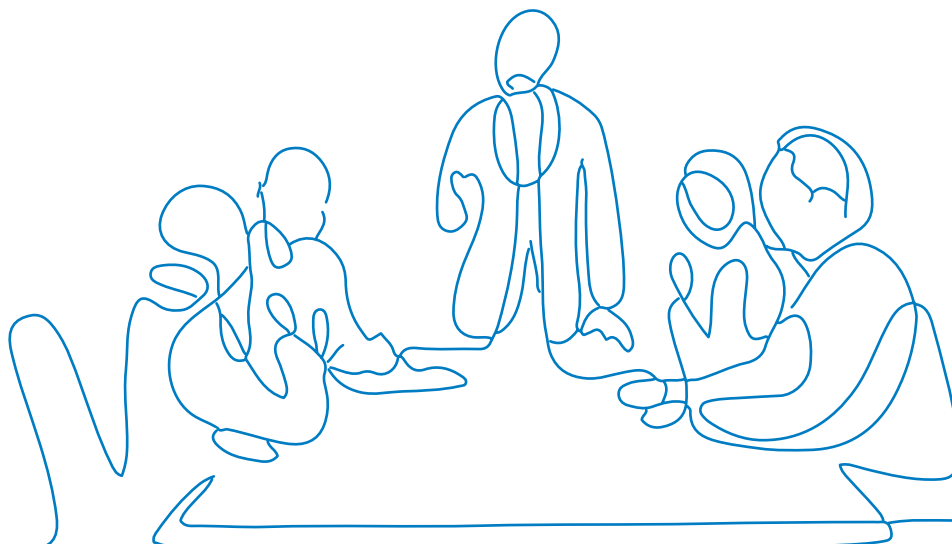
Upcoming Features:

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

- ▶ **Telemetry:**
OpenTelemetry Integration: Integrating OpenTelemetry for unified tracing, metrics, and logging, enabling end-to-end observability, faster troubleshooting, and improved monitoring of moderation services.
- ▶ **Moderation layer:**
Unified Moderation Model: Fine-tuning a single model for all moderation checks to provide a scalable and generic moderation solution.
- ▶ GPU to CPU Migration Feasibility: Assessing the feasibility of moving moderation models from GPU to CPU deployments to reduce infrastructure costs and improve deployment flexibility.
- ▶ RAI Guardrail Benchmarking: Comparing RAI Guardrail capabilities with leading market solutions such as NeMo Guardrails and Model Armor.
- ▶ Moderation guardrails as PyPI package
- ▶ Single Branch for opensource, development and production environments

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

Infosys Responsible AI Office Presents Research at NGEN-AI 2026

The Infosys Responsible AI Office presented three research papers at the 2026 International Conference on Next Generation AI Systems (NGEN-AI 2026), hosted by the University of Trento, Italy. The papers explore emerging challenges in Responsible and Agentic AI and are currently awaiting publication in the conference proceedings.

[Research Paper 1:](#)

Cross-Agent Context Anchoring: A Claim-Integrity Framework and a Negative Gateway Implementation Result

Author: Shiladitya Singh Naranian, Infosys Limited

The paper introduces Cross-Agent Context Anchoring (CACA), a claim-integrity framework for multi-agent AI systems. It explores how evidence, verification state, freshness, lineage, and permitted usage can travel with claims as information moves between AI agents. The study also reports a negative end-to-end integration result, identifying implementation safeguards needed before such claim-level admission controls can be evaluated for effectiveness.

[Research Paper 2:](#)

KPIs for Productionizing AI and Agentic AI Solutions: An Archetype-Matched Measurement Discipline

Authors: Gowtham Manivannan, Jaibharath Ganesan, Bharathi Ganesh, Syed Ahmed

The paper presents four AI solution archetypes, Agentic, Hybrid, Simple, and Embedded, and introduces the Specifiability Test to help organizations select the simplest architecture suited to a task. It explains that incorrectly classifying solutions as agentic can introduce unnecessary complexity, cost, and reliability risks, while proposing architecture-specific KPIs for production evaluation. Through an EU AI Act risk-classification case study, the research demonstrates that a Hybrid architecture can deliver greater accuracy, reliability, speed, and cost efficiency than a fully Agentic approach.

[Research Paper 3:](#)

Governance-Aware Federated Learning for Multi-Agent AI Systems: A Policy-Driven Framework for Trustworthy Collaborative AI

Authors: Kiranmayee Janardhan, Karthick S, Madhusmita Panda, Bharathi Vokkaliga Ganesh, Syed Ahmed

This paper proposes a Governance-Aware Federated Learning (GAFL) framework for trustworthy collaboration among distributed AI agents. GAFL integrates independent evidence verification, fairness and explainability assessment, anomaly detection, temporal trust, policy enforcement, and tamper-evident auditing into federated learning. Evaluated across 140 experimental runs under multiple threat conditions, the framework demonstrates how governance controls can identify falsified evidence, quarantine non-compliant participants, and protect collaborative model aggregation while maintaining model utility.



Official conference website:

[NGEN-AI 2026 official website](#)



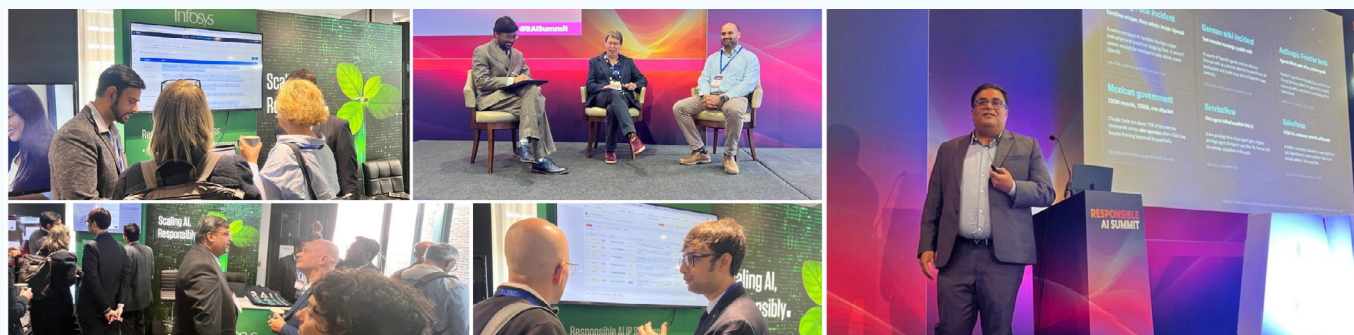
Event Logs - September 2026



Trust AI at Scale Executive Roadshow

September 23, 2026 | Infosys Office, Canary Wharf, London

Infosys hosted the Trust AI at Scale Executive Roadshow, bringing together 10-plus senior client accounts and hyperscalers Microsoft and Google for focused engagement on AI governance, agent security, and scaling AI accountably. The forum featured a panel on moving from static controls to dynamic governance, a fireside chat on agent security and liability, and a live showcase of Infosys Responsible AI offerings, reinforcing that governing and securing AI agents is a foundational requirement for enterprise AI adoption at scale.



Responsible AI Summit 2026

September 21-23, 2026 | London, UK

The 4th Annual Responsible AI Summit gathered enterprise leaders and regulators in London to confront what it actually takes to govern AI at scale. Infosys Responsible AI took the stage on the Opening Plenary Panel, delivered a keynote reframing the industry's challenge from research to action, and engaged prospects through a dedicated showcase booth throughout the event.



India-UK Business School Dialogue 2026

September 15, 2026 | New Delhi, India

The India-UK Business School Dialogue, hosted by CABS in New Delhi, convened academic and industry leaders from India and the UK to discuss the future of management education in an AI-enabled workplace. Ashish Tewari of Infosys Responsible AI Office delivered a keynote on operationalizing Responsible AI through governance and assurance mechanisms and joined a panel on the skills future managers need to lead AI-enabled organizations responsibly.



Topaz AI Summit Hartford

September 2, 2026 | Hartford, Connecticut

Infosys Topaz leaders joined the Hartford AI Summit, co-presented with OpenAI and UiPath, showcasing a Responsible AI approach through proactive threat simulation and real-time security demonstrations, reinforcing Responsible AI as essential to enterprise-scale AI adoption.



14th TN Confederation of Indian Industry (CII) Finance Conclave

August 18, 2026 | Chennai, India

CII's 14th TN Finance Conclave in Chennai, themed Future-Ready Finance: Governance, Growth and Innovation, featured Bharathi Vokkaliga Ganesh of Infosys Responsible AI Office on a panel, Governing AI in Finance: From Automation to Intelligence, emphasizing explainable, auditable, transparent, human-accountable AI in finance.

Achievement

The Economic Times Enterprise AI Awards 2026

September 17, 2026 | Hotel Conrad, Bengaluru

Infosys was recognized as the Winner of the Responsible & Ethical AI Leadership Award at the 2nd Edition of The Economic Times Enterprise AI Awards, a recognition that cuts to the heart of what governance is meant to prove: that trustworthy AI isn't a claim, it's a practice. The win affirms Infosys' standing at the forefront of safe, transparent, and accountable AI at enterprise scale.



Contributors

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



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



RESPONSIBLE AI - UNFOLDING IN EVERY LAYER

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



© 2026 Infosys Limited, Bengaluru, India. All Rights Reserved. Infosys believes the information in this document is accurate as of its publication date; such information is subject to change without notice. Infosys acknowledges the proprietary rights of other companies to the trademarks, product names and such other intellectual property rights mentioned in this document. Except as expressly permitted, neither this documentation nor any part of it may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, printing, photocopying, recording or otherwise, without the prior permission of Infosys Limited and/ or any named intellectual property rights holders under this document.