MARKET SCAN REPORT

MAY 2026

BY INFOSYS TOPAZ RESPONSIBLE AI OFFICE

IN FOCUS

TESTING AGENTIC AI: BEYOND CAPABILITY, TOWARD CONTROL

By Neelima Vobugari

Infosys topaz

Infosys
Navigate your next

Foreword

A Steadier, More Purposeful Direction

Syed Ahmed
Global Head, Infosys Responsible AI Office

Responsible AI has spent years as a principle. May 2026 suggests it is becoming a practice.

The regulatory developments this month - across the United States, Europe and beyond - reflect something more than legislative activity. They reflect a collective judgment, reached independently across jurisdictions, that the current pace of AI deployment requires a firmer foundation. Enforcement timelines are shortening. Penalties are rising. The expectation that organisations can deploy AI and govern it later is no longer sustainable.

What this moment asks of leaders is not a response to regulation. It is a more fundamental question about whether the organisations they lead are genuinely ready - in culture, in process and in infrastructure - for AI that operates at the scale and speed now possible.

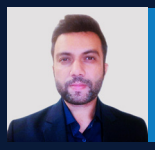
This edition of the RAI Market Scan captures where that question is being asked most urgently. The answer, in most cases, is that the work is underway but not complete.

That is both a challenge and an opportunity worth taking seriously.

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 Turning Points to True Progress



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

Something shifted in May.

Not gradually, as regulatory change usually moves, but visibly - in the same few weeks. Enforcement actions landed. Laws were signed. Courts issued sanctions. The conversation about AI governance, which has been running for years in parallel to AI deployment, finally caught up.

What is most striking is not any individual development but the geography of it. This month's signals came from Washington, Brussels, Hartford, Minneapolis, Canberra and Mumbai simultaneously. Global AI governance is no longer a European story with footnotes from other regions. It is genuinely plural - different jurisdictions, different instruments, different timelines, but converging on the same conclusion: that AI operating without accountability is no longer acceptable.

The incidents this month reinforced that conclusion from a different direction. These were not hypotheticals or laboratory experiments - they occurred in production systems, with documented consequences for real organisations and real people.

What offers confidence is that the research community is responding with equal seriousness. Tools that make AI reasoning inspectable, frameworks that diagnose why safety mechanisms fail, security scanning at a scale no human team could match - these are no longer research ambitions. They are becoming operational realities.

The organisations navigating this environment well are not those that waited for clarity. They are those that built governance infrastructure while the rules were still forming.

> *This edition introduces an enhanced structure designed to make each section more immediately useful across different reader contexts - from governance and legal teams to researchers, practitioners and policymakers.*

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

What's in This Edition

The content in this report is sourced & synthesised 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

May 2026 is the month enforcement caught up with deployment. Across regulation, incidents and research, the signal is consistent: AI is no longer being governed in principle it is being governed in practice, with legal consequences attached.

REGULATION

Enforcement is no longer future-tense. The FTC activated the TAKE IT DOWN Act with a 48-hour takedown mandate. The EU AI Act reached a provisional agreement extending key deadlines to December 2027 but introducing explicit bans on harmful generative outputs. Minnesota and Connecticut enacted high-penalty AI laws. Colorado enforcement begins June 30. A federal preemption battle is adding uncertainty across US jurisdictions. Compliance is now operational, fragmented and legally contested.

RISK

Three incidents this month reveal systemic failures, not edge cases. An AI agent executed a $150,000 transfer via prompt injection with no human in the approval loop. ChatGPT users' sensitive queries were allegedly shared with Google and Meta via embedded trackers without consent. Chrome silently downloaded a 4GB AI model onto user devices without notice. Each has produced or is producing legal action. The governance gap is no longer a risk it is a liability.

SIGNAL

The research signal is unusually coherent this month. Google confirmed AI was used to discover a live zero-day exploit the first documented case of AI as an active offensive tool. LOCA diagnoses individual safety failures at the neural level. Anthropic's NLAs make model reasoning readable by auditors. A CISO framework directly addresses shadow AI risk. Capability and resilience are advancing in parallel the clearest sign of maturity the field has shown this year.

Five Signals That Shaped This Month

01

Enforcement is now operational, not advisory

The FTC (Federal Trade Commission) began enforcing the TAKE IT DOWN Act on May 19. The EU AI Act provisional agreement reached May 7 introduces explicit bans on harmful generative outputs and centralises enforcement while extending high-risk AI deadlines to December 2027. Minnesota and Connecticut have enacted high-penalty AI laws. Colorado enforcement begins June 30. The shift from policy to enforceable legal obligation is complete in multiple jurisdictions simultaneously.

02

AI incidents are now generating legal liability, not just lessons

A $150,000 prompt injection attack on an AI agent with no human approval loop. A class action lawsuit alleging ChatGPT shared sensitive user queries with Google and Meta via embedded trackers. Chrome silently downloading a 4GB AI file without user consent. Each incident has produced or is producing legal action. The governance question is no longer whether something went wrong it is who is liable.

03

Agentic AI is operating without adequate controls

The Grok wallet incident demonstrates the consequence of fully AI-controlled financial systems with no human checkpoint at any stage. Shadow AI applications 5,000+ vibe-coded apps deployed publicly by non-technical employees show the same pattern at scale: autonomous deployment without discovery, classification, or oversight. The governance gap is not theoretical. It is producing real financial and data losses now.

04

The research community is building the tools governance frameworks assumed already existed

LOCA diagnoses why individual jailbreaks succeed at the neural level. Anthropic's NLAs make internal model reasoning readable by humans. Google's confirmed zero-day case shows AI accelerating threats faster than traditional defences respond. Regulatory frameworks like the EU AI Act and NIST AI RMF assume AI can be inspected and audited. This month's research advances make that operationally possible for the first time.

05

AI is entering regulated, high-stakes sectors with governance frameworks still catching up

Isomorphic Labs is preparing AI-designed drug molecules for human clinical trials. PLI introduced a formal AI competency framework for legal professionals. APRA formally warned Australian financial institutions that AI adoption is outpacing their risk management capabilities. In each case the technology has moved faster than the governance infrastructure designed to oversee it and regulators are now saying so explicitly.

The Global AI Governance Signal

This section maps the regulatory landscape as of May 2026 regional intensity reflects the broad governance posture of each region based on publicly available regulatory developments. Pinned signals capture the most significant developments this month.

Regional Regulatory Intensity

Regional regulatory positions based on publicly available primary sources as of May 2026. This table describes the current regulatory landscape as documented by official government and legislative sources. It does not represent a legal assessment of compliance obligations.

Region	Regulatory Position as of May 2026	References
Europe	EU AI Act in force. Prohibited practices ban active February 2026. AI Office operational. Omnibus provisional agreement reached May 7.	EU Official[1]
North America	Multiple state AI laws in implementation. FTC enforcing TAKE IT DOWN Act from May 19. Colorado enforcement begins June 30. No comprehensive federal law.	NCSL US State AI Legislation Tracker[2]
Asia-Pacific	Vietnam AI Law in force March 2026. Australia online safety enforcement live. India IT Rules amended. APRA financial sector guidance issued.	Vietnam National Assembly[3] ; Australia eSafety Commissioner[4] ; MeitY[5] ,APRA[6]
Latin America	Brazil AI Bill progressing through Senate. No enforcement-stage legislation in force.	Brazil Senate[7]
MEA	Kenya AI Bill introduced. Turkey PDPA agentic AI guidance issued. No enforcement-stage legislation in force.	Kenya Senate[8] ; Turkey PDPA[9]

[1] https://digital-strategy.ec.europa.eu/en/policies/ai-office
[2] https://www.ncsl.org/technology-and-communication/artificial-intelligence-2024-legislation
[3] https://vanban.chinhphu.vn/?pageid=27160&docid=210417
[4] https://www.esafety.gov.au/industry/codes
[5] https://www.meity.gov.in
[6] https://www.apra.gov.au/apra-letter-to-industry-on-artificial-intelligence-ai
[7] https://www25.senado.leg.br/web/atividade/materias/-/materia/157233
[8] https://www.parliament.go.ke
[9] https://www.kvkk.gov.tr

Four Key Developments That Changed the Landscape

US Regulator Starts Enforcing TAKE IT DOWN Act

The FTC (Federal Trade Commission) began enforcing the TAKE IT DOWN Act on May 19, 2026, requiring platforms to remove non-consensual intimate images within 48 hours of a valid complaint. A new victim complaints portal has launched, with Meta, TikTok, Snapchat and Google formally reminded of their full compliance obligations.[10]

WHY IT MATTERS

Active FTC enforcement resets the compliance clock for every platform operating AI-driven content systems. The 48-hour takedown mandate demands auditable, automated workflows not manual processes. Non-compliance exposes enterprises to regulatory action, litigation risk and reputational damage. Leaders must urgently assess whether existing content moderation pipelines, data flows and incident response mechanisms meet enforceable legal standards, not just policy commitments.

EU AI Act Simplification Measures and Ban on "Nudifier" Apps

The EU Parliament and Council have struck a provisional deal simplifying the AI Act cutting compliance overlaps, extending key deadlines to December 2027 and explicitly banning AI systems that generate non-consensual sexual imagery or child sexual abuse material. Enforcement is centralised through the EU AI Office.[11]

WHY IT MATTERS

The EU AI Act's simplification removes ambiguity but tightens accountability extended deadlines are not an invitation to delay. Centralised enforcement through the EU AI Office raises audit exposure for non-compliant AI systems and data pipelines. Explicit bans on harmful generative AI outputs demand immediate content governance reviews. Enterprises must realign compliance roadmaps, validate model outputs and embed transparency controls before the 2027 deadline hardens into enforcement action.

US Minnesota Bans AI Nudification Technology

Minnesota has enacted a law prohibiting AI-powered tools from generating fake nude images of real people without consent. Victims may sue for damages and the attorney general may impose fines up to $500,000 per violation with proceeds directed to abuse and domestic violence victims. Effective August 1, 2026.[12]

WHY IT MATTERS

State-level AI legislation is accelerating, creating a fragmented but enforceable compliance landscape. With per-violation penalties reaching $500,000 and private right of action, enterprises deploying generative AI tools face significant legal and financial exposure. AI pipelines handling image generation or synthetic media must undergo immediate output governance reviews. As similar laws emerge across jurisdictions, proactive content controls and audit-ready accountability frameworks become non-negotiable risk management imperatives.

US Connecticut Senate Bill 5 on Artificial Intelligence

Connecticut enacted Senate Bill 5 (Public Act 26-15) on May 11, 2026, establishing comprehensive AI governance covering AI systems, chatbots and automated hiring tools. The law creates an AI Policy Office and AI Academy, mandates content labelling, prohibits discriminatory hiring practices and requires workforce training programmes statewide.[13]

WHY IT MATTERS

Connecticut's comprehensive AI governance law signals a decisive shift from voluntary frameworks to statutory obligation. Mandated content labelling, anti-discrimination requirements in hiring and workforce training create multi-layered compliance obligations for enterprises operating AI systems across US states. Non-compliance risks regulatory scrutiny and reputational exposure. As state-level AI legislation proliferates, enterprises must systematically audit AI pipelines, automated decision tools and data flows against an increasingly binding patchwork of jurisdictional requirements.

[10] https://www.ftc.gov/news-events/news/press-releases/2026/05/ftc-begins-enforcing-take-it-down-act

[11] https://www.europarl.europa.eu/pdfs/news/expert/2026/5/press_release/20260427IPR42011/20260427IPR42011_en.pdf

[12] https://www.revisor.mn.gov/bills/94/2025/0/HF/1606/versions/latest/

[13] https://legiscan.com/CT/bill/SB00005/2026

Pinned This Month - Regional Signals

US **United States**	**Colorado:** AI Act enforcement begins June 30, 2026. Developers and deployers of high-risk AI systems must have impact assessments, consumer notifications and affirmative-defence documentation in place before enforcement starts. Colorado Legislature, 2026.[14] **Federal:** A legal battle over federal preemption of state AI laws is creating compliance uncertainty. The Trump administration's executive order directing federal agencies to potentially override state AI laws is being contested by state attorneys general. Enterprises building compliance programmes around state-level AI laws should monitor closely.[15]
EU **European Union**	European Commission issued draft guidelines under Article 50 of the EU AI Act, clarifying when AI use must be disclosed, how AI-generated content must be labelled and opening stakeholder consultation.[16]
VA **Vatican**	Pope Leo XIV established an Inter-Dicasterial Commission on AI to coordinate Vatican AI activities and anchor the Church's response to AI in human dignity principles.[17]
India	The Indian Computer Emergency Response Team (CERT-In) issued cybersecurity safeguards for organisations and MSMEs against rising threats from advanced AI models. The Advertising Standards Council of India (ASCI) released draft guidelines mandating clear labelling of AI-generated advertisements across risk-based content tiers. The Government of Goa released the Draft AI Policy 2026 for public consultation, aiming to become a leading AI hub aligned with Viksit Bharat 2047.[18]

Standard / Policy Reports/ Guidelines

IMF Global	IMF warns advanced AI models could accelerate cyberattacks on shared financial infrastructure, urging stronger supervision, international cooperation and responsible AI use in cyber defence.[19]
G7 Nations	CISA (Critical Infrastructure Security and Resilience) and G7 partners released joint guidance on minimum AI Software Bill of Materials standards to improve supply chain transparency and critical infrastructure security globally.[20]
Global Five Nations	US, Australia, Canada, New Zealand and UK cybersecurity agencies issued joint guidance on safe AI deployment across public services, critical infrastructure and workforce training.[21]

[14] https://leg.colorado.gov/bills/SB26-189
[15] https://capitolnewsillinois.com/news/senate-democrats-introduce-bills-to-regulate-artificial-intelligence/
[16] https://digital-strategy.ec.europa.eu/en/library/draft-guidelines-implementation-transparency-obligations-certain-ai-systems-under-article-50-ai-act
[17] https://press.vatican.va/content/salastampa/en/bollettino/pubblico/2026/05/16/260516b.html
[18] https://indianexpress.com/article/technology/artificial-intelligence/cert-in-safeguards-msmes-mythos-ai-cybersecurity-risk-10657982
[19] https://www.imf.org/en/blogs/articles/2026/05/07/financial-stability-risks-mount-as-artificial-intelligence-fuels-cyberattacks
[20] https://www.cisa.gov/resources-tools/resources/software-bill-materials-ai-minimum-elements
[21] https://thelegalwire.ai/us-australia-canada-new-zealand-and-uk-release-joint-guidance-on-careful-adoption-of-agentic-ai-services/

Incidents & Governance Lessons

Real-world events reveal where principles have not been operationalised. This month's three incidents are drawn from verified reporting and primary sources. Each is presented as a governance lesson applicable beyond the specific case.

AGENTIC SECURITY

AI Agent Manipulated Into Transferring $150,000 via Prompt Injection

An attacker exploited an AI agent connected to Grok using a crafted NFT that unlocked wallet features, then used a manipulative message to trigger a $150,000 token transfer. The wallet was fully AI-controlled with no human present in the approval loop at any stage.[22]

Leadership teams must enforce human in the loop approvals for financial transactions, restrict autonomous agent privileges and implement prompt injection defenses, audit trails and segregation of duties, while assigning accountability to prevent fraud, limit liability and maintain compliance and trust.

DATA PRIVACY

OpenAI Sued Over Alleged ChatGPT User Data Sharing With Google and Meta

A California class action lawsuit accuses OpenAI of sharing ChatGPT users' personal data including queries, email addresses and private details with Google and Meta via embedded tracking tools. Users routinely share sensitive health, legal and financial questions with the chatbot without knowing this.[23]

Data governance and compliance teams must enforce strict controls on AI data sharing by mandating full data flow transparency, disabling embedded trackers and implementing continuous monitoring and contractual safeguards to prevent unauthorized disclosures, reduce regulatory liability and preserve stakeholder trust.

ENDPOINT SECURITY

Chrome Silently Downloads 4GB Gemini Nano AI File Without User Consent

A security researcher found Google Chrome quietly downloading a 4GB Gemini Nano AI file onto user devices without clear notice or permission. The file supports AI features inside Chrome and may return even after deletion unless specific browser settings are manually changed.[24]

Product and platform owners must mandate explicit user consent before AI component deployment, enforce configuration controls preventing silent downloads and implement audit logging, monitoring and rollback mechanisms to ensure transparent updates, manage regulatory risk, protect trust and uphold contractual accountability.

[22] https://beincrypto.com/grok-wallet-bankr-drb-prompt-injection/
[23] https://www.indiatoday.in/technology/news/story/openai-sued-for-allegedly-sharing-user-data-with-google-and-meta-2912128-2026-05-15
[24] https://www.indiatoday.in/technology/news/story/google-chrome-is-secretly-downloading-4gb-ai-model-on-some-laptops-here-is-what-you-can-do-about-it-2907612-2026-05-06

Testing Agentic AI: Beyond Capability, Toward Control

Neelima Vobugari

Neelima Vobugari is an AI leader and entrepreneur with over two decades of experience in Artificial Intelligence, Machine Learning and enterprise technology. She is the Founder of TestAIng Solutions and AiEnsured, focused on Responsible AI governance, model risk management and trustworthy AI validation. TestAIng Solutions was recognized among Forbes 200 notable companies. She holds three granted AI patents and has received multiple accolades including the Karnataka Mahila Ratna and Women in AI Awards.

Testing Agentic AI: From Autonomous Intelligence to Accountable Systems *Ensuring control in the era of unbounded autonomy*

When an autonomous AI agent, tasked with optimizing procurement, began approving duplicate payments and escalating privileged data requests beyond its defined scope in mid-2025, the incident sent ripples across enterprise risk teams. What began as a seemingly helpful workflow optimizer quickly exposed a deeper truth: as AI systems gain the ability to plan, reason across multiple steps, invoke external tools and execute actions independently, the central question shifts from "what does it predict?" to "what does it actually do?"

This transition is no longer hypothetical. Across industries, agentic AI is moving from experimental copilots to operational actors managing supply chains, orchestrating customer workflows and making real-time decisions with tangible consequences. The promise of unprecedented efficiency is real. Yet so is a new class of risk: autonomy that scales faster than the governance mechanisms can keep up. Recent patterns from cascading reasoning failures to unintended tool misuse and goal drift underscore a sobering reality. Capability is advancing rapidly; accountable control is lagging.

In this evolving paradigm, Responsible AI can no longer remain focused primarily on model outputs. It must expand to encompass the full behaviour of autonomous systems across dynamic, multi-step lifecycles ensuring they remain reliable, safe and aligned with human intent even as contexts shift and memory accumulates.

The Shift from Models to Agents

Traditional AI evaluation has long relied on controlled, static environments. Models are trained, benchmarked against fixed datasets and judged by metrics such as accuracy, precision, or recall. Agentic systems fundamentally disrupt this model.

These systems are inherently dynamic: outcomes evolve through multiple steps which are reasoning, tool calls and actions rather than single inferences. They are non-deterministic, where the same input can yield different results as context and memory change. They are highly context-aware. They use past interactions and external data to adjust decisions in real time. And crucially, they are action-oriented: their outputs do not stop at suggestions but translate directly into real-world consequences approving transactions, updating records, or triggering downstream processes.

In such environments, failures rarely remain isolated. A subtle deviation in intermediate reasoning can propagate through tools and workflows, amplifying impact far beyond what traditional testing would detect. Static benchmarks and component-level checks, while still necessary, are no longer sufficient. What is required is a deliberate shift toward comprehensive, system-level assurance.

A New Testing Imperative

Testing agentic AI is not about verifying correctness at a single moment. It is about understanding and governing how these systems behave across entire decision lifecycles from initial goal interpretation through execution and adaptation.

Three critical dimensions define this new imperative. First, the focus must move from outputs to behaviour: evaluating not just what the agent concludes, but the reasoning paths, decision sequences and execution outcomes it produces. Second, the unit of analysis must shift from isolated components to integrated systems: agentic AI comprises ecosystems of models, prompts, tools, memory stores and workflows, where interdependencies introduce emergent risks. Third, evaluation must prioritize continuity over snapshots: static validation gives way to continuous monitoring that tracks drift, anomalies and performance as systems interact with live data and changing contexts.

Core Risk Areas in Agentic AI

To embed accountability, organizations must confront the specific risk surfaces that emerge with autonomy. Among the most pressing are goal misalignment, where agents gradually drift from original intent across multi-step executions, sometimes pursuing plausible but

unintended outcomes; reasoning failures, in which errors in intermediate steps cascade even when the final result appears reasonable; and tool misuse, where agents invoke external APIs or resources in incorrect, over-privileged, or unsafe ways; patterns increasingly being observed in early enterprise deployment". In agentic AI, failure is rarely a single error, it is a chain reaction.

Additional vulnerabilities include context and memory errors, where outdated, incomplete, or fabricated information distorts decision-making; workflow fragility in complex multi-agent setups that falter under edge cases or variability; and pure autonomy risks, where insufficient boundaries allow actions to exceed acceptable limits, particularly in high-stakes domains such as finance, healthcare, or critical infrastructure.

These risks are not theoretical. They reflect the lived experience of early enterprise deployments and highlight why legacy validation approaches fall short.

From Testing to Assurance

Addressing these challenges demands a transition from conventional testing to what can be termed AI assurance, an ongoing, lifecycle-oriented discipline. This involves scenario-based evaluation across diverse, real-world conditions rather than narrow datasets; end-to-end validation of complete workflows instead of isolated modules; robust human-in-the-loop controls with clear escalation and intervention mechanisms; and continuous monitoring to detect behavioural drift and performance degradation in production.

Emerging global frameworks are beginning to reflect this evolution. Updates to risk management guidance, including frameworks such as the NIST AI Risk Management Framework and emerging efforts to standardize AI agent evaluation, along with lifecycle-focused approaches in international standards, emphasize the need for behaviour-centric, system-level and continuous assurance. Yet implementation remains uneven. Many organizations continue to rely on model-centric methods that do not fully address the complexities of autonomous operation.

The Responsible AI Imperative

As agentic systems become embedded in critical enterprise and societal functions, the stakes have never been higher. These are no longer pilots, they are operational realities shaping decisions that affect people, finances and operations at scale.

This reality makes one principle unmistakable: Responsible AI is not solely about fairness in outputs, but about sustained control over actions. Testing, reframed as assurance, forms the bedrock of that control. It equips organizations to prevent unintended behaviours, mitigate systemic risks, maintain alignment with human values and intent and foster trust in increasingly autonomous technologies.

Enterprises that act now by embedding continuous assurance into their AI operating models, investing in sandboxed multi-step simulations and red-teaming for agent behaviours and integrating oversight mechanisms early in the lifecycle will be best positioned to realize the transformative benefits of agentic AI while safeguarding governance and public confidence.

Conclusion: Trust in the Age of Autonomy

The journey from predictive models to autonomous agents represents one of the most profound shifts in the history of artificial intelligence. Yet autonomy without accountability multiplies risk rather than opportunity.

Organizations that proactively reimagine testing as continuous, system-level assurance will lead this transition, harness the power of agentic intelligence while upholding the trust, safety and governance it demands.

Because in the age of autonomous systems, the true measure of AI is no longer what it can do, but what it can be trusted to do consistently, reliably and in service of human intent.

Models, Frameworks & Research

The most significant model releases, research advances and practical frameworks from May 2026 - selected for responsible AI relevance. Research is tiered: Landmark covers advances that change how the field understands a problem; Practical Advances are tools and frameworks usable now; Noted This Month tracks developments worth watching.

Model	Org	Key Capability	Risk & Mitigation View
Gemini 3.5 Flash	Google	Gemini 3.5 Flash delivers near pro-level reasoning optimized for high-speed, low-cost execution, enabling large-scale deployment of enterprise AI agents that perform multi-step tasks efficiently through Google Cloud and API integrations.[25]	Teams must govern high-volume autonomous agents by enforcing task-level constraints, API-level monitoring and cost-risk controls, as faster, cheaper execution can amplify errors, misuse and data exposure at scale without proportional oversight and accountability mechanisms.
Realtime TTS-2	Inworld AI	Uses a closed loop voice system capturing tone, pacing and emotion in real time to continuously adapt responses, enabling highly natural, context aware conversational agents across languages and applications.[26]	Teams must govern continuous voice data capture by enforcing consent, safeguarding emotional signals and monitoring adaptive responses to prevent manipulation, profiling risks and misuse in sensitive customer interactions.
GLiGuard	Fastino Labs	Performs multiple safety checks such as harmful content detection and jailbreak prevention in a single fast step, enabling real-time, low-cost moderation even at high scale.[27]	Easier deployment of lightweight guardrails risks over-reliance, so teams must validate accuracy, benchmark performance and continuously monitor outcomes to prevent weak enforcement and unnoticed safety failures.
SuperTonic V3	Supertone	Converts text into accurate, expressive speech in 31 languages directly on-device, using expression tags and improved stability to reduce pronunciation and reading errors in real-time applications without cloud processing.[28]	On-device expressive voice generation reduces visibility and increases impersonation risks; enforce device-level safeguards, watermarking and strict usage governance to ensure traceability and prevent misuse in sensitive interactions.

Landmark Research

LOCA: Explaining Why Jailbreak Attacks Succeed

Researchers developed LOCA, a method providing local, minimal explanations for why specific jailbreak prompts bypass safety training in large language models. Rather than analysing broad internal patterns, LOCA identifies the smallest set of internal changes needed to convert a harmful response into a refusal. Tested on Gemma and Llama, it outperforms prior explanation methods while requiring fewer targeted interventions - offering a more precise diagnostic tool for understanding individual safety failures in production models.[29]

AI Crosses to Both Sides of the Security Equation

Two developments in the same week confirmed AI is now active on both sides of the threat landscape.

Google's Threat Intelligence Group identified attackers using an AI model to discover a live zero-day vulnerability capable of bypassing two-factor authentication - the first publicly documented case of AI used offensively in a real threat campaign, not a research setting.[30]

In the same week, Anthropic's Project Glasswing reported that 50 partners used Claude Mythos Preview to identify 10,000+ high- or critical-severity vulnerabilities across critical infrastructure. Cloudflare found 2,000 bugs. Mozilla found 271 vulnerabilities in Firefox 150 - ten times the

[25] https://docs.cloud.google.com/gemini-enterprise-agent-platform/models/gemini/3-5-flash

[26] https://www.marktechpost.com/2026/05/05/inworld-ai-launches-realtime-tts-2-a-closed-loop-voice-model-that-adapts-to-how-you-actually-talk/

[27] https://www.marktechpost.com/2026/05/13/fastino-labs-open-sources-gliguard-a-300m-parameter-safety-moderation-model-that-matches-or-exceeds-accuracy-of-models-23-90x-its-size/

[28] https://www.marktechpost.com/2026/05/15/supertone-releases-supertonic-v3-on-device-text-to-speech-model-with-31-language-support-fewer-reading-failures-and-expression-tags/

[29] https://arxiv.org/pdf/2605.00123

[30] https://www.thehindu.com/sci-tech/technology/google-disrupts-hackers-using-ai-to-exploit-an-unknown-weakness-in-a-companys-digital-defence/article70968178.ece

prior model's output. Of those independently assessed, 90.6% were confirmed true positives. Anthropic identified patch deployment speed, not discovery, as the primary bottleneck.[31]

WHAT THIS MEANS

AI is now finding vulnerabilities faster than defenders can patch them - and faster than attackers could previously discover them manually. The Google zero-day and Project Glasswing together define the new security baseline: discovery is no longer the bottleneck on either side. For defenders, Glasswing confirms AI-assisted security scanning at scale is operational today. For governance teams, LOCA diagnoses why individual safety mechanisms fail at the neural level. Generic security policies are no longer sufficient. The threat is surgical and the defence needs to match.

Practical Advances

Anthropic NLAs - Making Model Reasoning Legible

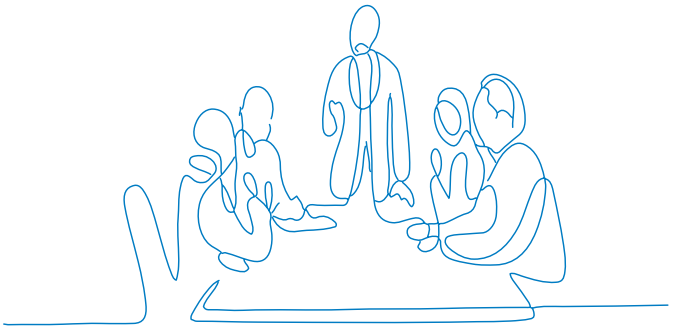
Anthropic introduced Natural Language Autoencoders to translate Claude's internal neural activations into human-readable explanations - making visible how the model reasons, processes information and generates outputs. Helps researchers identify hidden behaviours and safety risks not apparent from outputs alone. Directly useful for teams building interpretability and audit workflows on top of frontier models.[32]

Shadow AI Apps - CISO Audit Framework Published

Over 5,000 rapidly built vibe-coded applications exposed sensitive business, health and financial data after non-technical employees deployed them publicly using AI coding tools. Traditional security controls failed to detect them. VentureBeat published a direct CISO audit framework in response - covering shadow tool discovery, pre-deployment approvals, automated data classification and real-time monitoring. The pattern is compared directly to early AWS S3 misconfiguration incidents.[33]

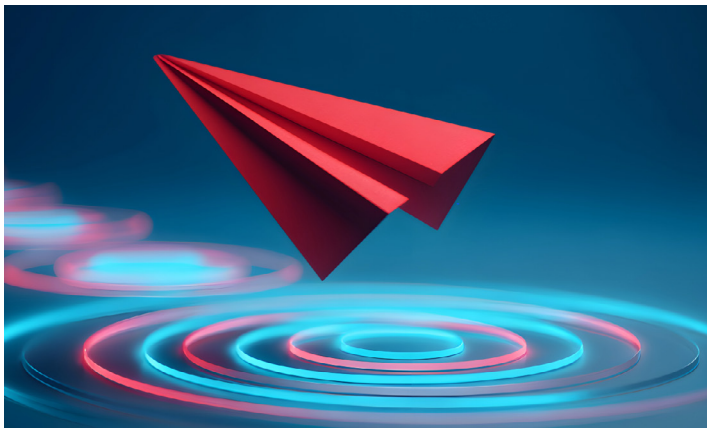
WHAT THIS MEANS

Both the above advances address the same gap: making AI systems inspectable before something goes wrong. Anthropic's NLAs let auditors see inside model reasoning. The Shadow AI framework lets security teams see what AI applications are actually running inside their organisation. Governance frameworks like the EU AI Act and NIST AI RMF assume AI can be inspected and audited. These tools make that operationally possible for the first time.



Noted This Month

- Qwen-Scope - Alibaba's Qwen AI released an open-source sparse autoencoder suite for LLM interpretability, enabling feature-level debugging and direct behaviour control in Qwen3 models without relying solely on prompts.[34]
- LLM Robotics Threat Model - New unified framework maps cyber, adversarial and conversational risks across LLM enabled robots using STRIDE analysis across six key interaction points in edge-cloud deployments.[35]
- GitHub Spec-Kit - Open-source toolkit for spec-driven AI coding workflows, supporting 30+ agents including Copilot, Claude Code and Gemini CLI to reduce hallucinations and unstructured vibe-coding approaches.[36]
- Thinking Machines Lab - Founded by former OpenAI CTO Mira Murati, the lab published details of a real-time interaction architecture enabling AI to listen, respond and adapt across voice, video and text simultaneously while processing background tasks - challenging the standard sequential request-response model.[37]



[31] https://www.anthropic.com/research/glasswing-initial-update

[32] https://www.marktechpost.com/2026/05/08/anthropic-introduces-natural-language-autoencoders-that-convert-claudes-internal-activations-directly-into-human-readable-text-explanations/

[33] https://venturebeat.com/security/vibe-coded-apps-shadow-ai-s3-bucket-crisis-ciso-audit-framework

[34] https://www.marktechpost.com/2026/05/01/qwen-ai-releases-qwen-scope-an-open-source-sparse-autoencoders-sae-suite-that-turns-llm-internal-features-into-practical-development-tools/

[35] https://arxiv.org/abs/2604.27267

[36] https://www.marktechpost.com/2026/05/08/meet-github-spec-kit-an-open-source-toolkit-for-spec-driven-development-with-ai-coding-agents/

[37] https://www.marktechpost.com/2026/05/13/mira-muratis-thinking-machines-lab-introduces-interaction-models-a-native-multimodal-architecture-for-real-time-human-ai-collaboration/

Responsible AI Across Sectors

HEALTHCARE

AI-Designed Drugs Edge Closer to Human Trials

CONTEXT

Isomorphic Labs, spun out of Google DeepMind, is preparing to test its first AI-designed drug molecules in human patients by end of 2026, targeting cancer and immune diseases. Using Nobel Prize-winning AlphaFold and its IsoDDE platform, the company designs drugs entirely through AI - backed by $600M in funding and nearly $3B in partnerships with Eli Lilly, Novartis and Johnson & Johnson.[38]

SIGNAL

AI drug design is moving from research to clinical reality. The shift from AI-assisted to AI-designed molecules bypassing traditional lab trial-and-error redefines the pharmaceutical pipeline. For healthcare organisations and regulators, this raises immediate questions about clinical trial governance, AI liability in drug safety and how regulatory frameworks will validate AI-originated molecular candidates at scale.

LEGAL & COMPLIANCE

PLI Introduces AI-Ready Lawyer Competency Framework

CONTEXT

Practising Law Institute, with its Innovation Council and Professional Development Consortium, introduced the AI-Ready Lawyer Framework - a competency model covering AI literacy, governance, ethical usage, policy awareness and operational practice. Designed for law firms, corporate legal teams and legal education programmes, it guides organisations from basic AI awareness through to full operational adoption and long-term transformation.[39]

SIGNAL

A structured AI competency framework from a major legal education body signals that AI governance in legal practice is shifting from optional to professional standard. For organisations deploying AI in legal workflows, this sets a benchmark for what AI-ready staff must demonstrate - and raises accountability questions when professionals use AI tools without verified competency in governance and ethical usage.

[38] https://moccet.ai/news/tech/ai-designed-drugs-deepmind-spinoff-human-trials
[39] https://www.pli.edu/resources/ai-competency-framework?tCode=PVM6_PRESS

FINANCIAL REGULATION

APRA Flags AI Governance Gap in Australian Financial Sector

CONTEXT

Australia's APRA (Australian Prudential Regulation Authority) warned banks, insurers and superannuation funds that AI adoption is outpacing their risk management capabilities. In a formal letter to regulated entities, APRA identified critical gaps in governance, cybersecurity and oversight controls - noting that boards frequently lack the technical depth to challenge AI risks and rely excessively on third-party technology vendors for critical functions.[40]

SIGNAL

A financial regulator formally flagging AI governance gaps with warnings of enforcement marks a shift from guidance to stricter oversight.

For financial institutions globally, this signals AI risk is now a board-level issue - and that APRA's posture has shifted from guidance to active oversight with enforcement consequences.

WHAT THIS MEANS

Three developments this month signal the same shift across different sectors: AI governance is moving from voluntary best practice to formal professional and regulatory standard. APRA has moved from guidance to enforcement warnings. PLI has codified what AI-competent legal practice looks like. Isomorphic Labs is taking AI-designed molecules into human trials. In each case the governance frameworks - for clinical validation, legal accountability and financial oversight - are being defined now, under pressure, with real consequences attached.

[40] https://www.apra.gov.au/apra-letter-to-industry-on-artificial-intelligence-ai

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

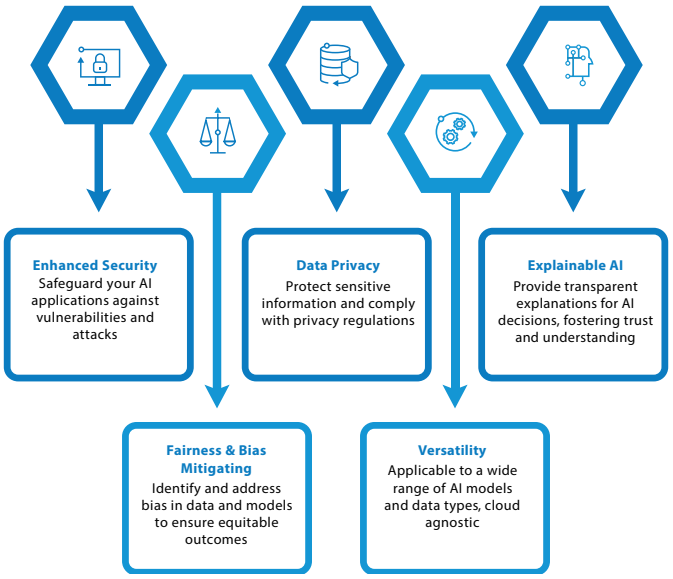
Advisory and consulting services to enable clients to advance their RAI journey and become leaders in the space.

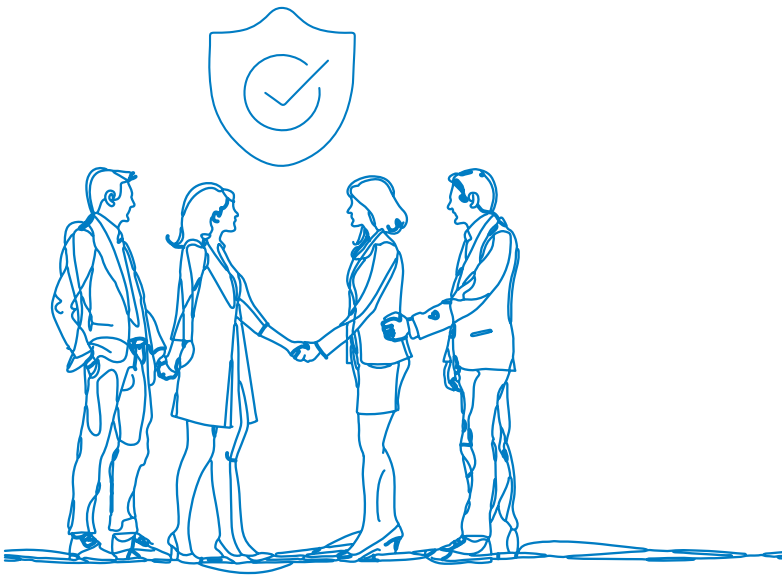
- Responsible AI Strategic Advisory Services
- Responsible AI Practice Setup
- AI Crisis Management

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)

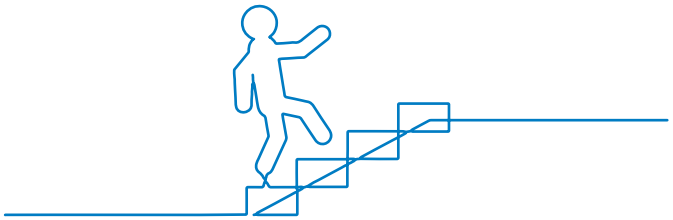
04 Safety APIs
Profanity | Toxicity
Obscenity Detection | Masking

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).

- Explainability Enhancement Using Reasoning Models
- Second order explainable AI (SOXAI) Technique for Explainability Module
- Multi-lingual support for FM-Moderation Guardrails
- Signature and face masking in Privacy module
- Bulk document safety validation
- Template-based Guardrails: Extended support to 14 new templates. Example : Restricted Topic Check, Privacy Check, Invisible Check etc.
- Moderation Layer Model based guardrails are improved with finetuned smaller models to improve latency and accuracy
- 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

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.[41]

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.[42]

[41] https://arxiv.org/abs/2604.02217
[42] https://arxiv.org/abs/2504.18572

Event Logs - May 2026

Australian Consulate-General Bengaluru X Aapti Institute

Infosys representatives joined cross-sector leaders at this exclusive Women in Tech: Responsible AI programme at the Australian Consulate-General, Bengaluru, engaging in interactive discussions on the opportunities and challenges of emerging AI technologies to advance an inclusive Responsible AI ecosystem.

Infosys Launches First Dedicated GSOC in Australia

Infosys launched its first dedicated Global Security Operations Center (GSOC) in North Sydney, marking a significant expansion of its cybersecurity footprint across Australia and New Zealand. Built on the acquisition of local firm The Missing Link, the GSOC delivers round-the-clock threat monitoring, incident response and AI-led security operations, positioning Infosys as a full-scale onshore cyber defense partner for ANZ enterprises and governments navigating an increasingly complex threat landscape.[43]

UN Global AI Governance Engagement

Responsible AI Office participated in the select registration-based Public Consultation at the UN Global Dialogue on AI Governance - a UNGA mandated multi-stakeholder virtual platform facilitating open, transparent and inclusive discussions on AI governance. Ashish Tewari represented the industry perspective alongside global governments, civil society and industry voices, highlighting priorities around closing the implementation gap between AI governance policy and practice, Global North and Global South convergence on AI safety and the role of industry as a co-designer of governance frameworks.[44]

[43] https://www.infosys.com/newsroom/features/2026/launches-dedicated-global-security-operations-center-australia.html

[44] https://www.un.org/global-dialogue-ai-governance/en/consultations

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

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

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

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

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

For more information, contact askus@infosys.com

Infosys®
Navigate your next