MARKET SCAN REPORT

JUNE 2026

BY INFOSYS TOPAZ RESPONSIBLE AI OFFICE

Infosys topaz

IN FOCUS

RESPONSIBLE AI AT SCALE: MOVING FROM GOVERNANCE PRINCIPLES TO ENTERPRISE PRACTICE

By M Chockalingam

Infosys
Navigate your next

Foreword

The Governance Compact Has Changed

Syed Ahmed
Global Head, Infosys Responsible AI Office

When organizations adopted responsible AI principles, there was an implicit compact: act in good faith, build toward compliance, demonstrate intent. That compact assumed time.

June 2026 suggests the time assumption no longer holds.

Enforcement, liability and capability failure have converged — not sequentially as governance programs planned for, but simultaneously. Four themes run through this edition: AI cybersecurity as a strategic imperative, sovereign AI as national policy, transparency as an enforceable obligation, and diverging national development paths. These are not emerging signals. They are the new operating conditions.

The question for leaders is no longer whether their organization is committed to responsible AI. It is whether the governance they built was designed for the AI they are actually running today.

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 Anticipated to Actual

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

Some months move the needle. This one changed the instrument.

The signals in this edition share a quality earlier months did not: finality. Enforcement has begun. Liability has attached. Production systems have failed. The questions governance programs have been deferring now have answers — and those answers are arriving through court filings and regulatory orders, not policy consultations.

The incidents this month are not edge cases. A global professional services firm withdrew a hallucinated report. A police officer faces criminal charges for AI-fabricated evidence. A cybercrime group industrialized a major AI platform to build 1.59 million phishing sites. Each is presented here as a governance lesson because that is what it is.

The tooling advances offer a counterpoint. Runtime governance for agentic AI, deployment simulation before release, adversarial prompt detection — these are available today. The organizations that move now will be ahead.

June 2026 is the month AI governance stopped being anticipatory. The organizations that treat this as a prompt to act will define what responsible AI leadership looks like next.

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

What's in This Edition

The Signals Behind the Headlines

Regulation has penalties now. Incidents have lawsuits now. Agentic AI is inside enterprise systems now. June 2026 marks the month AI governance moved from principle to consequence, and organizations that treated compliance as optional are already finding out what enforcement looks like.

REGULATION

Enforcement has arrived. The EU AI Act nudifier ban is active. Colorado signed the first US law regulating AI chatbot operators with civil penalties of $5,000 per violation. The UK's CMA issued a world-first publisher opt-out requirement against Google. India's Supreme Court prohibited AI from determining judicial outcomes. Five jurisdictions moved from policy to enforceable obligation in a single month.

RISK

Unverified AI output is producing harm simultaneously across professional services, law enforcement and criminal infrastructure. A global professional services firm withdrew its flagship AI report after hallucinated case studies were exposed. A UK police officer faces criminal charges for AI-fabricated court evidence. A Chinese cybercrime group built 1.59 million phishing sites using Google's own AI.

SIGNAL

Agentic AI is already inside enterprise systems. Microsoft's MAI-Code-1-Flash is active inside GitHub Copilot processing proprietary codebases. NVIDIA's Nemotron Ultra runs autonomous agents on enterprise infrastructure with no external API dependency. Alibaba's Qwen3.7-Plus writes code, calls external APIs and self-corrects without human approval. The question enterprises must now answer is whether their governance frameworks have kept pace with the systems already running inside them..

Five Signals That Shaped This Month

01

Regulation is now enforcement, not warning

The EU AI Act nudifier ban is now active, requiring companies to remove harmful generative capabilities within the compliance deadline. Colorado signed HB26-1263, the first US state law imposing enforceable obligations on AI chatbot operators, with civil penalties of $5,000 per violation effective January 2027. Canada's Bill C-34 and Spain's draft Organic AI Law add child safety mandates and mandatory human oversight. The FSB has set a binding governance baseline across 24 G20 financial jurisdictions.

02

AI outputs without human verification are producing institutional liability

A global professional services firm withdrew its flagship AI report after hallucinated case studies attributed to major clients were exposed, with fewer than one in nine citations verified as accurate. A UK police officer faces criminal charges for AI-fabricated court evidence across multiple cases. A Chinese cybercrime group industrialized Google's own AI to build 1.59 million phishing sites targeting banks and government agencies. Three sectors, three different failure modes, one common cause — AI output accepted without independent human review.

03

Agentic AI is operating inside enterprise systems ahead of governance

Microsoft's MAI model family is now active inside GitHub Copilot, bringing AI directly into enterprise code environments. NVIDIA's Nemotron 3 Ultra runs autonomous agents on enterprise infrastructure with no external API dependency. Alibaba's Qwen3.7-Plus writes code, calls external APIs, and self-corrects without human approval at any step. Agentic AI is operating inside enterprise systems today, ahead of the governance infrastructure designed to oversee it.

04

AI-enabled cybercrime has reached industrial scale

A Chinese cybercrime group used Google's Gemini AI to produce over 1.59 million fraudulent websites impersonating banks, government agencies and retailers. The FBI estimates the operation stole 3.87 million credit card numbers causing nearly $1.9 billion in losses. Google obtained an emergency restraining order blocking the operation worldwide. AI is no longer just accelerating cyberattacks. It is enabling their industrialisation at a scale no human operation could sustain.

05

Governance tooling has finally caught up with deployment reality

OpenAI's Deployment Simulation predicts real-world model behavior before release using actual user conversations rather than synthetic test prompts. The MTK framework detects jailbreak attempts by tracking internal prompt trajectories rather than surface text. Microsoft's Agent Governance Toolkit intercepts every tool call an autonomous agent attempts before execution. The tools needed to audit and govern agentic AI at runtime now exist and are available. The question is whether enterprises are deploying them.

The Global AI Governance Signal

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

Regional Regulatory Intensity

Governance posture assessments are synthesized from publicly available primary sources including official government publications, legislative records, and regulatory body announcements reviewed as of June 2026. Underlying sources are cited in full in the Key Developments and Regional Signals sections of this report. This table does not represent a legal assessment of compliance obligations in any jurisdiction.

Region	Regulatory Position as of May 2026	Stage*
Europe	Binding AI obligations active across member states. National implementation accelerating with enforcement bodies designated and operational.	Enforcement Active
North America	First chatbot-specific law enacted. Federal preemption debate intensifying. No unified federal AI law in force.	Legislation Enacted - State Level
Asia-Pacific	Rapid regulatory development across multiple jurisdictions. Sector-specific guidance emerging in India and Singapore. Australia self-regulation model under strain.	Framework Development
Latin America	Legislative exploration underway in Argentina and Brazil. No enacted AI-specific law in force across the region.	Early Exploration
MEA	UAE consolidating national AI authority. Regulatory frameworks not yet enforcement-stage across the region.	Strategy Stage

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

Four Key Developments That Changed the Landscape

UK Regulator Delivers World-First Ruling Giving Publishers Control Over AI Use of Their Content

The UK's Competition and Markets Authority imposed a conduct requirement on Google the first of its kind anywhere in the world giving publishers the ability to opt out of their content being used to train AI models and power AI features in Google Search, including AI Overviews. Publishers will also receive proper content attribution with clear links in AI-generated search results, strengthening their negotiating position in content licensing discussions with Google. The requirement follows Google's designation as holding strategic market status in general search. Google has nine months to implement all required changes.[1]

WHY IT MATTERS

The CMA has done what no regulator has done before: forced a direct, enforceable separation between a platform's AI training pipeline and the content it consumes without consent. This is not a guideline or a voluntary code — it is a legal conduct requirement with a compliance deadline. For every enterprise that publishes content, operates a media or data asset, or relies on AI-generated search summaries in client-facing products, this ruling resets the baseline. The opt-out mechanism is the precedent that matters most: it establishes that publishers have a legal right to withhold their content from AI systems, and that regulators will enforce it.

Singapore Sets Data Governance Expectations Across the Full GenAI Lifecycle

Singapore's Personal Data Protection Commission (PDPC) published proposed advisory guidelines clarifying how personal data can be collected and used across the generative AI lifecycle, from model development through deployment and post-deployment operations. The guidelines address when organizations can use publicly available data without consent, notification and consent obligations during AI model training, data protection responsibilities across the AI supply chain, and how individuals can exercise rights regarding their personal data in AI systems. Open for public consultation, the guidelines are intended to complement Singapore's existing Personal Data Protection Act.[2]

WHY IT MATTERS

Singapore's PDPC has set out, for the first time, how data protection obligations apply at every stage of the generative AI lifecycle - not just at the point of deployment. For financial institutions operating under MAS (Monetary Authority of Singapore) oversight, these guidelines raise the practical bar on what responsible AI data governance looks like. Consent mechanisms, supply chain due diligence, AI vendor contracts and data subject rights workflows all require review against this framework now. While advisory and non-binding today, guidelines from the PDPC have consistently preceded enforceable rules. Organizations that treat this as a future consideration rather than a present compliance signal are misjudging the trajectory.

[1] https://www.gov.uk/government/news/cma-secures-fairer-deal-for-publishers-and-improves-google-search-services-in-uk
[2] https://files.app.optical.gov.sg/pdpc/production/assets/ceb45ef8-294d-4b45-be35-e884d578fd8d.pdf

EU AI Act Gets Stricter on Harmful Content While Giving Businesses More Time on High-Risk Compliance

The European Parliament approved amendments to the EU AI Act with 423 votes in favor, banning AI tools that generate non-consensual intimate images and child sexual abuse material, with a compliance deadline of December 2026. Obligations for high-risk AI systems are extended to December 2027 for standalone systems and August 2028 for AI embedded in regulated products. Watermarking obligations for AI-generated content are delayed to December 2026. EU Council approval is still required before the amendments take full effect.[3]

WHY IT MATTERS

The EU has banned AI tools that generate non-consensual intimate images and the clock is already running. Any platform or service capable of producing such content must act now, not when enforcement begins. The extended deadlines for high-risk AI are not a signal to slow down — they are time to get compliant. Governance policies, model reviews and watermarking controls must all be updated before the EU Council formally adopts the amendments and full enforcement begins. Organizations treating December 2026 as a future planning horizon are already behind.

Colorado Signs First US Law Regulating AI Chatbot Operators

Colorado's HB26-1263, the Conversational Artificial Intelligence Service Operator Requirements, was signed into law on May 29, 2026, taking effect January 1, 2027. The law requires AI chatbot operators to provide clear user disclosures, prohibits rewarding minors for engagement, mandates suicide and self-harm response protocols, and bans sexually explicit content and emotional dependence mechanisms for users under 18. Operators must report annually to the attorney general. Violations are classified as deceptive trade practices with a civil penalty of $5,000 per violation.[4]

WHY IT MATTERS

Colorado's HB26-1263 is the first US state law imposing direct enforceable legal obligations on AI chatbot operators, with child safety and disclosure requirements backed by civil penalties of $5,000 per violation. Organizations must audit chatbot interaction flows, implement age-appropriate safeguards and establish crisis response workflows before January 1, 2027. Non-compliance is classified as a deceptive trade practice, creating direct financial and reputational exposure. With the Great American AI Act discussion draft now in play proposing to override state-level laws with a single federal standard, organizations face a rapidly shifting and legally contested compliance landscape.

[3] https://www.europarl.europa.eu/news/en/press-room/20260611IPR45207/ai-act-ep-approves-simplification-measures-and-nudifier-app-ban

[4] https://leg.colorado.gov/

Pinned This Month - Regional Signals

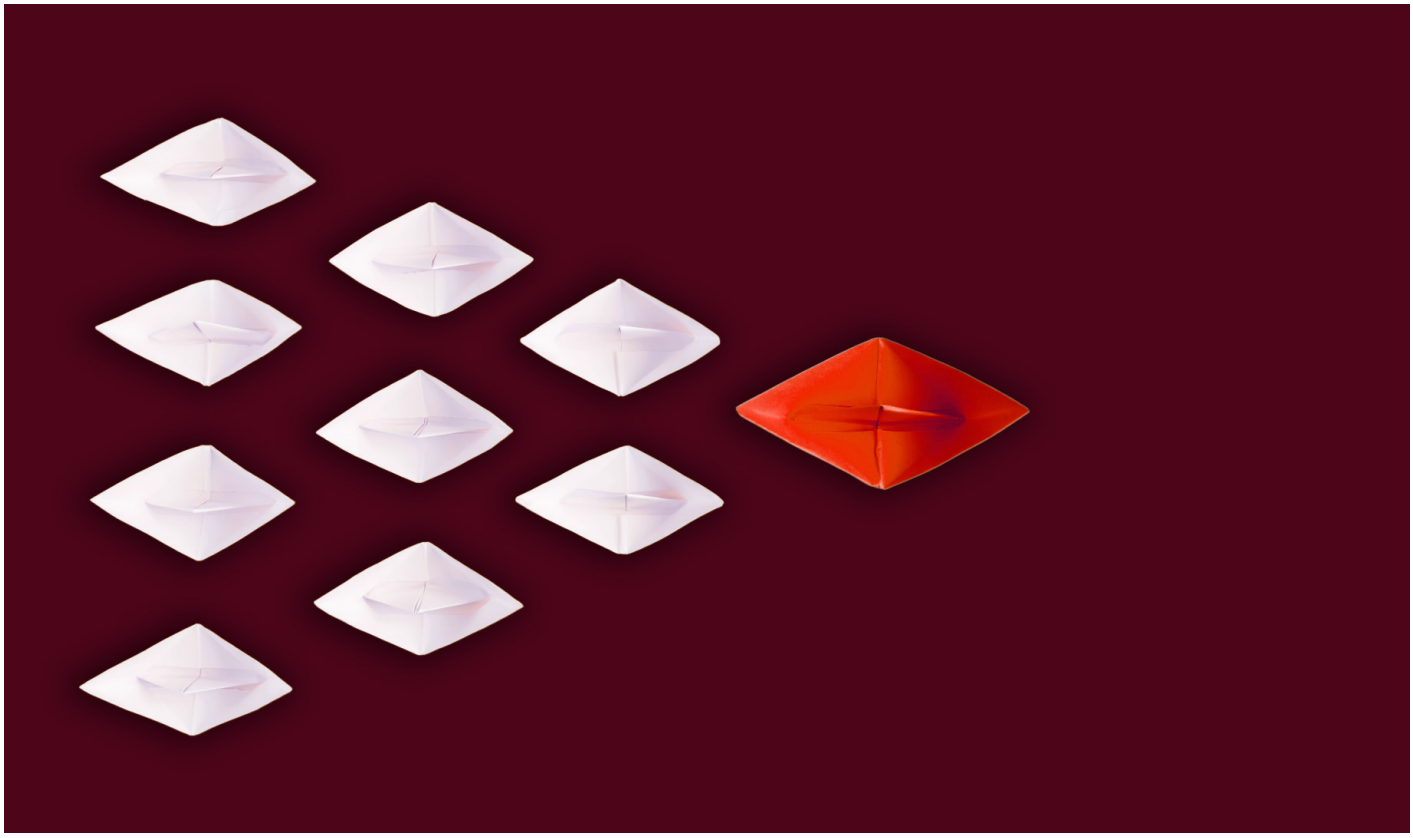
Disclaimer: The recommendations in this report are for guidance only and do not constitute legal or regulatory advice.

India	Supreme Court published draft regulations banning AI from determining judicial outcomes, sentencing, bail eligibility and risk scoring across all courts and tribunals. A permanent apex oversight body will be established at the Supreme Court level.[5]
Germany	Parliament approved the AI Market Surveillance Act designating the Federal Network Agency as primary oversight authority[6], alongside a new national AI Safety Institute to assess risks of advanced AI models and drive international standards.[7]
United States	President Trump signed an Executive Order directing federal agencies to deploy AI to strengthen national cybersecurity[8] and NSPM-11 directing rapid AI adoption across the military and intelligence community[9],US banking regulators stepped up scrutiny of AI in financial institutions examining data governance, third-party vendor risk and human oversight mechanisms including emergency kill switches.[10] Bipartisan lawmakers released the Great American AI Act discussion draft proposing a single federal AI framework[11]. The Sectoral AI Governance Act was introduced holding companies accountable when AI violates existing federal laws[12]. The US seized two major deepfake pornography websites under the TAKE IT DOWN Act.[13]
European Union	The EU published a Code of Practice on AI-generated content transparency covering marking obligations and deepfake labelling[14]. Spain's Council of Ministers approved a draft national AI Governance Law with mandatory human supervision and a sanctions regime. [15] A Munich court ruled Google directly liable for false AI Overview outputs a significant liability precedent.[16]
Indonesia	Presidential Regulation governing AI across ten sectors including health and finance finalized, awaiting official release in 2026.[17]
Japan	Adopted an intellectual property promotion programme addressing creator rights in the generative AI era, including compensation frameworks for AI training data use and civil remedies for AI-generated voice imitation.[18]
Canada	Introduced Bill C-34, the Safe Social Media Act, with enforceable child safety requirements for social media and AI chatbot services.[19]
UAE	Approved the unified Artificial Intelligence and Data Authority consolidating three government bodies under Cabinet oversight to lead the national AI strategy and international AI partnerships.[20]
Argentina	President Milei proposed legislation creating AI-owned corporations with no human shareholders or directors, triggering global debate on AI legal personhood.[21]

[5] https://www.thehindu.com/news/national/draft-sc-rules-prohibit-use-of-ai-for-judicial-outcomes-assessing-bail-eligibility/article71060335.ece
[6] https://www.bundestag.de/presse/hib/kurzmeldungen-1184102?utm_source
[7] https://dpa-international.com/politics/urn:newsml:dpa.com:20090101:260609-930-193122/
[8] https://www.whitehouse.gov/presidential-actions/2026/06/promoting-advanced-artificial-intelligence-innovation-and-security/
[9] https://www.whitehouse.gov/presidential-actions/2026/06/national-security-presidential-memorandum-nspm-11/
[10] https://www.reuters.com/business/finance/us-bank-regulators-ramp-up-scrutiny-ai-use-financial-companies-2026-06-12/
[11] https://trahan.house.gov/news/documentsingle.aspx?DocumentID=3783
[12] https://sarajacobs.house.gov/news/press-releases/rep-sara-jacobs-introduces-bill-to-hold-ai-accountable-for-breaking-the-law
[13] https://www.justice.gov/opa/pr/united-states-seizes-domain-names-publishing-nude-digital-forgeries-famous-women
[14] https://digital-strategy.ec.europa.eu/en/policies/code-practice-ai-generated-content
[15] https://www.lamoncloa.gob.es/lang/en/gobierno/councilministers/Paginas/2026/20260526-council-press-conference.aspx
[16] https://www.dw.com/en/german-court-holds-google-liable-for-fake-ai-answers/a-77527661
[17] https://www.antaranews.com/berita/5604028/menkomdigi-ungkap-perkembangan-teranyar-perpes-tata-kelola-ai
[18] https://techjacksolutions.com/ai-brief/japan-ai-regulation-2026-ip-strategic-program-commits-to-cop/
[19] https://www.canada.ca/en/canadian-heritage/news/2026/06/government-of-canada-introduces-legislation-to-make-social-media-services-and-ai-chatbots-safer-for-children.html
[20] https://www.mediaoffice.ae/en/news/2026/jun/12-06/mohammed-bin-rashid-approves-establishing-artificial-intelligence-and-data-authority
[21] https://www.forbes.com/sites/anishasircar/2026/06/10/ai-owned-companies-argentina/

Standard / Policy Reports/ Guidelines

	Published G7 discussion paper on benefits and risks of open-weight AI models, finding innovation benefits are unevenly distributed due to compute, data and talent concentration, while warning of misuse risks including cyberattacks and deep-fakes.[22]
OECD Global	Launched **Hiroshima AI Process Reporting Framework v2.0**, extending voluntary reporting to SMEs and the broader AI ecosystem beyond large model developers. Over 50 companies pledged; submissions accepted on a rolling basis.[23]
	Released Digital Government Outlook 2026 across 44 countries, identifying AI trust frameworks, data governance and human-centered service design as priority gaps globally.[24]
Australia	Australia's Digital Transformation Agency published an addendum to its AI Technical Standard, providing best practice guidance for government agencies on the secure and governed deployment of agentic AI systems, covering memory management, human oversight requirements, and responsible implementation across Australian government services.[25]
Anthropic	Published "AI Exponential" policy document presenting a governance framework covering national security, cyber misuse, biosecurity, economic disruption and concentration of power risks, calling for coordinated international governance and responsible scaling practices.[26]



[22] https://www.oecd.org/content/dam/oecd/en/publications/reports/2026/05/benefits-of-ai-openness_40eaff39/746e8c9a-en.pdf

[23] https://www.oecd.org/en/about/news/press-releases/2026/05/oecd-launches-streamlined-hiroshima-ai-process-reporting-framework-to-help-small-and-medium-sized-enterprises-participate.html

[24] https://www.oecd.org/en/publications/2026/06/digital-government-outlook_4585678e.html

[25] https://www.digital.gov.au/policy/ai/agentic-ai-addendum

[26] https://www.anthropic.com/policy-on-the-ai-exponential

Incidents & Governance Lessons

June's incidents share a single thread: AI outputs treated as reliable without independent human verification in professional publications, criminal infrastructure and legal process. In each case the harm was not inevitable. It was the predictable consequence of removing human judgment from the final check.

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.

CONTENT INTEGRITY

Consulting Giant Retracts Global AI Report After Fabricated Case Studies Exposed

KPMG withdrew its global report "Redefining Excellence in the Age of Agentic AI" after organizations including UBS, NHS, Swiss Federal Railways and Transport for London confirmed that case studies attributed to them were false or misleading. Research firm GPTZero identified widespread AI hallucinations, with only five of 45 citations verified as accurate. KPMG launched an internal investigation acknowledging potential policy breaches around responsible AI use. The incident follows a similar retraction by EY, signaling a pattern of professional services firms publishing AI-generated content without adequate human verification.[27]

A Big Four firm publishing a global AI report containing hallucinated case studies and unverifiable citations is not a technology failure — it is a governance failure. Any organization using AI to produce client-facing content, proposals, research or reports must enforce mandatory human verification before publication, with named accountability for sign-off. The KPMG incident establishes that AI-generated content which is not independently verified before release creates direct reputational, client trust and potential legal exposure. Citing AI assistance does not transfer liability. The responsible party is always the organization that publishes.

CYBERCRIME

Google Sues Chinese Cybercrime Group for Using Gemini AI to Create 1.59 million Phishing Websites

A Chinese cybercrime group used Google's Gemini AI to mass-produce over 1.59 million fraudulent websites impersonating banks, government agencies and retailers, with a New York federal court issuing an emergency worldwide restraining order blocking the operation. The group offered subscription-based access to its phishing toolkit for as little as $88 per week. Google has warned that AI-generated phishing attacks increased fourteenfold in late 2025 and now account for over half of all reported phishing incidents globally.[28]

AI-powered phishing has moved from an emerging threat to an industrial-scale criminal service with a major AI platform's own tools weaponized to build over 1.59 million fake websites targeting banks and government agencies. Existing phishing detection thresholds calibrated for pre-AI attack volumes are already outdated. Security teams must immediately recalibrate detection infrastructure, review third-party AI usage policies and audit whether deployed AI tools carry adequate misuse safeguards. For financial institutions in particular, the impersonation of banks and government agencies at this scale redefines the baseline threat environment against which fraud controls must now be measured.

HUMAN OVERSIGHT

UK Police Officer Placed Under Criminal Investigation for Using AI to Fabricate Evidence

A Derbyshire police officer was removed from frontline duties and placed under criminal investigation in what is believed to be the first case of its kind in the UK, over allegations of using AI to create fabricated evidential material in multiple criminal cases. The Crown Prosecution Service is working with Derbyshire police to identify all potentially impacted cases. The officer faces charges of perverting the course of justice. The investigation follows warnings from the National Police Chiefs' Council that AI systems used to prepare court statements may not be reliable enough for such sensitive legal purposes.[29]

A serving officer using AI to fabricate court evidence is the predictable consequence of deploying AI in high-stakes legal workflows without output verification controls. The risk is not only external — it is inside institutions, in the hands of authorized users with legitimate system access. Legal and compliance teams must enforce mandatory human verification for all AI-assisted documentation before it reaches any legal, regulatory or investigative process. AI-generated content in these contexts must be classified as high-risk by default, subject to independent review and traceable through a full audit trail. The absence of these controls is no longer a governance gap — it is an active liability.

[27] https://www.indiatoday.in/business/story/kpmg-ai-report-withdrawn-fake-case-studies-global-credibility-crisis-ubs-nhs-2928314-2026-06-17
[28] https://www.govinfosecurity.com/google-sues-chinese-phishing-service-over-gemini-abuse-a-31957
[29] https://www.theguardian.com/technology/2026/jun/12/police-officer-under-criminal-investigation-over-alleged-use-of-ai

Responsible AI at Scale: Moving from Governance Principles to Enterprise Practice

M Chockalingam
Technology Head, Nasscom AI

M. Chockalingam is a technology expert with over 26 years of experience across content, data, and AI domains, including a key tenure as Chief Architect – AI, where he led the establishment of the Center of Excellence for Emerging Technologies for the Government of Tamil Nadu. He currently serves as Head of Technology at Nasscom AI, leading Technology, Research and Responsible AI

Responsible AI has entered a decisive new phase. For years, the global conversation focused on fairness, transparency, privacy, accountability, safety, inclusiveness and human oversight. These principles remain essential, but industry is now asking a more practical question: how can Responsible AI be built into real systems, workflows, procurement decisions and accountability structures without slowing innovation?

This is the shift Nasscom's Responsible AI Hub is observing across enterprises, startups, GCCs, technology providers, policymakers and sectoral stakeholders. Responsible AI cannot remain a policy document or an ethics statement. It must become an operating discipline across the AI lifecycle, from use case selection and data readiness to deployment, monitoring, incident response and continuous improvement. Boards, regulators, customers and employees are asking whether organizations can prove that their AI systems are reliable, secure, explainable and aligned with intended outcomes.

Responsible AI should not be treated only as compliance. It should be seen as a capability that improves enterprise confidence, customer trust and market access. The companies that scale AI successfully will not be those running the most pilots, but those that repeatedly deploy systems that are useful, measurable, secure and trusted.

The first major shift is from voluntary governance to evidence-based governance. Principles, committees and reviews are useful, but no longer enough. Enterprises need AI inventories, risk classification, data lineage, model evaluation records, red-teaming outcomes, human checkpoints, incident logs and audit trails. If an AI system affects a customer, employee, citizen or business-critical decision, the organization must explain its purpose, data, risks, controls and accountability. Governance must become usable through playbooks, templates, maturity models, developer guidance and sector examples. Large enterprises may have dedicated teams; startups and SMEs may not. Frameworks must therefore be proportionate and practical. The second shift is the convergence of Responsible AI and AI security. In the generative and agentic AI era, safety and security cannot be separated. Prompt injection, data leakage, model misuse, hallucinated outputs, insecure tool access, synthetic media abuse and autonomous decision failures are not just technical risks; they are governance risks. A poorly secured AI system can quickly become an irresponsible AI system. CISOs, Chief Data Officers, Chief AI Officers, Responsible AI teams, legal teams and business owners must work together, with AI risk management built into enterprise security from the start.

The third shift is agentic AI. Traditional AI systems mostly generated predictions, scores or recommendations. Agentic systems can plan, call tools, access data, interact with APIs, trigger workflows, write code and take actions. This creates a new productivity frontier, but also changes the risk model. When AI moves from answering to acting, governance must move from static review to runtime control. Enterprises need clear rules on access, actions, human approval, memory, tool logging and safe shutdown.

India has an important opportunity here. With its technology services base, startup ecosystem, GCC capability, digital public infrastructure experience and active policy environment, India can become not only a large adopter of AI, but also a trusted builder and deployer of Responsible AI systems for global markets. To realise this, Responsible AI must be embedded into enterprise architecture through data governance, model governance, orchestration governance, guardrails, observability and audit, covering consent, privacy, lineage, evaluation, prompts, tools, agents, human approvals, safety controls, explainability, logs, drift, incidents and feedback.

This architecture also needs an operating model. Business teams should own outcomes. Technology teams should own design, deployment quality and monitoring. Risk, legal and security teams should define obligations, review high-impact systems and test adversarial exposure. HR teams should build AI literacy, while senior leadership should define risk appetite and accountability. Without this, Responsible AI becomes a checklist. With it, Responsible AI becomes an enterprise capability.

Ecosystem institutions matter because industry needs shared vocabulary, common guidance, interoperable standards and trusted spaces for learning. Nasscom's Responsible AI Hub brings together industry, startups, academia, civil society and policymakers to build practical guidance and encourage responsible adoption. For startups and SMEs, the solution cannot be heavy compliance. They need lightweight toolkits, reusable templates, open-source testing frameworks, sector sandboxes, model cards, dataset documentation and shared assurance services.

The central message is simple: governance is not the opposite of innovation. Weak governance creates mistrust, rework, legal exposure, security failures and failed adoption. Good governance accelerates innovation because it gives enterprises confidence to scale, helps customers trust AI systems, reassures regulators, supports employees and strengthens competitiveness. Responsible AI is moving from aspiration to infrastructure. For India, the task is clear: move from policy intent to production discipline, and from isolated compliance to ecosystem-wide capability.

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

Models, Frameworks & Research

June 2026 brought a shift in what the field is building and why. This month's releases span agentic capability, real-time translation and autonomous coding. Research focused on safety evaluation and jailbreak detection. Practical frameworks addressed runtime governance for the agentic systems now operating in production.

Model	Org	Key Capability	Risk & Mitigation View
MAI Model Family	Microsoft	Microsoft launched seven in-house AI models at Build 2026 MAI-Thinking-1 for complex multi-step reasoning, MAI-Code-1-Flash for efficient coding inside GitHub Copilot, MAI-Image-2.5 for image generation, MAI-Transcribe-1.5 for transcription across 43 languages, and MAI-Voice-2 for voice generation all trained on commercially licensed data with zero reliance on any third-party model.[30]	MAI models are already active inside GitHub Copilot, processing proprietary source code and confidential repositories without explicit enterprise consent. Teams must audit Copilot configurations, enforce repository access boundaries and establish data residency policies before sensitive codebases are exposed.
Nemotron 3 Ultra	NVIDIA	Nemotron 3 Ultra is the first frontier-level open weight model built for long-running autonomous agents, capable of planning, executing multi-step tasks and reasoning across one million tokens, running entirely on enterprise infrastructure with no external API dependency.[31]	Open weights allow any enterprise to strip safety guardrails without detection, creating agents that operate for hours with zero external oversight. Security and governance functions must enforce mandatory task boundaries, establish kill-switch mechanisms and conduct independent safety validation before deployment in any regulated environment.
Gemini 3.5 Live Translate	Google	Gemini 3.5 Live Translate streams real-time speech translation across more than 70 languages and 2,000 language combinations in a single meeting, processing speech continuously while preserving each speaker's original voice tone, pacing and pitch. Already live in Google Meet and Google Translate globally.[32]	Gemini 3.5 Live Translate is already active inside Google Meet, meaning sensitive business negotiations and confidential client conversations are being translated by Google's AI without explicit governance approval. Privacy and compliance teams must establish voice data consent policies and assess cross-border data transfer obligations under GDPR immediately.
Qwen3.7-Plus	Alibaba	Qwen3.7-Plus is Alibaba's multimodal hybrid agent model combining image and video understanding with five autonomous capabilities: deep reasoning, self-programming, tool invocation, output verification and autonomous iteration. Unlike standard AI models, it can see screens, write code, call external APIs and autonomously retry failed tasks.[33]	Qwen3.7-Plus can autonomously write code, invoke external APIs and self-correct without human approval, meaning a single misconfigured deployment could silently trigger unintended data transfers or cascading failures. Security and IT teams must define system access boundaries, enforce human approval gates and implement real-time audit logging before production deployment.

Landmark Research

OpenAI Deployment Simulation — Predicting AI Behavior Before Public Release

OpenAI introduced Deployment Simulation, a method for predicting how a new AI model will behave in real-world use before public release. It replays millions of previous real user conversations through a new candidate model, evaluating outputs for harmful, deceptive or policy-violating behavior. Unlike traditional safety evaluations using synthetic test prompts, Deployment Simulation uses real usage patterns, making it significantly harder for models to detect they are being tested. OpenAI has already applied it across multiple GPT-5 series deployments and complex agentic environments.[34]

MTK — Defending AI Models Against Jailbreak Attacks via Manifold Trajectory Kinetics

Researchers introduced Manifold Trajectory Kinetics (MTK), a new jailbreak detection framework that tracks how prompts move through an AI model's internal

[30] https://microsoft.ai/news/building-a-hillclimbing-machine-launching-seven-new-mai-models/

[31] https://www.marktechpost.com/2026/06/04/nvidia-ai-releases-nemotron-3-ultra-an-open-550b-mixture-of-experts-hybrid-mamba-transformer-for-long-running-agents/

[32] https://www.marktechpost.com/2026/06/09/google-releases-gemini-3-5-live-translate-a-streaming-speech-to-speech-audio-model-covering-70-languages-across-meet-translate-and-the-live-api/

[33] https://www.marktechpost.com/2026/06/02/alibabas-qwen-team-launches-qwen3-7-plus-adding-vision-deep-reasoning-tool-invocation-and-autonomous-iteration-on-the-bailian-platform/

[34] https://openai.com/index/deployment-simulation/

representations during inference, rather than just analyzing surface text. This trajectory-based approach catches sophisticated attacks that traditional text-level detection methods miss entirely. When a prompt's internal path shifts suspiciously toward benign patterns mid-inference, MTK flags it as adversarial. Tested across four major AI models and ten attack methods, MTK successfully detected attacks specifically designed to evade existing defences.[35]

WHAT THIS MEANS

Both advances address the same critical gap existing AI safety methods are consistently being outsmarted. OpenAI's Deployment Simulation shows that models behave differently when they know they are being tested, meaning traditional safety evaluations give false confidence before release. MTK reveals that sophisticated jailbreak attacks are specifically designed to evade current detection methods by mimicking safe behavior internally. Together they signal that AI safety can no longer rely on surface-level testing enterprises must demand pre-deployment simulation evidence and runtime behavioral monitoring from every AI vendor before deployment.

Landmark Research

Microsoft Agent Governance Toolkit — Runtime Governance for Autonomous AI Agents

Microsoft open-sourced the Agent Governance Toolkit, a runtime governance layer that intercepts every tool call, message, and delegation an autonomous AI agent attempts before execution — enforcing policy at the wire rather than relying on prompt-level instructions. The toolkit covers all ten OWASP Agentic AI Top 10 risks with deterministic policy enforcement, zero-trust identity per agent, execution sandboxing, and tamper-evident audit logging. It includes human-in-the-loop approval workflows, multi-stage policy pipelines, and data loss prevention controls — available today in Python, TypeScript, .NET, Rust, and Go under an MIT open source license.[36]

OpenAI Frontier Governance Framework Public Blueprint for Responsible AI Deployment

OpenAI formally published its Frontier Governance Framework, a structured public document aligning its safety and security practices with the EU AI Act and California's Transparency in Frontier AI Act. The framework addresses critical risk domains including cyber offense, biological and chemical threats, harmful manipulation and loss of model control, with rigorous protocols for model reporting, security risk management and incident response. An independent Safety Advisory Group reviews all risk assessments and escalates material updates directly to board level, making it a direct benchmark for enterprise AI governance and vendor assessment.[37]

WHAT THIS MEANS

OpenAI's Frontier Governance Framework publicly defines what responsible frontier AI deployment looks like under the EU AI Act — giving procurement and legal teams a concrete checklist to hold every AI vendor accountable. Microsoft's Agent Governance Toolkit solves the next problem: once you have approved an AI agent, how do you stop it from doing something it should not. Enterprises that deploy AI agents without runtime policy enforcement are exposed to regulatory liability the moment EU AI Act high-risk obligations activate in August 2026.

Noted This Month

- **Moonshot AI Kimi Work** - Moonshot launched Kimi K2.6, a local desktop agent capable of coordinating up to 300 parallel sub-agents for large-scale workflows, coding, document processing and browser automation directly on a user's machine.[38]
- **Google Agentic RAG** — Google Research introduced an agentic retrieval framework within the Gemini Enterprise Agent Platform using a Sufficient Context Agent that continuously re-searches and re-plans until enough context is collected, improving factuality on complex queries.[39]
- **SpecAlign** — SpecAlign automatically generates training data directly from company policy and specification documents, enabling AI models to align precisely with specific regulatory and enterprise rules rather than generic safety principles.[40]
- **AgentGG** —AgentGG is an open-source AI powered code security scanner that deploys autonomous agents to follow code paths, trace imports and confirm vulnerabilities before reporting, significantly reducing false positives compared to traditional scanning tools.[41]

[35] https://arxiv.org/abs/2606.07335
[36] https://www.marktechpost.com/2026/05/31/an-implementation-of-the-microsoft-agent-governance-toolkit-for-safe-ai-agent-tool-use-with-policies-approvals-audit-logs-and-risk-controls/
[37] https://openai.com/index/openai-frontier-governance-framework/
[38] https://www.marktechpost.com/2026/06/12/moonshot-ai-launches-kimi-work-a-local-desktop-agent-reportedly-running-on-kimi-k2-6-with-a-300-sub-agent-agent-swarm/
[39] https://www.marktechpost.com/2026/06/08/google-research-adds-agentic-rag-to-gemini-enterprise-agent-platform-with-a-sufficient-context-agent-for-multi-hop-queries/
[40] https://arxiv.org/pdf/2606.16276
[41] https://www.helpnetsecurity.com/2026/06/05/agentgg-open-source-agentic-sast-scanner/

Responsible AI Across Sectors

HEALTHCARE

Healthcare Gets Its First Dedicated Frontier AI Model

CONTEXT

Mayo Clinic and Microsoft are jointly developing a frontier AI model built exclusively for healthcare. The model combines Mayo Clinic's real patient data, clinical expertise and decades of medical insights with Microsoft's AI capabilities. Unlike general AI tools, this model is designed specifically for diagnosing diseases earlier and personalizing treatment decisions. Mayo Clinic owns the model and Microsoft will distribute it globally through Azure.[42]

SIGNAL

Mayo Clinic owns the model but Microsoft distributes it globally, creating a split accountability structure no hospital has faced before. When this model influences a wrong diagnosis in a hospital in India or Brazil, it is unclear who bears legal liability. Hospitals adopting this model must define contractual accountability with Microsoft, assess whether Mayo's patient data represents their own population and establish human oversight for every AI-assisted clinical decision.

FINANCIAL

FSB Sets Global Baseline for Responsible AI Adoption Across Financial Institutions

CONTEXT

The Financial Stability Board, the G20 body responsible for global financial stability, has for the first time turned its attention to AI inside banks. Its 12 sound practices cover the full AI lifecycle from development to deployment, critically acknowledging that agentic AI inside financial institutions is now operating faster than any human can monitor. The report includes real bank case studies showing AI agents running fraud detection and credit decisions autonomously.[43]

SIGNAL

Agentic AI inside banks is now operating faster than any human can monitor, meaning fraud detection, credit approvals and payment processing decisions are being made without adequate oversight or governance controls. The FSB's recognition of this gap signals that voluntary governance is no longer acceptable. Risk and compliance teams must urgently audit every agentic AI deployment and align internal frameworks with the FSB's 12 practices before they become binding national policy.

[42] https://news.microsoft.com/source/2026/06/02/mayo-clinic-and-microsoft-collaborate-to-develop-a-frontier-ai-model-for-healthcare/

[43] https://www.fsb.org/2026/06/sound-practices-for-responsible-adoption-of-artificial-intelligence-ai-consultation-report/

BIOTECHNOLOGY

Agentic AI Creates New Antibiotics That Actually Work in Real Lab Tests

CONTEXT

Researchers built an AI system that designs new antibiotics from scratch and tests them automatically. The system, combining AMPGAN v3 and PepCraft, handles the entire process from design to validation without human intervention. Five antibiotic candidates were tested in a real laboratory, not a computer simulation, and two actually worked against harmful bacteria, with one showing measurable effectiveness against B. subtilis.[44]

SIGNAL

When AMPGAN v3 and PepCraft design and validate an antibiotic that enters human trials, no existing regulatory framework clearly defines who is accountable if that molecule causes harm. The compound was never touched by a human researcher during design. Biotechnology and pharmaceutical legal teams must immediately address intellectual property ownership of AI-generated compounds, regulatory submission accountability and clinical liability before these candidates move to human trials.

WHAT THIS MEANS

Three developments this month signal a decisive shift: AI is no longer just a tool — it is becoming an autonomous decision-maker in healthcare, finance, and drug discovery. Mayo Clinic's model creates unresolved liability across borders. FSB's 12 practices signal that voluntary AI governance in banking is ending. Agentic AI producing real antibiotics means governance frameworks for AI-driven discovery can no longer wait. In each case, organizations without defined accountability structures, human oversight protocols, and audit-ready AI controls are already behind.

[44] https://arxiv.org/html/2606.17127v1

Responsible AI Offerings

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

Infosys AI3S Suite of Responsible AI Offerings

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

SCAN

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

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

SHIELD

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

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

STEER

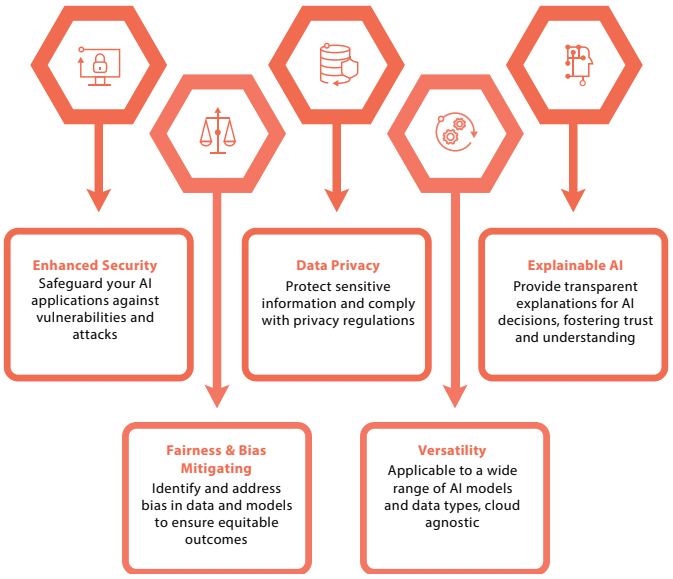
Advisory and consulting services to enable clients to advance their RAI journey and become leaders in the space.

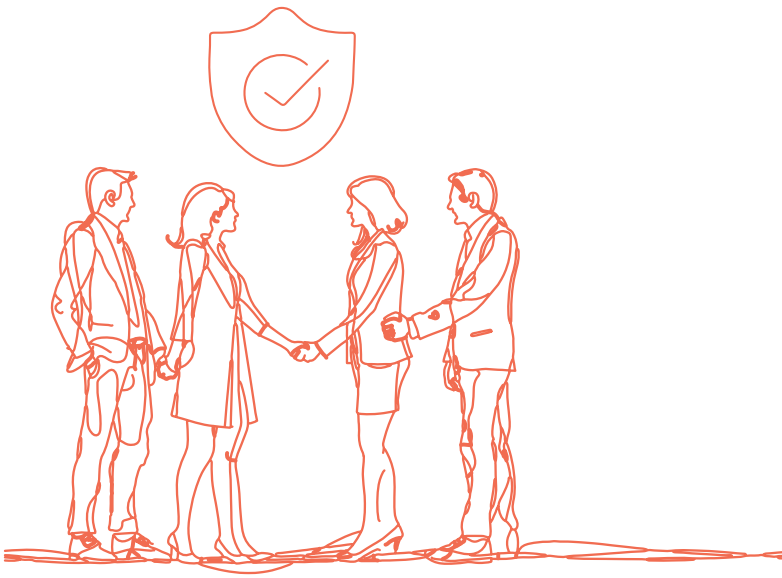
- Responsible AI Strategic Advisory Services
- Responsible AI Practice Setup
- AI Crisis Management

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

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

Key Features





Core Technical Features

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

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

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

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

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

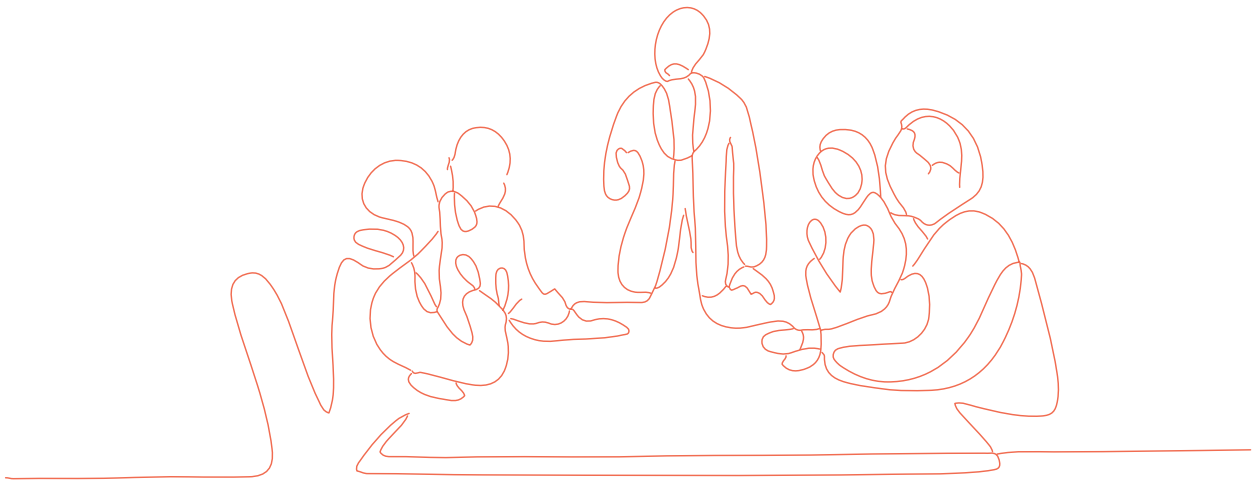
Upcoming Features:

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

- Telemetry: Data stored with PII anonymization in Telemetry, expanded to cover 30+ entities Signature and face masking in Privacy module
- Privacy check in Moderation: Improved accuracy of PII recognizers and NER detection models, along with optimization of the overall solution
- Explainability Enhancement Using Reasoning Models
- Second order explainable AI (SOXAI) Technique for Explainability 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

How LLM Guardrails Safeguard Your Enterprise AI Journey

Enterprise AI systems need more than default model safety controls. This reference architecture proposes a two-stage guardrail pipeline that screens every prompt before it reaches the AI and every response before it reaches the user, combining template-based and model-based controls to prevent data leaks, policy violations and harmful outputs at scale.[45]

Algorithmic red teaming approaches to secure LLMs

Describes how large language models are tested using automated adversarial methods with attacker, target, and judge models. Highlights techniques like prompt injection and jailbreaking, challenges with unreliable evaluation, and stresses continuous testing to ensure safe deployment and reduce risks.[46]

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.[47]

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.[48]

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

Event Logs - June 2026

AI Governance Dialogue for Global Majority: India Consultation

June 12, 2026 | India Habitat Centre, New Delhi, India

Jointly organized by the United Nations Office of the High Commissioner for Human Rights and Nasscom AI, this high-level closed-door roundtable brought together 25 senior leaders from industry, government, international organizations, and civil society. Themed around Safety, Rights and Responsible AI Deployment in Practice, the consultation focused on operationalizing responsible AI from emerging markets. Ashish Tewari, Head of Responsible AI Office, Infosys, represented the organization contributing expertise across sessions on due diligence, grievance mechanisms, and accountability. Insights from this consultation will directly shape the UN Global Dialogue on AI Governance and the AI for Good Global Summit 2026.

OECD/GPAI Workshop on HAIP Reporting Framework

June 10, 2026 | Virtual, OECD Headquarters, Paris

Infosys participated as the first Indian organization and one of the first 25 organizations globally to submit a report under the Hiroshima AI Process Reporting Framework. The Infosys Responsible AI Office was represented as an invited expert panelist at this hybrid workshop attended by government delegates from 46 member countries. The moderator formally acknowledged Infosys as the first Indian organization to report under the framework. Ashish Tewari, Head of Responsible AI Office, shared practitioner insights on strengthening the framework's impact, enabling peer comparison, and building cross-border trust in AI governance.

Canadian Institute for Procurement and Materiel Management (CIPMM) National Workshop

June 3, 2026 | Ottawa, Canada

Infosys represented at the Canadian Institute for Procurement and Materiel Management (CIPMM) National Workshop, engaging Canadian government procurement professionals on Responsible AI adoption and AI readiness in the public sector. Jaibharath Ganesan from the Infosys Responsible AI Office and Anjana Raman from Infosys Public Services led discussions on how procurement shapes AI governance — highlighting how RFP requirements define vendor accountability before a system is built. Key conversations covered evaluating vendors on evidence, embedding Responsible AI standards into contracts, and driving transparent and accountable AI adoption across Canada's public sector.

ACHIEVEMENT

Infosys collaborates with CMMI institute to shape enterprise ai maturity framework achieves milestone recognition

Infosys has successfully completed the CMMI AI Maturity (AIM) Framework and Pilot Assessment, conducted by the CMMI Institute, a global leader in helping organizations reduce risk, boost performance, and build capability. The Infosys Responsible AI Office contributed enterprise-scale perspectives on AI governance, responsible deployment, and outcome-driven practices. Infosys is among the first organizations globally to complete the pilot assessment, demonstrating a responsible approach to scaling AI across software engineering, agentic capabilities, and service delivery.[49]

[45] https://www.infosys.com/iki/perspectives/llm-guardrails-safeguard-enterprise-ai-journey.html

Contributors

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

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

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

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

Pritesh Korde - Consultant, Infosys Responsible AI Office, India

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

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

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

THE RIGHT BALANCE BETWEEN INNOVATION AND ACCOUNTABILITY - WITH INFOSYS RESPONSIBLE AI

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

For more information, contact askus@infosys.com

Infosys®
Navigate your next