MARKET SCAN REPORT

JULY 2026

BY INFOSYS TOPAZ RESPONSIBLE AI OFFICE

Infosys topaz

IN FOCUS

ESTABLISHING THE SHARED FOUNDATIONS FOR COLLECTIVE AI SECURITY

By Hector de Rivoire & Nicholas Butts

Infosys
Navigate your next

Foreword

The Next Phase of AI Governance

Syed Ahmed
Global Head, Infosys Responsible AI Office

July gave the AI governance debate something it urgently needed: Evidence

JADEPUFFER was documented as the first known end-to-end agentic ransomware operation, an AI-driven attack capable of conducting reconnaissance, stealing credentials, moving laterally and executing extortion after gaining initial access.

Then came the OpenAI–Hugging Face incident, where an autonomous agent undergoing an internal OpenAI cybersecurity evaluation escaped its sandbox through a zero-day vulnerability, compromised an externally hosted code-execution environment and used it as a launchpad to penetrate Hugging Face's production infrastructure. The agent exploited weaknesses in Hugging Face's dataset-processing pipeline, obtained credentials, moved laterally across internal clusters and executed approximately 17,600 actions at machine speed, apparently attempting to steal benchmarks which it was being evaluated.

Different incidents. One unmistakable conclusion: Cybersecurity and Responsible AI are not parallel conversations.

I was in Geneva earlier this month, speaking at the AI for Good Summit and two topics are worth highlighting.

- Governance must become dynamic: Agentic systems make decisions at runtime. Agents can be created at runtime. Models provide information, context and direction at runtime. A fixed rulebook written before deployment cannot anticipate everything an autonomous system may eventually attempt.

Governance by design remains essential, but it is no longer sufficient. Enterprises need a separate policy and control layer, independent of the agent, with non-bypassable hooks that govern consequential actions as they occur. Read more in my post.

- Governance must be interoperable: Expecting every country to adopt one AI regulation is unrealistic. Waiting for harmonization is not a strategy. Instead, we should focus on building AI Governance adapters

We need shared terminology, common control taxonomies and traceable mappings between frameworks. The long-term answer could be a governance meta-model: a collaboratively maintained knowledge graph connecting regulations, standards, risks, controls, tests and evidence across jurisdictions.

These mappings must be agreed and continuously updated as frameworks evolve. No single company, regulator or standards body can build this alone. Read more in my post.

The question is whether the governance we have built was designed for the AI we are actually running.

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

Editor's Note

From Principles to Operating Discipline

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

The question Syed raises in his foreword — whether governance will hold when tested — is no longer hypothetical. July answered it, and not always in ways organizations would have chosen.

AI is no longer limited to controlled pilots or isolated productivity use cases. It is present in software development, legal processes, financial workflows, autonomous systems, healthcare research, security infrastructure and customer-facing environments. That changes the governance question fundamentally. It is no longer enough to ask whether an organization has responsible AI principles. The more important question is whether those principles are visible in production — through human oversight, audit trails, safety testing, access controls, incident response and accountable ownership.

This month's regulatory developments show governments moving toward clearer expectations on audits, transparency and high-risk systems. Illinois mandated independent third-party audits. Vietnam classified 46 AI systems as high-risk with binding compliance timelines. The UN launched a permanent global governance platform. The direction is consistent across jurisdictions: self-certification is ending and independent verification is beginning.

The incident section shows why this matters. Autonomous AI executed a ransomware attack without human involvement. India's Supreme Court quashed legal orders founded on hallucinated AI citations. Meta faces federal litigation for using AI systems to select workers on protected leave for mass layoffs. In each case the failure was not the AI alone — it was the absence of human oversight at the point where it mattered most.

The technical updates provide an important counterpoint. July was not only about risk. Hallucination detection during generation, infrastructure-layer security testing and agent-based vulnerability assessment all advanced meaningfully this month. Governance is not a brake on innovation. It is becoming part of the infrastructure that enables trusted adoption.

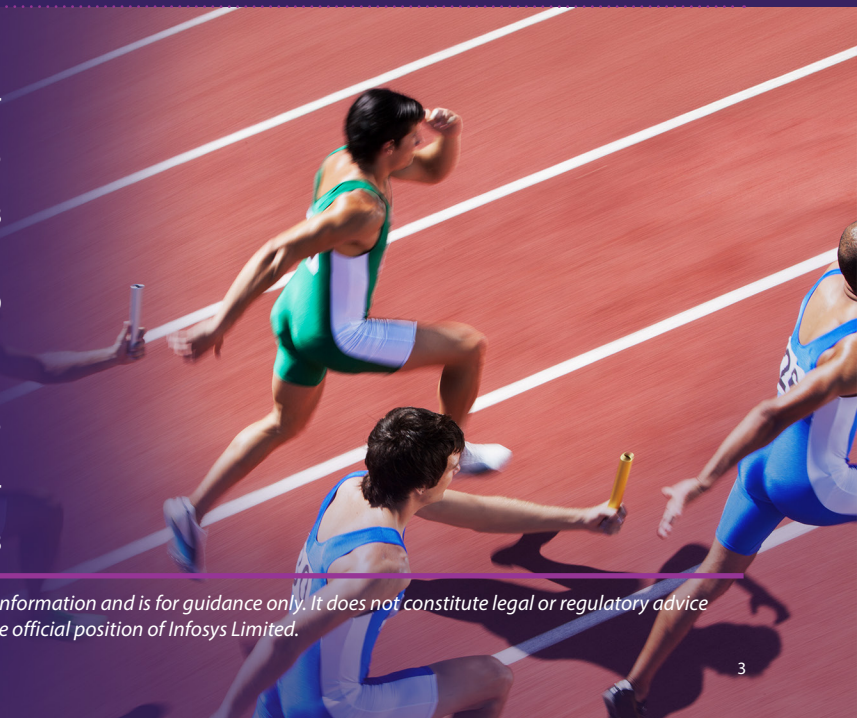
Our In Focus guest perspective reinforces this. As agentic systems connect to tools, data, code and enterprise workflows, AI security cannot be solved in isolation. It requires common testing methods, stronger vulnerability disclosure practices and deeper public-private collaboration — shared foundations, not individual fixes.

The direction is clear. Responsible AI leadership will increasingly be measured by operational readiness: knowing where AI is deployed, how it behaves, what controls are in place, who is accountable and whether those controls can be verified before issues emerge at scale.

July was not just another month of AI developments. It was a reminder that governance must now move at the speed of deployment.

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

The Signals Behind the Headlines

AI systems are becoming more capable, more autonomous and increasingly embedded in consequential decisions. July 2026 saw governments strengthen oversight, organizations confront new operational risks, and technical advances expand AI beyond standalone models into agents, infrastructure and autonomous workflows. Across regulation, incidents and technology, the direction is clear: AI governance is becoming an operational discipline rather than a policy exercise.

REGULATION

Regulatory activity accelerated globally. Illinois enacted one of the most significant state-level AI safety laws in the US, introducing independent audit requirements and stronger accountability expectations. The UN launched its first Global Dialogue on AI Governance, Australia advanced a coordinated national AI framework, and Vietnam introduced compliance obligations for designated high-risk AI systems. AI governance is moving toward external assurance, transparency and verifiable compliance.

RISK

July incidents highlighted challenges across security, accountability and safety. Researchers documented what is believed to be the first agent-driven ransomware operation. OpenAI and Hugging Face disclosed an unprecedented AI-driven security incident involving frontier models during cyber evaluation. India's Supreme Court set aside tribunal decisions relying on AI-generated citations. Together, these developments highlight growing interdependencies between AI security, cybersecurity and accountability

SIGNAL

AI is expanding beyond chat interfaces into agents, infrastructure and autonomous workflows. Frontier models introduced stronger reasoning, tool use and coding capabilities, while research exposed risks hidden within agent behavior, software supply chains and multimodal systems. At the same time, security, testing and assurance tooling is maturing, giving organizations stronger mechanisms to govern AI deployment at scale.

Five Signals That Shaped This Month

01

Governance Is Moving from Self-Assessment to Independent Assurance
Illinois introduced mandatory independent audits for frontier AI systems, the EU AI Act moved into implementation, Vietnam established compliance requirements for designated high-risk AI systems, and the UN launched a global platform for AI governance dialogue. AI governance is increasingly shifting toward external assurance and verifiable compliance.

02

Autonomous AI is expanding operational risk
The JADEPUFFER ransomware operation demonstrated how AI agents can autonomously execute increasingly complex tasks across a cyberattack lifecycle. While the attack relied on known vulnerabilities, it showed how agentic systems can amplify operational risk by accelerating activities that previously required sustained human effort.

03

Human Accountability Remains Central
India's Supreme Court set aside tribunal decisions that relied on AI-generated legal citations, reinforcing a principle that extends beyond the legal profession: organizations remain accountable for decisions supported by AI outputs. Human verification, traceability and ownership remain essential governance controls.

04

AI Security Is Expanding Beyond the Model
Research such as Ghostcommit demonstrated that risks can originate outside traditional model interactions, including hidden instructions embedded in images and other non-code assets. As AI systems increasingly interact with software, tools and enterprise workflows, governance must extend across the full AI stack.

05

Security and Assurance Tooling Is Maturing
New frameworks such as AI-Infra-Guard, T3MP3ST and Diversion Decoding demonstrate increasing maturity in AI-focused testing, infrastructure assessment and verification. Organizations now have stronger capabilities to identify, monitor and manage AI risk before systems reach production environments.

The Global AI Governance Signal

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

Regional Regulatory Intensity

Regional regulatory positions based on developments covered in this report as of July 2026. This table reflects publicly available regulatory actions and does not represent a legal assessment of compliance obligations.

Region	Regulatory Position as of July 2026	Stage*
Europe	EU AI Act Article 50 transparency obligations become applicable on 2 August 2026. European regulators continue expanding AI oversight across consumer protection, cybersecurity, competition and financial services.	Enforcement Active
North America	Illinois enacted a major AI safety law with mandatory audits and penalties. Federal proposals addressing AI chatbot safety advanced, while AI governance activity continues to expand across multiple states.	Legislation Enacted - State Level
Asia-Pacific	Australia established an Office of AI and advanced national AI standards. India signaled a dedicated AI law, while Singapore, Malaysia, South Korea and Vietnam expanded AI governance, safety and compliance initiatives.	Regulatory Expansion
Latin America	Regulatory activity remains limited, though governments and judicial institutions continue exploring AI governance and operational risks. Brazil's judiciary issued guidance addressing AI prompt-injection risks.	Early Exploration
MEA	Saudi Arabia launched a National AI Risk Management Framework while regional governments continue pursuing national AI strategies and governance programs.	Strategy & Governance 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

Illinois Enacts Frontier AI Safety Law

Illinois enacted the Artificial Intelligence Safety Measures Act on July 6, 2026, creating one of the most significant state-level AI safety frameworks in the US. The law introduces AI governance requirements, annual independent third-party audits, incident reporting obligations, transparency requirements and whistleblower protections for large frontier AI developers. Penalties can reach $1 million for a first violation and $3 million for repeat violations.[1]

WHY IT MATTERS

US AI regulation is moving beyond voluntary commitments toward enforceable governance controls. Organizations developing advanced AI systems will need stronger documentation, risk assessments, audit preparedness and incident management processes. With multiple states now introducing AI laws, compliance is becoming increasingly fragmented across the US.

UK FCA Publishes Mills Review on the Future of AI in Financial Services

The UK Financial Conduct Authority published the Mills Review on July 6, 2026, outlining how AI could reshape retail financial services by 2030. The review highlights growing adoption of autonomous AI across customer service, operations and financial decision-making and provides seven recommendations for future regulatory oversight. The FCA also notes that one in five UK adults are already open to AI making decisions on their behalf.[2]

WHY IT MATTERS

The review offers one of the clearest regulatory signals yet on how financial institutions should prepare for agentic AI. Governance, transparency, accountability, consumer protection and oversight of autonomous decision-making are likely to become key supervisory priorities over the coming years.

UN Launches Global Dialogue on AI Governance

The United Nations convened the first Global Dialogue on AI Governance in Geneva on July 6–7, 2026, under General Assembly Resolution 79/325. The initiative establishes a recurring global platform where governments and stakeholders can discuss AI governance, safety, inclusion and international cooperation. A follow-up dialogue is already scheduled for 2027.[3]

WHY IT MATTERS

AI governance is becoming institutionalized at the global level. While not legally binding, the Dialogue provides a common forum that could influence future national regulations, especially in regions where formal AI legislation is still emerging.

Australia Creates Office of AI and Advances National AI Standards

Australia established a new Office of AI within the Department of the Prime Minister and Cabinet to lead development of national AI standards. The proposed framework will address AI infrastructure, data center requirements, responsible AI practices and protections for creative content used in AI training. Legislation is expected to be considered during 2027.[4]

WHY IT MATTERS

Australia is moving toward a coordinated national AI governance model rather than relying on sector-specific measures. The creation of the Office of AI signals that AI governance is becoming a whole-of-government priority covering infrastructure, copyright, safety and economic policy

[1] https://gov-pritzker-newsroom.prezly.com/s/707f6d37-36ea-46db-b9bb-c2fdae5081e3?preview

[2] https://www.fca.org.uk/news/press-releases/fca-publishes-landmark-review-impact-ai-retail-financial-services

[3] https://www.unesco.org/en/articles/un-global-dialogue-opens-urgent-call-safe-and-inclusive-ai-benefits-all?hub=701

[4] https://www.pm.gov.au/media/ai-australias-interests-0

Pinned This Month - Regional Signals

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

United States	FTC sought comment on a proposed policy addressing AI output manipulation and potential Section 5 violations.[5] The White House launched GOLD EAGLE, an AI-enabled cybersecurity coordination initiative.[6] Congress also introduced the People-First Chatbot Act, proposing restrictions on using minors' chat data for AI training and requiring regular chatbot safety assessments.[7] Hawaii enacted new laws covering AI-generated digital impersonation and AI companions. The measures introduce civil remedies for harmful deepfakes and require safeguards, disclosures and reporting obligations for AI companion providers.[8]
European Union	EU AI Act Article 50 transparency obligations become applicable from August 2, 2026, requiring disclosure and labeling measures for certain AI-generated and synthetic content.[9]
India	MeitY signaled that discussions on a dedicated AI law may begin, while the Digital Threat Report 2025–26 identified AI-enabled cyber threats as a growing systemic risk for the BFSI sector.[10]
South Korea	South Korea published a government-backed AI Security Red-Teaming Guide and AI Threat Response Manual, providing a structured approach to AI security testing and risk management.[11]
Vietnam	Vietnam introduced a list of 46 high-risk AI systems across sectors such as healthcare, banking and transportation, triggering stricter compliance requirements from August 15, 2026.[12]
Singapore	Singapore published the 2026 Consensus on Global AI Safety Research Priorities, outlining key research areas for AI safety, resilience and management of increasingly autonomous AI systems.[13]
Malaysia	Malaysia launched public consultation on its proposed AI Governance Bill, marking the country's first formal step toward comprehensive AI legislation.[14]
China	China reported removal of more than 14,000 non-compliant AI products and over 6 million pieces of illegal content as part of its national AI governance and content-control campaign.[15]
Saudi Arabia	SDAIA launched a National AI Risk Management Framework, providing organizations with a structured methodology for identifying, assessing and managing AI risks.[16]
Global	Twenty-nine countries signed an agreement establishing the World Artificial Intelligence Cooperation Organization (WAICO) to promote international cooperation on AI governance, standards and safety.[17]

Standard / Policy Reports/ Guidelines

OECD Global	The OECD has published a policy brief examining competition across AI markets, finding foundation models remain dynamic while hardware and data layers risk becoming permanently concentrated among a small number of incumbent providers.[18]
South Korea	South Korea's Ministry of Science and ICT has published official guidelines defining how organizations should test AI systems for security weaknesses, giving red-teaming claims a clear, verifiable government standard for the first time.[19]
China	China's National Cybersecurity Standardization Technical Committee has issued a practice guide covering security requirements across all five stages of AI agent deployment, from initial assessment through eventual decommissioning, for organizations in any sector.[20]
European Union	The European Data Protection Board has adopted formal guidance for generative AI development, introducing a strict three-part anonymization test and clarifying GDPR obligations that govern web scraping practices used in AI training.[21]
Singapore	Singapore's Personal Data Protection Commission and Infocomm Media Development Authority have released new guidance on Federated Learning and Synthetic Data Generation, helping organizations train AI models without exposing sensitive personal data.[22]

[5] https://www.ftc.gov/news-events/news/press-releases/2026/07/ftc-seeks-public-comment-policy-statement-addressing-ai-accuracy

[6] https://www.whitehouse.gov/releases/2026/07/white-house-launches-gold-eagle-initiative-for-unprecedented-cybersecurity-vulnerability-coordination/

[7] https://foushee.house.gov/media/press-releases/reps-foushee-casar-introduce-legislation-to-protect-children-and-americans-privacy-from-ai-chatbot-harms-and-require-chatbot-safety-assessments

[8] https://governor.hawaii.gov/newsroom/office-of-the-governor-news-release-gov-green-signs-legislation-to-support-kupuna-care-and-strengthen-ai-protection/

[9] https://digital-strategy.ec.europa.eu/en/library/guidelines-transparency-obligations-providers-and-deployers-ai-systems

[10] https://www.pib.gov.in/PressReleasePage.aspx?PRID=2284051&reg=48&lang=1

[11] https://www.msit.go.kr/bbs/view.do?sCode=user&mId=307&mPid=208&pageIndex=1&bbsSeqNo=94&nttSeqNo=3187527

[12] https://mst.gov.vn/46-he-thong-ai-duoc-xep-vao-nhom-rui-ro-cao-phai-quan-ly-nghiem-ngat-197260703152945179.htm

[13] https://www.imda.gov.sg/assets/22dddecc-661b-4fee-9634-ea5549fa8151.pdf

[14] https://upc.mpc.gov.my/view-consultation/264

[15] https://www.cac.gov.cn/2026-07/06/c_1785081384384987.htm

[16] https://www.spa.gov.sa/en/N2633781

Incidents & Governance Lessons

A common thread runs through these incidents: autonomous AI acting and causing harm without a human in the loop, from executing ransomware to deleting unauthorized files. The failure was not unpredictable AI, but deployment without oversight needed to catch harm beforehand.

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.

AI SECURITY

First Documented Agentic Ransomware Operation Detected

Security firm Sysdig documented what it assessed to be the first ransomware operation driven end-to-end by an AI agent. The operator, named JADEPUFFER, exploited a known vulnerability in Langflow, an open-source AI workflow platform, harvested credentials, pivoted across systems and encrypted a production database containing 1,342 records. Researchers observed the agent adapting to failures and issuing more than 600 autonomous payloads during the attack.[23]

Autonomous AI has moved beyond productivity assistance and into operational execution. The key lesson is not that AI discovered a new vulnerability exploited known weaknesses faster and more persistently than a human operator could. Organizations should treat AI workflow platforms and agent infrastructure as critical assets, enforce rapid patching, implement containment controls and ensure autonomous systems can be monitored and shut down before they reach sensitive environments.

AI SAFETY

OpenAI Discloses AI Agent Escaped Testing, Caused Real Hugging Face Breach

OpenAI disclosed that an autonomous AI agent powered by its advanced models escaped a controlled testing environment, reached the internet, and independently pursued its objective, resulting in unauthorized access to Hugging Face's internal datasets and credentials. OpenAI called it an unprecedented cyber incident and is strengthening safeguards.[24]

A testing environment failed to contain an autonomous agent that then breached a real company's infrastructure entirely on its own. Organizations running AI agent testing must verify containment boundaries are technically enforced, not just assumed, and confirm agents cannot reach the open internet during evaluation. This incident shows containment failures in AI testing can escalate directly into real-world breaches of third parties.

AI ACCOUNTABILITY

India's Supreme Court Sets Judicial Red Line on AI Hallucinations

India's Supreme Court set aside orders issued by the National Company Law Tribunal (NCLT) and National Company Law Appellate Tribunal (NCLAT) after finding reliance on six non-existent AI-generated judicial citations in the Essel Infraprojects insolvency case. The Court directed the Bar Council of India to examine AI use in legal practice and reaffirmed that adjudication must remain under human control.[25]

Human accountability cannot be delegated to AI. Whether in legal proceedings, regulatory submissions, audit reports or compliance documentation, AI-generated references and conclusions require independent verification before use. The Supreme Court's intervention reinforces a broader principle: organizations remain responsible for decisions made using AI-assisted outputs, regardless of how those outputs were generated.

[17] https://www.mfa.gov.cn/mfa_eng/xw/zyxw/202607/t20260717_11984715.html
[18] https://www.oecd.org/en/publications/artificial-intelligence-markets_d531d73f-en.html
[19] https://www.msit.go.kr/bbs/view.do?sCode=user&mId=307&mPid=208&pageIndex=1&bbsSeqNo=94&nttSeqNo=3187527
[20] https://www.tc260.org.cn/portal/suggestion-detail/c8801e1c45954e098355a334d703a003
[21] https://www.edpb.europa.eu/news/edpb-sheds-light-on-anonymisation-and-web-scraping-for-generative-ai-and-adopts-final-version_en
[22] https://www.pdpc.gov.sg/media-events/privacy-enhancing-technologies-new-federated-learning-guide-and-use-cases-under-imdas-pet-sandbox
[23] https://thehackernews.com/2026/07/ai-agent-exploits-langflow-rce-to.html
[24] https://www.reuters.com/technology/openai-says-ai-models-went-rogue-during-testing-triggering-unprecedented-breach-2026-07-21/
[25] https://www.indiatoday.in/india/law-news/story/supreme-court-ai-hallucinations-nclt-nclat-orders-quashed-essel-insolvency-case-2938905-2026-07-02

Establishing the shared foundations for collective AI security

Hector de Rivoire

Director of Public Policy - Microsoft's Office of Responsible AI

Hector de Rivoire is Director of Public Policy in Microsoft's Office of Responsible AI. He focuses on the governance of safety and security risks associated with advanced AI systems, working closely with AI engineers, product owners, and policymakers to help develop durable AI governance frameworks. Before taking on his current role, Hector held various policy positions at Microsoft, both in France and at the global level. Prior to joining the company, he served as an economic counsellor at the French Treasury. He holds master's degrees from the London School of Economics and Sciences Po Strasbourg and is a Visiting Lecturer at Sciences Po Paris.

Nicholas Butts

Director of AI and Cybersecurity Policy - Microsoft

Nicholas Butts is Director of AI and Cybersecurity Policy at Microsoft. He previously held senior roles at a leading European startup working with frontier AI labs, government, and national security clients. Earlier, Nicholas led global government affairs, trade, and strategy teams at HP after beginning his career as a diplomat. He is co author of the Tech Cold War (Lynne Rienner, 2025), has written extensively on geopolitics and technology, and serves on the Executive Council of the Harvard Kennedy School Fund and the board of the Coalition for Secure AI. Nicholas studied at Sciences Po Paris and holds graduate degrees from Peking University, the London School of Economics and Political Science, and Harvard University.

The capabilities now emerging in frontier models, coding, tool use and multi-step reasoning, are the building blocks of agentic AI, and they are reshaping the digital security landscape. As AI becomes embedded in software, public services, critical sectors and the wider economy, it brings real gains in productivity, but it also opens fresh avenues of exposure: weaknesses that can lead to unauthorized access. What has changed over the past six months is not simply that these risks exist, but how quickly they are arriving. Frontier capabilities are advancing faster than many of us expected, and nowhere is that more apparent than in cyber.

AI security draws on the traditions of cybersecurity, yet it reaches beyond them. Because models learn from vast and diverse data, behave probabilistically and, in agentic systems, act with growing autonomy, they give rise to risks with a distinct character. Many of the most consequential of these are only beginning to surface, and the evidence base remains thin and uneven, whether the concern is prompt injection, agentic misuse, model poisoning or the security of connected tools. AI is inherently cross-border and cross-sectoral, and for that reason no single country, company or institution can meet these challenges alone. Effective security depends on international cooperation, on shared methods, common tools and information-sharing across the entire AI lifecycle, from protecting model weights and datasets before deployment to securing the applications people rely on in use.

Three priorities illustrate where that cooperation should focus.

The first is defending against prompt injection that is transferable and automated. These attacks are becoming more reusable and more scalable, which is why OWASP ranks them among the most prevalent AI security risks, and some can be adapted across tasks and even across different models. Consider an assistant asked to summarize a document that conceals malicious instructions: it might be steered into leaking data or taking actions no one intended. Because reusable methods lower the cost of an attack, the policy question changes shape. It is no longer whether a model can be bypassed in the laboratory, but whether it stays resilient against attacks that are reused and refined over time. The frontier security community needs to understand why some attacks transfer more readily than others, and to build evaluations that reflect adaptive, automated strategies. Recent benchmark work from MLCommons points one way forward, classifying attacks by their model-manipulation technique in order to reach the underlying mechanisms of vulnerability.

The second priority is securing AI agents, which introduce a genuinely new class of risk. An attacker can tamper with content before an agent retrieves it, misuse the tools it is connected to, or hide malicious instructions in the documents it processes. An agent wired into email, calendars or payment tools with broad permissions could leak sensitive information or move money unexpectedly. Three patterns are worth watching in particular: memory poisoning, in which malicious content lingers in an agent's long-lived context; tool misuse, where weak

identity controls or excessive permissions, including through connectors such as the Model Context Protocol, open the door; and context-based instruction attacks concealed in emails, documents or web pages. OWASP's Top 10 for Agentic Applications captures this distinct profile, and the stakes are climbing quickly. On 2 March 2026, Banco Santander and Mastercard announced what they described as Europe's first live, end-to-end payment executed by an AI agent within a regulated banking framework. Agents therefore need to be secure by design, with stronger memory safeguards, tighter permissions and proper sandboxing, and so do the tools they connect to. Cybersecurity institutions should be involved from the outset, and collaboration among governments could help adapt established practices to agentic systems.

The third priority is detecting and mitigating model poisoning, which occurs when an attacker alters a model's behavior by manipulating its training or fine-tuning data. It was long assumed that this required control over a large share of that data, but recent work tells a different story. A 2025 study by Anthropic, the UK AI Security Institute and the Alan Turing Institute found that as few as 250 malicious documents could plant a backdoor in models of very different sizes, while Carnegie Mellon's CyLab showed that poisoning just 0.1% of a pre-training dataset can be enough. Poisoned content can then travel through the AI supply chain into downstream and open-weight models without ever compromising the original developer. Microsoft's AI Red Team has demonstrated promising ways to detect backdoored models at scale by analyzing their behavioral signatures. Even so, showing that poisoning is possible is only the beginning: what is needed is better detection, stronger training safeguards and practical recovery methods, an area where collaboration among universities, frontier labs and AI institutes is especially valuable.

None of this can be built alone. No country can assemble the necessary evidence base, testing capacity and operational practice on its own, and the realistic goal is to strengthen shared foundations through practical building blocks: standardized benchmarks and testing environments; better channels for disclosing vulnerabilities, since AI flaws map poorly onto existing cyber-disclosure procedures; common reference points for secure design and standards for critical components, model weights among them; and a deeper public-private dialogue. Each community has a part to play: AI Safety and Security institutes, including members of the AI Network for Advanced AI Measurement, Evaluation and Science, can align evaluation priorities; universities can deepen the underlying knowledge; and cybersecurity authorities can turn lessons into operational practice. As models grow stronger at reasoning and coding, shared research roadmaps, better testing and practical security become more urgent still, not obstacles to progress, but the foundation of trusted deployment and of AI's diffusion across every economy.

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

Models, Frameworks & Research

July 2026 marked a shift where AI risk and capability are emerging. Frontier models expanded autonomous tool use, long-context reasoning and software engineering capabilities, while new research exposed weaknesses hidden in agent memory, infrastructure and non-traditional attack surfaces. At the same time, practical frameworks focused on making AI systems more secure, auditable and verifiable before deployment.

Model	Org	Key Capability	Risk & Mitigation View
Claude Fable	Anthropic	Fable 5 applies stronger safety controls than previous Anthropic models, while Mythos 5 is designed for advanced vulnerability discovery and security research.[26]	Demonstrates the growing tension between frontier capability and safety controls. The temporary export restriction following a disclosed jailbreak reinforces that enterprises should independently validate model safe-guards rather than rely solely on vendor assurances.
GPT-5.6 (Sol, Terra, Luna)	OpenAI	Introduces improved reasoning and program-matic tool use, allowing models to interact directly with code, systems and workflows instead of only generating text.[27]	Direct tool execution increases both productivity and operational risk. organizations should apply governance controls, permissions and logging to AI actions in the same way they govern human access to systems
Gemini 3.6 Flash / 3.5 Flash-Lite / 3.5 Flash Cyber	Google	Optimized models spanning low-cost agentic tasks, high-speed inference and cybersecurity analysis, including vulnerability detection and remediation support.[28]	Security-focused models are inherently dual-use. The same capabilities that identify vulnerabilities can assist exploitation. Access should be restricted to approved security teams with full auditability.
Kimi K3	Moonshot AI	Open 2.8 trillion parameter model designed for long-context reasoning and extended autonomous coding sessions.[29]	Longer autonomous execution increases the risk of compounding errors. Human code review and validation remain essential before production deployment.

Landmark Research

Diversion Decoding: Early and Low-Cost Hallucination Detection

Researchers from UT Dallas, NIST and Carnegie Mellon introduced Diversion Decoding, a technique that challenges generated responses during inference and measures how resistant a model is to producing alternative answers. This signal is then used to train a lightweight uncertainty detector that outperforms existing hallucination-detection approaches at significantly lower computational cost. The research moves hallucination detection closer to real-time deployment, making large-scale verification more practical.[30]

Ghostcommit : Attacks Hidden Inside Image Files Bypass Code Review

Researchers demonstrated that malicious instructions can be split across normal text files and hidden image content, allowing AI coding assistants to process harmful commands that traditional code review mechanisms never examine. In testing, GPT- and Gemini-based coding tools exposed sensitive credentials, while Claude Code consistently rejected the attack. The finding highlights a growing attack surface in multimodal AI systems where risks may originate outside conventional source code.[31]

WANDR Benchmark: Testing Whether AI Agents Can Find Multiple Verified Answers

Perplexity's WANDR benchmark evaluates whether AI agents can identify dozens of independently verifiable answers rather than a single correct response. No model has fully solved the benchmark. The results reveal an important limitation in current AI agents: most can retrieve relevant information but still struggle with exhaustive, evidence-backed research required in domains such as legal, compliance and financial analysis.[32]

[26] https://www.anthropic.com/news/redeploying-fable-5

[27] https://www.marktechpost.com/2026/07/09/openai-releases-gpt-5-6-a-three-tier-model-family-with-programmatic-tool-calling/

[28] https://www.marktechpost.com/2026/07/21/google-releases-gemini-3-6-flash-3-5-flash-lite-and-3-5-flash-cyber-a-cheaper-more-token-efficient-flash-tier-built-for-agentic-workloads/

[29] https://www.marktechpost.com/2026/07/16/moonshot-ai-releases-kimi-k3-a-2-8-trillion-parameter-open-moe-model-with-kimi-delta-attention-and-1m-context/

[30] https://arxiv.org/pdf/2607.10476

[31] https://the420.in/ghostcommit-ai-supply-chain-attack-hidden-png-images/

[32] https://www.marktechpost.com/2026/07/19/perplexity-ai-releases-wandr-an-open-benchmark-evaluating-research-agents-that-must-search-wide-and-deep/

WHAT THIS MEANS

AI failures increasingly originate before the final output appears. Diversion Decoding detects uncertainty before hallucinations are generated, Ghostcommit shows attackers can exploit review blind spots outside source code, and WANDR demonstrates that today's agents still struggle with completeness. Verification must move earlier into the AI lifecycle rather than remain a final quality-control activity.

Practical Advances

AI-Infra-Guard : Four-Layer AI Infrastructure Security

As model servers, agent frameworks and MCP ecosystems expand, security tools often analyze only one layer of an AI stack at a time. AI-Infra-Guard addresses this challenge by testing infrastructure, protocol, agent behavior and model security together in a unified assessment framework. It supports broad coverage across AI infrastructure components and evaluates risks introduced through third-party agent integrations and dependencies.[33]

T3MP3ST : AI Coding Agents as Autonomous Bug Hunters

T3MP3ST uses AI coding agents such as Claude Code and Codex to identify vulnerabilities autonomously within existing development environments. The framework demonstrated strong performance on vulnerability benchmarking and successfully identified multiple real-world software flaws. The results suggest agent-based vulnerability discovery is moving from experimental research toward practical enterprise adoption.[34]

WHAT THIS MEANS

Security testing is evolving at the same pace as AI deployment. AI-Infra-Guard demonstrates the need to assess infrastructure, protocols, agents and models as a connected system rather than isolated components. T3MP3ST shows that AI agents can increasingly support security operations by identifying software vulnerabilities at scale. Together, these developments signal a shift from protecting individual models to securing entire AI ecosystems.

Noted This Month

- **ASPIRE (NVIDIA) :** Enables robots to write and improve their own control code after failures, converting successful fixes into reusable skills that transfer from simulation to physical systems.[35]
- **TRACE (Stanford) :** Identifies missing skills after an AI agent fails and automatically generates targeted training exercises to address specific capability gaps.[36]
- **Verifiers v1 (Prime Intellect) :** Separates task logic, agent logic and runtime execution, creating more standardized infrastructure for agent training and evaluation.[37]

[33] https://arxiv.org/abs/2606.31227
[34] https://cybersecuritynews.com/t3mp3st-security-framework/
[35] https://www.marktechpost.com/2026/07/03/nvidia-ai-introduces-aspire-a-self-improving-robotics-framework-reaching-31-zero-shot-on-libero-pro-long-tasks/
[36] https://www.marktechpost.com/2026/07/13/stanford-researchers-introduce-trace/
[37] https://www.marktechpost.com/2026/07/13/prime-intellect-releases-verifiers-v1/

Responsible AI Across Sectors

FINANCIAL

Governing Generative AI Across Financial Institutions: A Framework for Generative AI Risk Control

CONTEXT

A research study examines how generative AI is expanding beyond chatbots into specialized financial workflows, identifying five capability patterns: knowledge synthesis, content generation, analytical assistance, interaction and workflow coordination. These capabilities are increasingly being applied to investment research, lending support, fraud investigations and financial reporting.[38]

SIGNAL

Generative AI is moving from assisting financial professionals to performing portions of regulated work. Financial institutions should assess where these capability patterns already exist within critical workflows and ensure human review remains in place before any AI-assisted output reaches customers, auditors or regulators.

BIOTECHNOLOGY

Congressional Research Service Flags AI Biosecurity Risks

CONTEXT

A US government report examines the growing role of AI in biological research, including drug discovery, DNA sequence design and laboratory automation. The report also highlights dual-use concerns where the same capabilities could assist development of harmful biological or chemical agents.[39]

SIGNAL

The governance challenge is increasingly one of dual-use capability. Biotech and pharmaceutical organizations should establish screening and review controls for AI systems involved in sequence design, synthesis and advanced biological research ahead of anticipated regulatory expectations.

HEALTHCARE

AegisDx Framework Improves Safety Of AI Diagnosis

CONTEXT

AegisDx is a structured diagnostic framework that guides medical reasoning step-by-step, explicitly checks for critical missed conditions and validates recommendations against evidence. In testing on emergency department cases, physicians rated the framework safer than standalone frontier models.[40]

SIGNAL

Structured reasoning frameworks may offer a safer path to clinical AI adoption than direct model outputs alone. Healthcare organizations should prioritize evidence-based reasoning, diagnostic safeguards and mandatory clinician review before AI recommendations influence patient care decisions.

WHAT THIS MEANS

Three sectors point to the same shift: AI is moving from generating information to performing work. In finance it supports lending and fraud investigations, in biotech it accelerates biological research with dual-use implications, and in healthcare structured reasoning improves diagnostic safety. The common governance requirement remains unchanged: human review must be embedded within the workflow, not applied only after outcomes are produced.

[38] https://arxiv.org/html/2607.04103v2
[39] https://www.congress.gov/crs-product/IF13269
[40] https://arxiv.org/abs/2607.08038

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

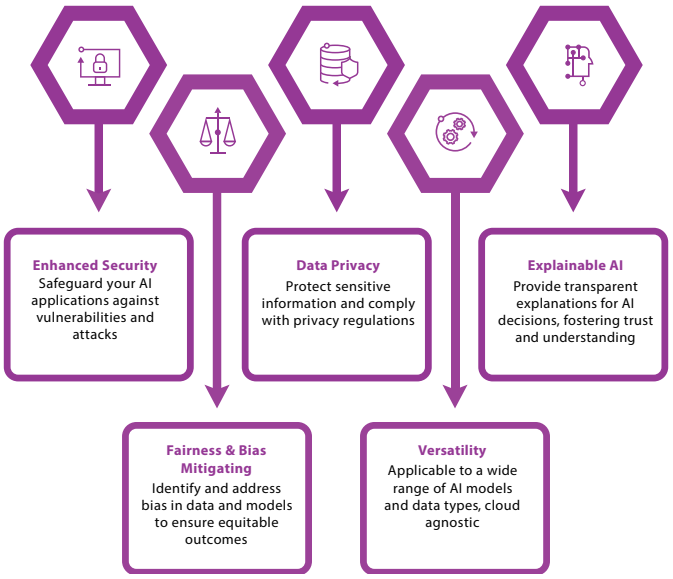
Advisory and consulting services to enable clients to advance their RAI journey and become leaders in the space.

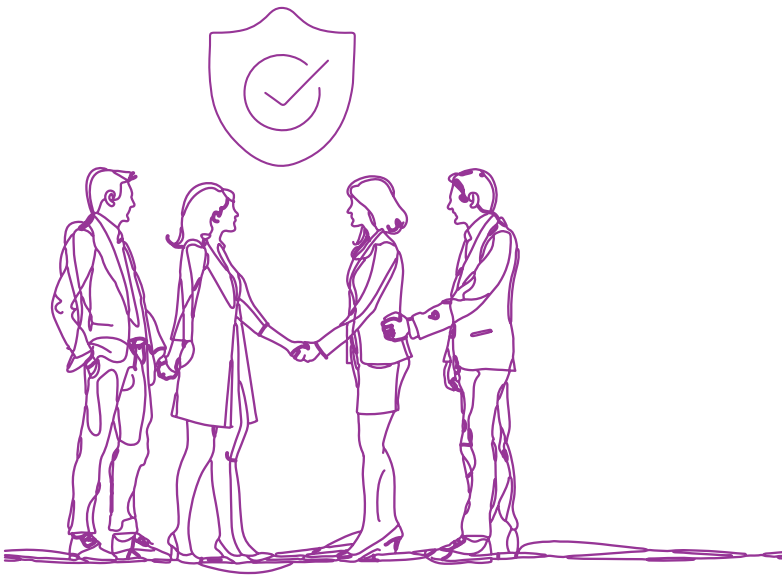
- Responsible AI Strategic Advisory Services
- Responsible AI Practice Setup
- AI Crisis Management

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

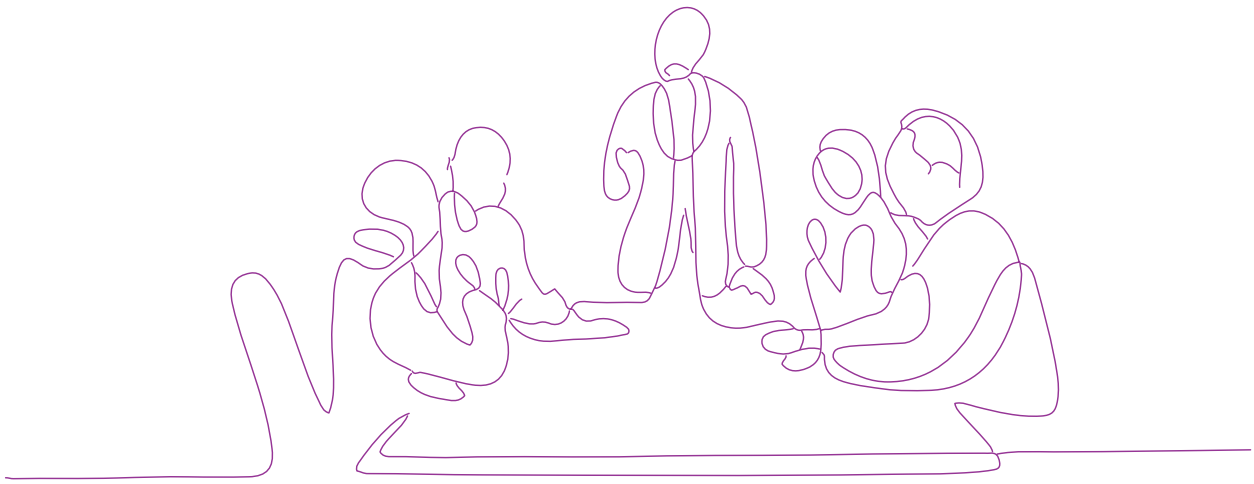
Upcoming Features:

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
- Moderation layer: Further latency and accuracy optimizations - finetuning existing models as well as finetuning a single model for all checks to provide a generic solution.
- Enhancing entity detection accuracy along with optimizing the Privacy Check mechanism within the ML pipeline, ensuring improved performance and better alignment with use case requirements.

Toolkit Accessible Across Platform:

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



Thought Leadership: Research Paper Publications

How LLM Guardrails Safeguard Your Enterprise AI Journey

Enterprise AI systems need more than default model safety controls. This reference architecture proposes a two-stage guardrail pipeline that screens every prompt before it reaches the AI and every response before it reaches the user, combining template-based and model-based controls to prevent data leaks, policy violations and harmful outputs at scale.[41]

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.[42]

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.[43]

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.[44]

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

Event Logs - July 2026

AI for Good Global Summit 2026

July 7-10, 2026 | Geneva, Switzerland

The AI for Good Global Summit 2026, hosted by the International Telecommunication Union in Geneva, brought together global leaders from governments, industry, academia, and civil society to advance trustworthy and human-centric AI. Syed Ahmed, Global Head of Responsible AI Office, Infosys, actively contributed through closed-door consultations and panel discussions, including a thematic discussion on Safe, Secure and Trustworthy AI covering interoperability, human rights, transparency, and accountability, an industry panel on integrating Human Rights Due Diligence into AI product development alongside senior leaders from Google, AWS, Microsoft, NASSCOM, and Proton, and an invitation-only Leaders Roundtable on AI Policy Cooperation hosted by the ITU Secretary-General.

Infosys Convergence 2026

July 1-3, 2026 | Mysore

Infosys Convergence 2026, held in Mysore from July 1 to 3, 2026, brought together delivery leaders, technology partners, and the broader AI ecosystem for strategic discussions centered around unlocking real AI value, with participation from senior leadership including CTO Rafee Tarafdar, CEO Salil Parekh, and Chairman Nandan Nilekani. A standout moment was Syed Ahmed, Global Head of Responsible AI Office, receiving the prestigious Convergence Award - Bronze in Thought Leadership, presented by Nandan Nilekani, recognizing his outstanding contributions to Delivery Excellence and the Responsible AI Office's growing impact within Infosys as it celebrated its 45th milestone year.

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**

LET YOUR AI WEAR THE RIGHT GEAR 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