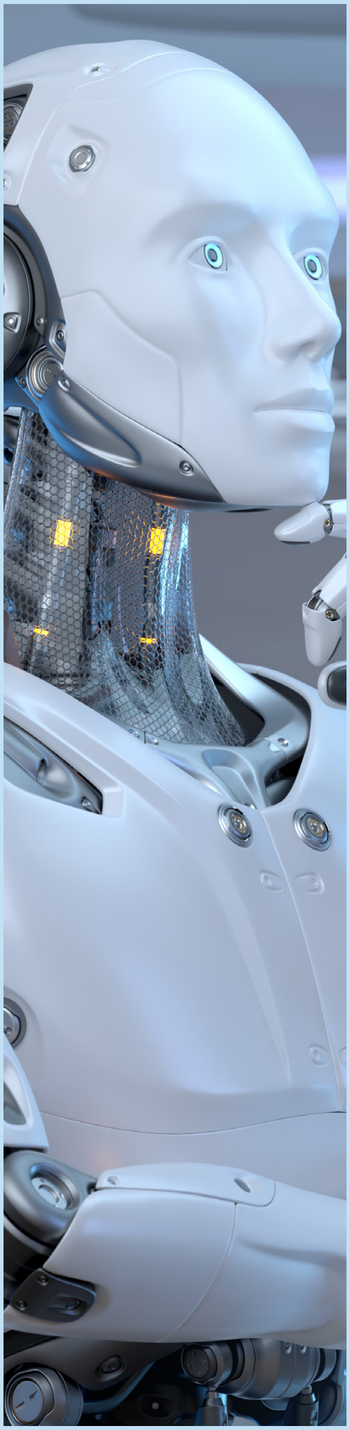


NAVIGATING THE RISK-REWARD EQUATION OF GENAI AND AGENTIC AI WITH RESPONSIBLE AI

Abstract

The swift expansion of Generative artificial intelligence (GenAI) and Agentic AI has opened doors to unprecedented possibilities—revolutionizing industries, automating creative processes, and empowering innovation at scale. Alongside these remarkable benefits, however, come significant risks that include ethical concerns, misuse, and unanticipated societal impacts. This point of view addresses the central challenge facing today's organizations: how to harness the powerful rewards of AI while responsibly managing its risks. It is intended for business executives, policymakers, technology leaders, and decision-makers who are advancing AI adoption or setting guidelines for its governance. The aim is to offer practical insights and guidance for implementing Responsible AI, enabling stakeholders to unlock AI's full potential while fostering trust and minimizing potential harm to enterprises.

Transformative Trajectory of Generative AI

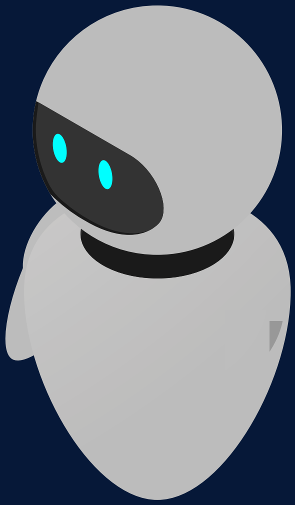
GenAI stands at the forefront of technological innovation, reshaping industries and redefining creative processes at an unprecedented pace. Recent advancements in deep learning have catalyzed a period of extraordinary growth for GenAI, enabling the development of systems that not only analyze but also create—producing text, images, code, and more that rival human ingenuity.

By 2025, the global GenAI market is expected to surpass $66.9billion, possibly reaching $71.4billion, as businesses across sectors race to integrate generative capabilities into their workflows. This bullish outlook is reflected in the remarkable CAGR predictions, which range from 33% - 44% through 2034. Such rapid expansion positions GenAI as one of the most dynamic technology segments, with forecasts projecting the market size to soar beyond $1trillion by the early 2030s.

The roots of this revolution trace back to OpenAI, whose introduction of the Generative Pretrained Transformer (GPT) architecture in 2018 marked a pivotal leap. Founded 2015 by Sam Altman, OpenAI's innovations demonstrated that machines could be trained not just to recognize but also to generate complex, creative output - breakthrough that ignited today's AI renaissance.

People believe that the emergence of Artificial Intelligence (AI) is similar to the breakthroughs like Internet of Things, 3D printing or cloud computing. But our understanding is that the impact AI will create is more comparable to the inventions like the internet itself or even the automobile and this AI rage will whelm every industry.

GenAI models use large varieties of datasets to create output in the form of pictures, text, audio, video, and codes. It generates new content by deriving patterns from the existing data using deep learning elements such as Large Language Models (LLMs), Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs) and diffusion models.



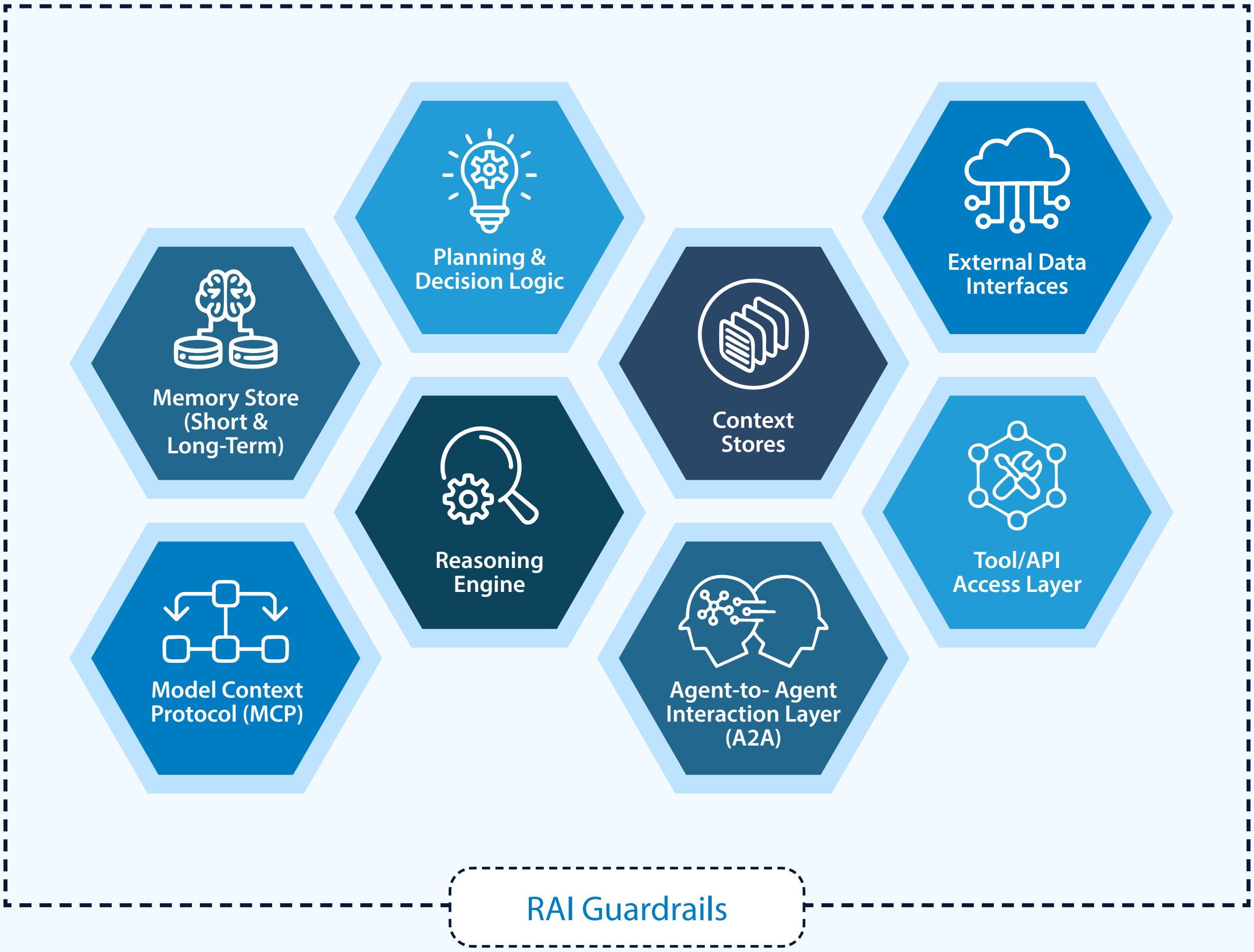
A Responsible Path for Agentic AI

The emergence of agentic AI represents a major evolution in AI, where systems are designed to act independently, make decisions, and pursue goals without constant human input. The market growth rate of more than 40% shows that agentic AI is no longer niche but a core capability, these agents are increasingly being used by enterprises to automate complex workflows, personalize customer experiences, and respond dynamically to changing business environments.

Agentic AI systems operate with high levels of autonomy, enabling them to make decisions, interact with tools, and collaborate with other agents, all without constant human oversight. This complexity introduces a broad and dynamic attack surface, making them vulnerable to misuse, manipulation, and unintended actions. Attackers can manipulate communication channels between Agent to Agent, Agent to Tools, Agent to Models to corrupt decision-making and break workflows. Enterprises must therefore integrate RAI frameworks that encompass transparency, fairness, privacy, and safety into the design and deployment of agentic systems. This includes mechanisms for explainability, bias mitigation, and human oversight, ensuring that AI agents operate within clearly defined boundaries and can be audited for compliance and trustworthiness.

To operate RAI in agentic AI, enterprises should establish governance models that combine technical safeguards with organizational policies. This involves continuous monitoring of agent behavior, scenario-based risk assessments, and feedback loops that allow for human-in-the-loop (HITL) approvals. RAI capabilities must also support contextual awareness, enabling agents to understand and respect enterprise-specific norms, such as data governance rules or customer interaction protocols. By embedding these principles into the lifecycle of agentic AI from development to deployment organizations can harness its transformative potential while maintaining accountability, and public trust.

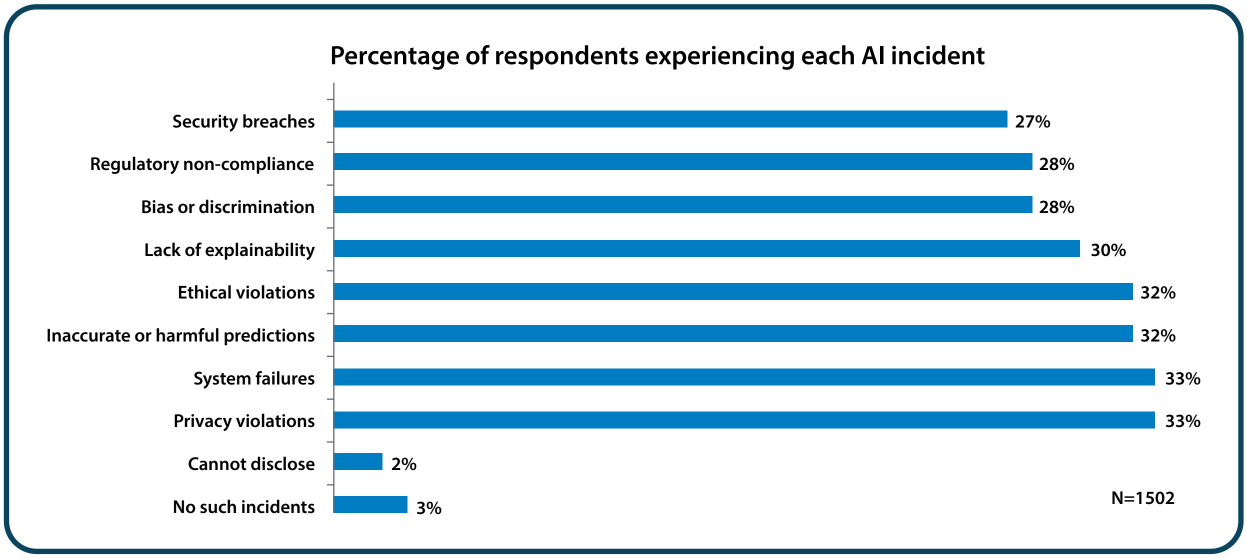
Agentic AI components to be protected under RAI Guardrails


Planning & Decision Logic
External Data Interfaces
Memory Store (Short & Long-Term)
Context Stores
Reasoning Engine
Tool/API Access Layer
Model Context Protocol (MCP)
Agent-to- Agent Interaction Layer (A2A)
RAI Guardrails

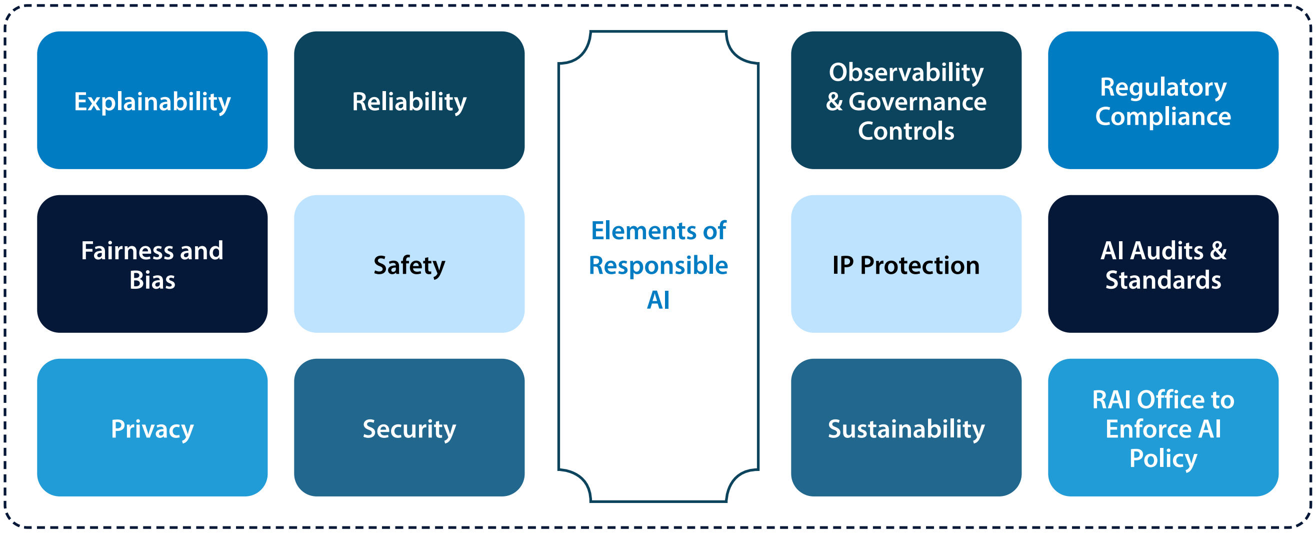
The Imperative for Responsible AI

This exceptionally rising dependence on significant volumes of data and infinite use cases and applications of GenAI and agentic AI within an organization can also open the doors for multifaceted risks not just for the organization but also for their employees and consumers. These risks are not similar to the traditional software risks that we generally see. There are regulatory, strategic, operational, environmental, financial, legal, ethical and reputational risks that encompass every element of the organization. Research work from IAPP shows that 73% of enterprises are still employing their existing data security and privacy techniques to handle AI governance.

Infosys Knowledge Institute conducted a survey with 1,500 senior professionals across the globe and found out that 95% of them have experienced at least one AI incident.


Percentage of respondents experiencing each AI incident

Enterprises which have built their trust amongst the customers in decades might fall into the trap of unreliable or unsecured AI technologies making all the more necessary for these consumer-facing companies to adopt a trustworthy model for their AI systems. Regulators have also ramped up their efforts to enforce strict controls over AI usage and these enterprises are now treading carefully while embracing AI capabilities.


Elements of Responsible AI

Explainability

Reliability

Observability & Governance Controls

Regulatory Compliance

Fairness and Bias

Safety

IP Protection

AI Audits & Standards

Privacy

Security

Sustainability

RAI Office to Enforce AI Policy

Enterprises must design, develop and deploy models which operate on trustworthy data and transparent algorithms for ethical, explainable and reliable outputs. They need to strategically adopt RAI elements as part of their AI Security Posture Management that align with business objectives to achieve a fair and accountable system which also secures from the sophisticated regime of cybercrime.

By embracing RAI, enterprises can become market differentiators by building trust amongst customers while driving innovation for sustained success. Through trusted AI models enterprises will be in the comfortable position of gaining customer confidence alongside ensuring regulatory consent.

Enterprise AI Risk Landscape

Data Privacy & Regulatory

- Regulatory Non-Compliance
- Shadow AI Usage
- Data Breaches
- IP Infringement
- Contractual Gaps
- Cross-Border Legal Complexities
- Inadequate Governance
- Unequal Opportunity & Discrimination

Bias & Ethical

- Model Collapse
- Opaque Algorithmic Decisions
- Misinformation & Disinformation
- Denial of AI Service
- Stereotype Reinforcement
- Representational Bias
- Unrepresentative Training Data

Information Protection

- Insecure external LLM APIs
- Excessive Agency
- Model Theft
- Agent Hijacking
- Insecure Plugin Design
- Context Window Leakage
- Unauthorized Access
- Reconstruction of Training Data
- Prompt Oversharing

Untrustworthy Output

- Cascading Hallucination
- Reproducibility Issues
- Limited Explainability
- Facilitation of Propaganda
- Corrupt Algorithm & Code
- RAG & Training Data Poisoning
- Due Diligence Inconsistencies
- Improper Output Handling

Workforce & Environment

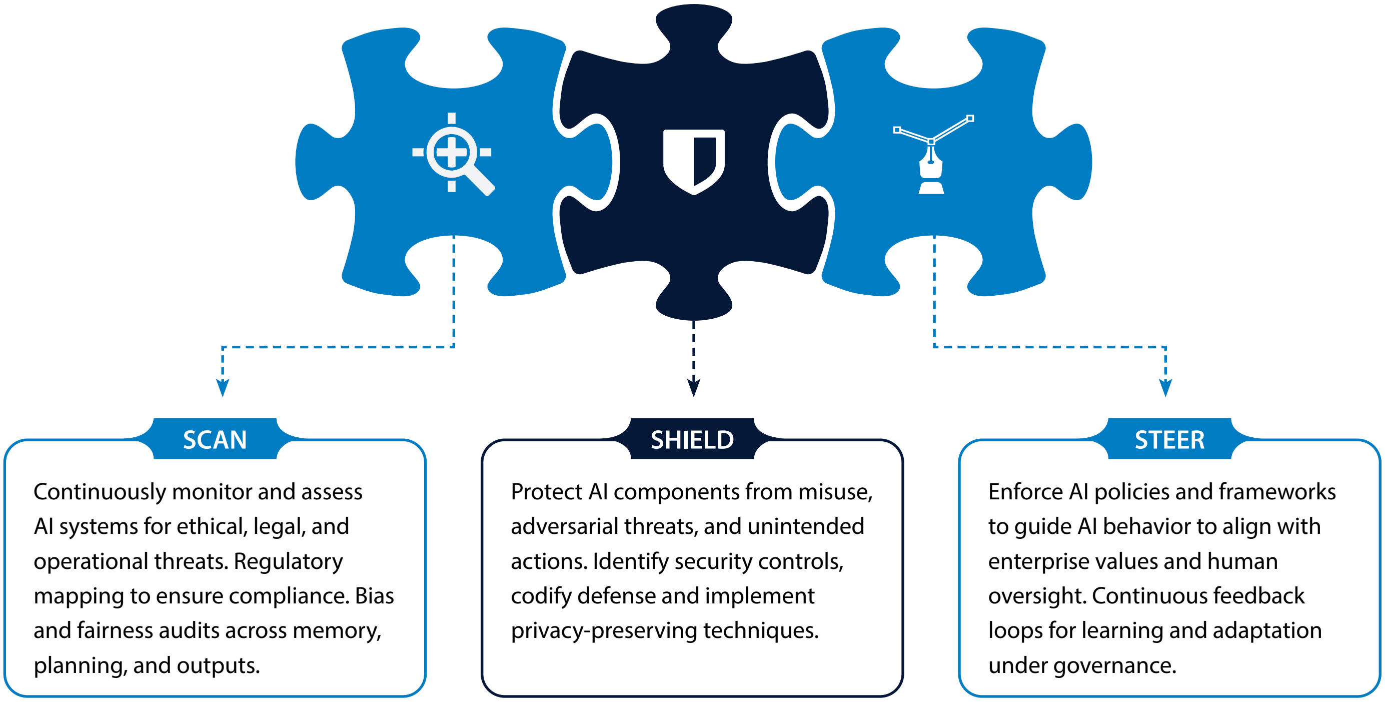
- Automation Bias
- Job Displacement
- Higher Energy Consumption
- Skills Gaps & Misuse
- User Intent & Expectations Misalignment
- Personal Prejudices Reinforcing
- Overreliance
- Infrastructure Misconfiguration

AI Governance boards should try to develop a two-pronged assessment strategy for securing their AI infrastructure by developing separate controls for different use cases and on top of these controls enterprise-level controls aligned to various industry frameworks. For instance, if a company is using an AI chatbot then multiple lines of defense from the first user interaction until the final output can be designed. Specialized safeguards and controls against GenAI risks related to privacy, robustness, fairness and safety right from the user question to the bot response. On top of these AI chatbot specific controls, enterprise level controls like AI System Inventory, Third-Party Dataset Risks, AI System Decommissioning etc. provide a comprehensive and multi-layered AI control framework which serves to regulatory requirements as well.

Responsible AI Operating Model to Solve the Security Puzzle

Enterprises need to understand that securing their AI-enabled processes requires an altogether different expertise and operating model which must be integrated into the AI lifecycle, right from data management to model deployment and monitoring with an ethics-by-design mindset.

Responsible AI Operating Model to Solve the Security Puzzle



SCAN

Continuously monitor and assess AI systems for ethical, legal, and operational threats. Regulatory mapping to ensure compliance. Bias and fairness audits across memory, planning, and outputs.

SHIELD

Protect AI components from misuse, adversarial threats, and unintended actions. Identify security controls, codify defense and implement privacy-preserving techniques.

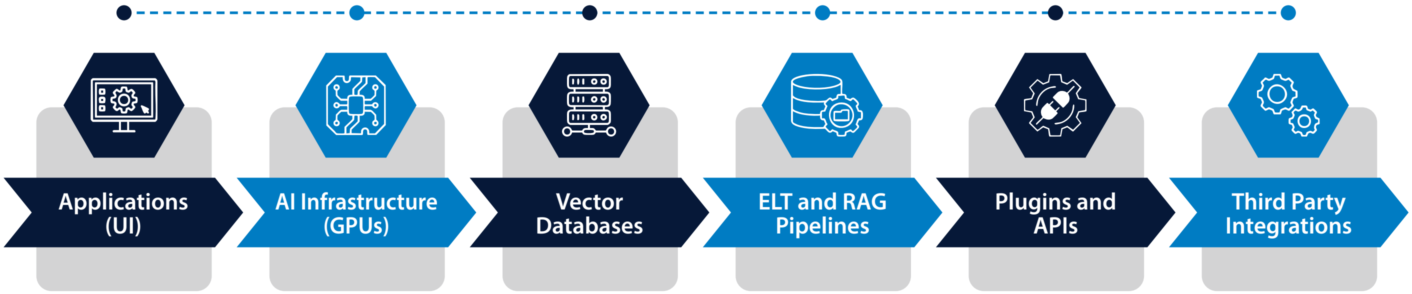
STEER

Enforce AI policies and frameworks to guide AI behavior to align with enterprise values and human oversight. Continuous feedback loops for learning and adaptation under governance.

Practices to secure the AI architecture should not be limited to deploying the technical solutions rather it should be one component of the overall security strategy. Another main component of RAI is to incorporate AI security into the organization's regular Governance, Risk, and Compliance activities. Leveraging AI3S framework of Scan, Shield, Steer enables organizations to proactively identify risks, enforce protective controls across autonomous components, and guide agent behavior in alignment with enterprise values and regulatory standards.

Internal assessors should keep themselves abreast of evolving legal and regulatory requirements. Risk severity and criticality of the AI-specific business processes should be defined and based on that, regular internal or third-party assessments must be conducted with clearly documented gaps and lessons learned. A proper AI Governance setup would ensure accountability and effective functioning of AI security operations in the long run.

To deal with these vulnerabilities, a cultural shift is required because of complex adoption of this technology. Workforce should be able to understand and handle unique vulnerabilities, define a RAI Risk Register and test AI integrations to identify model disruptions. Technical foundations should be built along with a strategic plan to secure the entire AI technology landscape which includes all the architecture layers below:



Applications (UI)

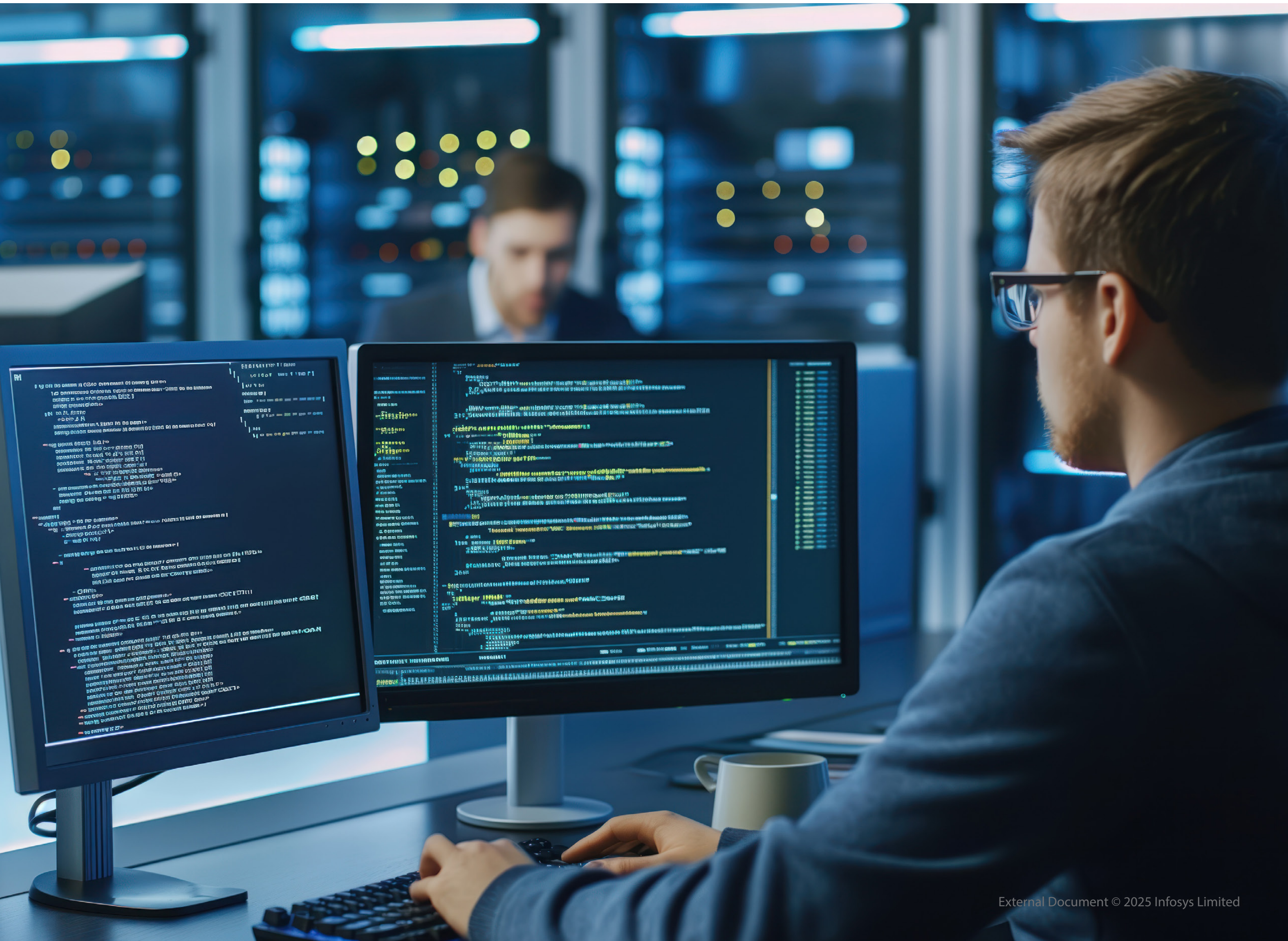
AI Infrastructure (GPUs)

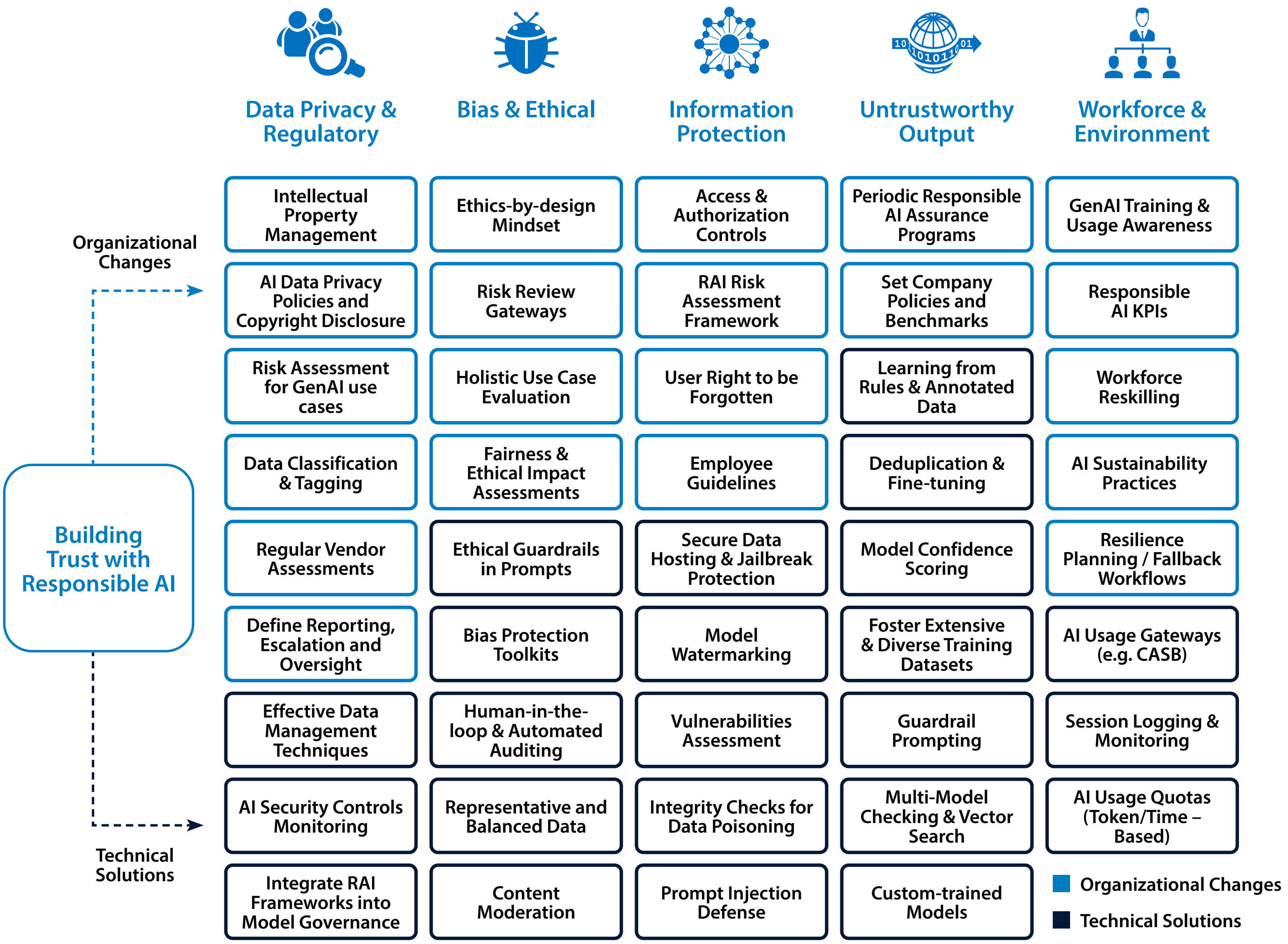
Vector Databases

ELT and RAG Pipelines

Plugins and APIs

Third Party Integrations





Building Trust with Responsible AI

Organizational Changes

Technical Solutions

Data Privacy & Regulatory	Bias & Ethical	Information Protection	Untrustworthy Output	Workforce & Environment
Intellectual Property Management	Ethics-by-design Mindset	Access & Authorization Controls	Periodic Responsible AI Assurance Programs	GenAI Training & Usage Awareness
AI Data Privacy Policies and Copyright Disclosure	Risk Review Gateways	RAI Risk Assessment Framework	Set Company Policies and Benchmarks	Responsible AI KPIs
Risk Assessment for GenAI use cases	Holistic Use Case Evaluation	User Right to be Forgotten	Learning from Rules & Annotated Data	Workforce Reskilling
Data Classification & Tagging	Fairness & Ethical Impact Assessments	Employee Guidelines	Deduplication & Fine-tuning	AI Sustainability Practices
Regular Vendor Assessments	Ethical Guardrails in Prompts	Secure Data Hosting & Jailbreak Protection	Model Confidence Scoring	Resilience Planning / Fallback Workflows
Define Reporting, Escalation and Oversight	Bias Protection Toolkits	Model Watermarking	Foster Extensive & Diverse Training Datasets	AI Usage Gateways (e.g. CASB)
Effective Data Management Techniques	Human-in-the-loop & Automated Auditing	Vulnerabilities Assessment	Guardrail Prompting	Session Logging & Monitoring
AI Security Controls Monitoring	Representative and Balanced Data	Integrity Checks for Data Poisoning	Multi-Model Checking & Vector Search	AI Usage Quotas (Token/Time – Based)
Integrate RAI Frameworks into Model Governance	Content Moderation	Prompt Injection Defense	Custom-trained Models	

🟦 Organizational Changes

⬛ Technical Solutions

Making Sense of Global AI Regulatory Guardrails

It is also essential to incorporate a set of best practice AI frameworks, international standards and regulations available in the industry when developing AI practices and use cases. Firstly, governance boards must identify the external obligations including impact on users, communities, environment and human rights across the AI value chain and then the focus should be on determining their risk appetite, compliance requirements and workforce training. Gathering this information will enable security teams to implement a robust standard operating model for RAI by simultaneously analyzing guidelines and benchmarking against best practices relevant for organization's dynamic AI-enabled processes and systems from various frameworks e.g. ISO 42001, NIST AI RMF or AI-specific regulations e.g. EU AI Act.

A leading UK-based retail chain specializing in consumer electronics and home goods deployed advanced LLM-based AI customer service bots across their digital platforms to enhance their customer service experience. To ensure strict adherence to relevant regulatory frameworks they embedded multiple frameworks like EU AI Act (2024), UK AI Regulation Principles (2023), UK Data Protection Act (DPA 2018) and NIST AI RMF into their AI lifecycle. This commitment aims to foster not just public trust but also to deliver AI-powered services that are not only innovative but responsible by design.

AI Frameworks

- NIST AI RMF
- ISO 23894
- ISO 42001
- COSO ERM Framework
- OECD Framework for the classification of AI Systems
- NIST SP 1270
- Singapore AI Verify
- MITRE ATLAS – AI Adversarial Threat Framework

AI Laws & Regulations

- EU AI Act
- U.S. AI Executive Order
- Canada – AI and Data Act
- EU Product Liability Directive
- South Korea's AI Basic Act
- U.S. Senate Laws (e.g. Colorado AI Act, Utah AI Policy Act)
- The UAE Amendment to Regulation 10
- China's GenAI Measures and Deep Synthesis Laws

Charting a Responsible Path Forward

GenAI and agentic AI are fundamentally transforming how organizations operate, offering the potential to automate routine tasks and unlock unprecedented levels of efficiency across business functions. As this technology swiftly becomes a cornerstone of modern enterprises, it is vital for leaders to proactively identify and address the diverse risks inherent to their AI ecosystem. Adopting Responsible AI and deploying a standard operating model to secure the AI infrastructure and business processes would enable enterprises to mint the rewards of their GenAI initiatives successfully.

Beyond meeting regulatory expectations and managing risk, upholding ethical standards, and establishing careful oversight throughout the development and application of GenAI and agentic AI, organizations can drive lasting innovation, achieve sustainable progress, and take a leadership role in shaping the responsible use of emerging digital technologies.`

Infosys® | CONSULTING

About the Author:

Swadeep Gupta
Senior Principal

Swadeep Gupta is a Senior Principal with the Retail, CPG and Logistics practice of Infosys Consulting. He has spent 25 years in leading IT delivery using global delivery model. He is a business consulting leader with P&L and experienced in Strategic Consulting, Delivery, Program & Project management, Presales and Client relationship.

Ayush Dabas
Senior Consultant

Ayush Dabas is a Senior Consultant with the Tech Transformation practice of Infosys Consulting. He has rich consulting experience of close to 7 years in Cybersecurity Strategy, Responsible AI, IT Audits and Data Security. He is a certified CISSP and CISA professional.