Deloitte. | cisco

THE SOVEREIGN SHIELD SECURING MODERN INDIA'S AI ECOSYSTEM

July 2026

Table of contents

Foreword

Today, we stand at a defining moment. The vision of Viksit Bharat 2047, a confident, resilient and fully developed India, is no longer just a dream. It is taking shape every day as digital progress unfolds across the country. At the core of this change is Artificial Intelligence (AI). AI is set to strengthen our healthcare systems, transform agriculture, modernise finance and bring quality education to every corner of India.

However, as we build this future, driven by the ambitious goals of the India AI Mission, such as expanding access to compute and nurturing local innovation, we must recognise one simple truth: there is no digital sovereignty without digital security.

For years, security has been treated as a challenge, something that slows things down. In an age shaped by generative AI and autonomous systems, that conception must evolve. Security is the safety rail that lets us innovate at full speed. The India AI Mission's focus on safe and trusted AI reflects this understanding. It recognises that AI can uplift our economy only when citizens trust the systems built for them.

This is where responsible AI and ethics become critical. The standards emerging today, guided by NITI Aayog and reflected in our evolving regulations, are not merely compliance requirements. They form the value system of our technological journey. When we speak about fairness, privacy, or explainability, we refer to the democratic principles that define India. We are making sure that the AI influencing loans, hiring and public services treats every person with dignity, transparency and respect.

Adopting these practices is the true measure of our readiness for Viksit Bharat. A developed nation is defined by how responsibly it uses AI technology. As leaders, our responsibility goes beyond deploying AI models. We must build resilience and ethics into the foundation of our AI ecosystem, by design, not as an afterthought.

This joint point of view is a guide for that journey. It outlines how we can match the pace of innovation with the need for safety and trust. It calls on government, enterprises and start-ups to come together with a shared commitment: build a Cyber Surakshit Bharat that is secure, inclusive and deeply human.

Let us begin the next chapter of India's digital journey with purpose and confidence.

Gaurav Shukla
Partner and Leader – Cyber, Deloitte South Asia

Arun Shetty
Sr. Director & CTO, Cisco India & South Asia

Securing India's enterprise AI ecosystem and data centres through integrated governance and Zero Trust

India's rapid digital transformation has positioned the nation as a global leader in adopting advanced technologies, with AI emerging as a cornerstone of economic growth and national competitiveness. Today, over six million professionals[2] are engaged in technology and AI-related roles, fuelling an ecosystem that is expanding at an unprecedented pace.

As AI systems become deeply interwoven with critical data assets, large-scale compute infrastructure and essential services, the integrity and security of the AI ecosystem assume strategic importance. The evolution from GenAI to agentic AI signals a paradigm shift from systems that produce to systems that decide. As we enter this next frontier, building a resilient and trustworthy AI environment is a national priority rooted in data sovereignty, infrastructure resilience and the protection of economic and societal interests.

The promise of AI is undeniable, but does our current reality live up to that ambition?

The current state of AI adoption

India's technology sector is characterised by immense scale, with annual revenues projected to cross USD280 billion by 2030. The country hosts over 1,800 Global Capability Centres (GCCs), of which more than 500 are focused on AI development and deployment.[3] This thriving ecosystem is further powered by a remarkable talent pool, with AI skill penetration 3.09 times the global average recorded between 2015 and 2021.[4] Consequently, the adoption rate within key economic sectors is exceptionally high, with 87 percent of enterprises actively using AI solutions in areas such as Banking, Financial Services and Insurance (BFSI), industrial and automotive, consumer goods, retail and healthcare.

Despite this acceleration in adoption, an analysis of organisational maturity reveals a significant risk. On the NASSCOM AI Adoption Index, India scores 2.45 out of 4, classifying the nation at the "enthusiast" level. This position highlights a critical maturity paradox: governance and risk management have not kept pace with widespread adoption. Approximately 70 percent of companies allocate less than 10 percent of their IT budget to AI projects,[5] indicating a clear misalignment between the speed of deployment and the resources dedicated to securing these systems.

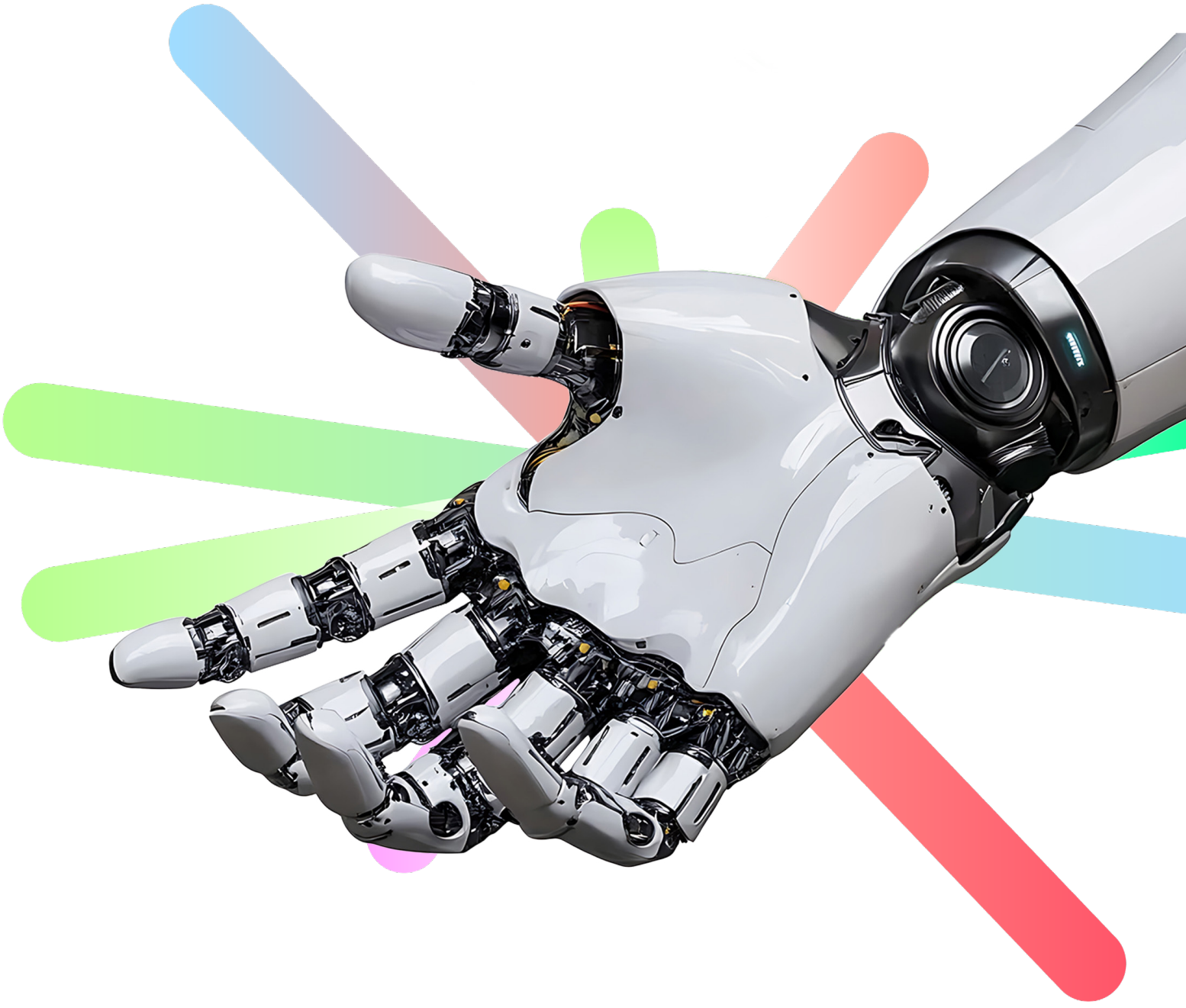
This enthusiastic but under-governed deployment posture directly translates into an expansive and highly vulnerable attack surface. When AI systems are rolled out without rigorous, dedicated security controls, they become susceptible to complex application-layer attacks and infrastructure compromises that target the reliability and repeatability of the models these applications rely on. Furthermore, the complexity of AI risk management, involving diverse stakeholders across product management, data and model engineering, infrastructure, legal, and compliance teams, can lead to fragmented ownership and unclear responsibilities. It can exacerbate the risks inherent in low-maturity environments. This deficiency in dedicated resources and coordinated governance structure compounds the challenge presented by the acute 18 percent skill gap identified in the cybersecurity domain across India.[6]

[2]Static PIB

[3]India AI Mission

[4]AI Skill Penetration

[5]NASSCOM AI Index



[6]Government initiatives to strengthen cyber ecosystem

Key challenges in securing AI

The promise of AI is immense, but so is the complexity of securing it. Unlike traditional digital systems, AI introduces new forms of dependency on data, compute and continuous learning pipelines often spanning multiple organisational and operational boundaries. While AI is everywhere, the frameworks to safeguard it remain fragmented and underdeveloped.

Fragmented ownership across the AI value chain 1

AI spans domains such as data engineering, model development, cloud infrastructure, cybersecurity, legal and product teams, yet no single function traditionally owns end-to-end accountability.

This fragmentation results in:

- Gaps in decision rights, where no team is explicitly accountable for model security, data lineage or ethical deployment
- Inconsistent control maturity, as each function applies its own interpretation of security, fairness and risk requirements
- Slow incident response, as breaches affecting AI pipelines require coordination across multiple siloed teams

In global benchmarks, this "multi-owner, no-owner" dynamic consistently emerges as the primary cause of AI security failures.

Limited lineage, transparency and observability 2

AI systems depend on evolving datasets, third-party models, parameterised code and distributed infrastructure. The resulting opacity creates challenges such as:

- Unknown data provenance, making it difficult to validate whether training data is authorised, high-quality, or free of embedded bias or malicious content
- Low visibility into model behaviour, especially in blackbox or deep learning systems, where decisions cannot easily be explained or monitored
- Insufficient monitoring of pipelines, where traditional logs capture system activity but not model drift, data skew or adversarial anomalous behaviour

Global incidents such as model poisoning and unintended bias amplifications show that a lack of transparency is one of the most systemic AI risks.

Security retrofits instead of secure-by-design 3

Most enterprises push AI projects rapidly from proof-of-concept to production, prioritising innovation speed over structured risk management. This results in the following situations:

- Models getting deployed without formal threat modelling, making them vulnerable to prompt injection, data extraction and inference manipulation
- Late addition of security, with controls bolted on after the architecture is locked, which increases cost and reduces effectiveness
- Inconsistent hardening of model endpoints, often leaving Application Programming Interface (APIs) exposed to misuse, over-permissioned access or unmonitored integrations with enterprise systems

This "race to production" mirrors early cloud adoption patterns, but AI systems carry more profound downstream consequences.

Immature regulatory alignment and compliance complexity 4

AI governance is evolving globally, with differing interpretations of what constitutes responsible, safe and compliant AI. Enterprises today face the following concerns:

- Ambiguity in regulatory obligations, particularly around data protection, model explainability, consent and cross-border data flows
- Growing expectations for auditability, where firms must demonstrate how models were trained, validated, tested and monitored
- Sector-specific requirements, especially in finance, healthcare, public sector and critical infrastructure, which demand precise risk-based controls for AI

This regulatory momentum requires organisations to embed governance deeply into AI design rather than retroactively justifying model behaviour.

Expansion of the technical attack surface 5

AI introduces unique security hotspots that differ substantially from traditional IT threats.

- Training pipelines that process sensitive or proprietary datasets can be targeted for poisoning or unauthorised access.
- Model files and registries can be tampered with, altered or substituted during promotion stages.
- Inference systems can be manipulated through adversarial prompts, crafted inputs or inference attacks that extract Intellectual Property (IP) or sensitive insights.
- Third-party and open-source dependencies dramatically increase supply-chain exposure in Machine Learning Operations (MLOps).

Without targeted defences, such as AI firewalls, red-teaming, model fingerprinting or provenance tools, organisations cannot reliably safeguard models and their data across their end-to-end lifecycle.

Centralised infrastructure creating single points of failure 6

AI workloads often consolidate compute in high-value environments such as Graphics Processing Unit (GPU) clusters, shared cloud platforms or sovereign data centres. This centralisation creates:

- Dense lateral-movement risk, where a breach in one container or node could compromise entire training clusters
- Dependency on specialised hardware, making firmware, drivers and accelerators critical components of the attack surface
- Operational fragility, as disruptions in power, cooling or network fabric have an outsized impact on AI availability

Lessons from global cloud-scale incidents indicate that AI infrastructure must be secured at the hardware, network and orchestration layers, in addition to the application level.

Talent and skills gap in AI security 7

AI security is a convergence of Machine Learning (ML), cybersecurity, distributed systems and regulatory expertise. Most organisations face multiple hurdles, including:

- Lack of dedicated AI security roles, relying instead on traditional cyber teams unfamiliar with adversarial ML or model governance
- Limited access to red-teamers with AI threat experience, affecting the ability to test and validate models rigorously
- Cross-functional training struggles, leading to uneven capability across the enterprise

This skills gap is one of the most significant global inhibitors to the secure scaling of AI.

These challenges are further amplified as AI infrastructure scales and centralises. The rapid proliferation of AI data centres, high-performance compute clusters and sovereign AI platforms introduces new dependencies on shared infrastructure, power, connectivity and specialised hardware. This convergence of fragmented governance and centralised AI infrastructure highlights the urgency of rethinking how AI systems are secured at scale.

AI data centre proliferation

The foundation of India's AI ambition rests upon a rapidly expanding physical infrastructure. India's data centre capacity is projected to double from 0.9 GW in 2023 to approximately 2 GW by 2026.[7] This growth, driven largely by digitisation and data localisation requirements, makes the physical and network integrity of these data centres a matter of critical national security.

AI data centres rely heavily on High-Performance Computing (HPC) technologies, primarily GPU clusters connected by high-speed, low-latency fabrics such as InfiniBand or Ethernet, utilising Remote Direct Memory Access (RDMA). While RDMA provides superior performance by allowing the Network Interface Card (NIC) to bypass the host Central Processing Unit (CPU) for memory access, its adoption introduces profound security challenges.

The design of RDMA originated in HPC environments that traditionally had a high degree of internal trust or relied on physical isolation for classified jobs. In contrast, a modern AI data centre is a shared, multi-tenant cloud environment where users and workloads are inherently untrusted. RDMA's design fundamentally alters security assumptions, potentially introducing unauditable data-access paths and undermining the reliability guarantees of traditional Remote Procedure Calls (RPCs). Securing this new architecture requires a radical shift from conventional network security models to hardware-enforced, Zero Trust principles applied directly to the high-speed fabric.

The significant expansion of AI infrastructure, coupled with the regulatory pressure of the India Digital Personal Data Protection (DPDP) Act, means that security investment must pivot strategically. The focus must move beyond application and traditional network security to securing the physical and logical fabric of these domestic AI data centres, mandating a Zero Trust approach from the firmware layer up to the API gateway.

[7] India Data Centers

The compliance backbone driving AI adoption

As AI adoption accelerates, the regulatory landscape is evolving to keep pace. The objective is clear: create guardrails that enable innovation while safeguarding data, infrastructure and societal interests. India's regulatory momentum reflects this balance, with frameworks designed to enforce accountability without stifling progress.

The security of India's AI future is inextricably linked to the regulatory framework, spearheaded by the DPDP Act, 2023, as well as certain published guidelines, such as the Reserve Bank of India's (RBI) FREE AI, Securities and Exchange Board of India's (SEBI) five-point AI Rulebook and the India AI Governance Guidelines. This legislation governs the processing of digital personal data within India and applies even to processing outside the country's territory if goods or services are provided to data principals within India.

The DPDP Act establishes a foundation for responsible data handling, mandating consent-driven processing, purpose limitation and stringent breach reporting. For AI systems that thrive on vast datasets, compliance is existential.

The framework places strong emphasis on data localisation, potentially requiring that specified categories of personal and traffic data, along with AI models built on such data, be stored locally. This strategic move simultaneously addresses national security concerns regarding cross-border data exposure and fosters domestic market growth. By mandating local hosting, the government incentivises the development of India specific models, favouring domestic providers over foreign alternatives.

As the National Centre for AI and Cybersecurity's (NCAIC) AI Governance Framework and CERT-In's AI Security Directives shape responsible AI adoption, the integration of AI systems into public administration also raises unique ethical considerations for a large, pluralistic democracy. GenAI systems, if not properly governed, increase risks such as deepfakes, which can lead to fraud, reputational harm and gender-based abuse. Therefore, any deployment of AI must uphold the principles for responsible AI outlined in recent policy discussions: safety, reliability, equality, inclusivity, privacy, security, transparency and accountability.

Reshaping India's critical infrastructure resilience will require a multilayered strategy aligned with frameworks such as the National Institute of Standards and Technology's Artificial Intelligence Risk Management Framework (NIST AI RMF), Zero Trust architectures and sector-specific cyber standards under India's National Critical Information Infrastructure Protection Centre (NCIIPC). As the country strengthens the AI security foundations, the next strategic imperative lies in examining the growing vertical attack surface where sector-specific AI dependencies introduce unique, high-impact vulnerabilities that demand equally tailored defence models.

Regulations and governance frameworks provide the scaffolding for trust. Translating these expectations into effective safeguards requires a clear understanding of how AI systems are built, deployed and operated in practice.

Across the AI lifecycle, where do key risks truly reside, and how exposed is the AI stack today?

Understanding the AI attack surface

AI security is a multi-layer challenge. Every component in the AI ecosystem, including user-facing applications and the physical infrastructure, introduces unique risks.

At the top lies the application layer, where APIs, prompts and integrations serve as entry points for adversarial manipulation. Under this layer, the ML models trained on vast datasets are vulnerable to poisoning, inversion and extraction attacks.

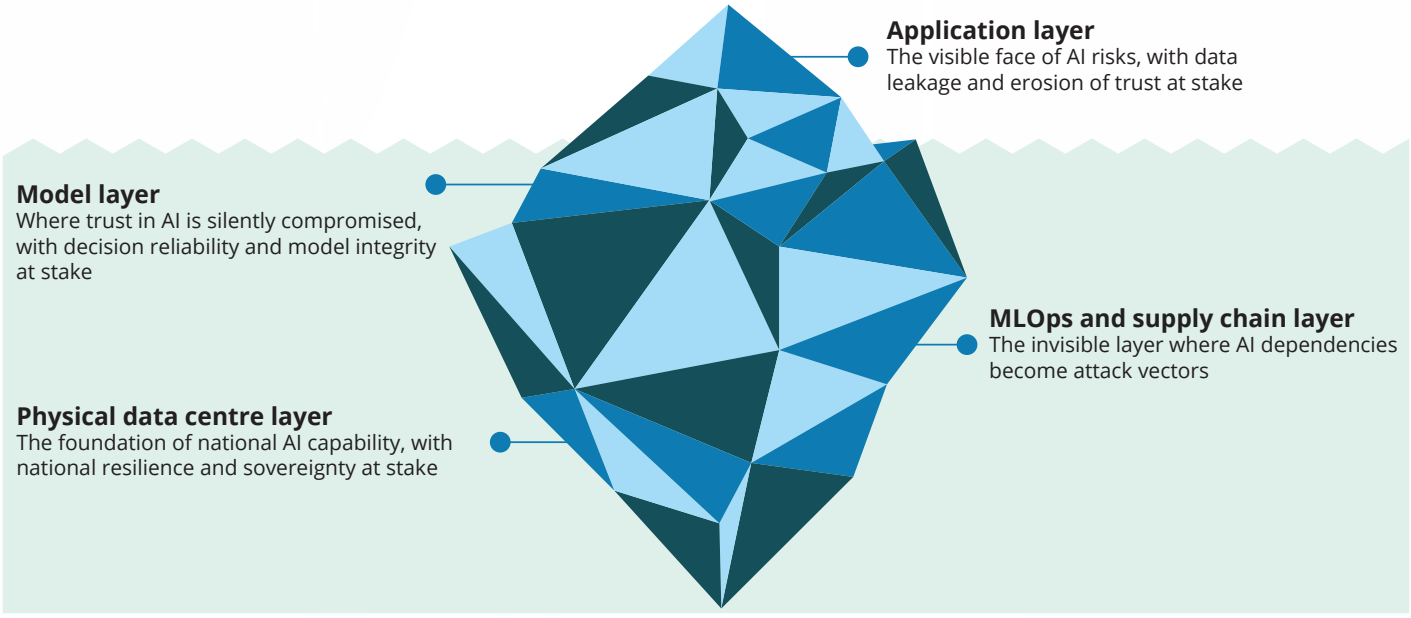
Further down, the MLOps pipeline, the orchestration of data, code and deployment presents risks through compromised registries, insecure Continuous Integration (CI)/Continuous Deployment (CD) workflows and supply chain exploits.

At the foundation level, the physical data centre and the hardware backbone are vulnerable to firmware tampering and geopolitical risks. To effectively defend against these threats, organisations must adopt a structured approach informed by global threat intelligence frameworks.

Application layer: GenAI and agentic AI at the edge
At the edge where users, APIs and agents interact, risk concentrates in a handful of high-impact patterns. These include:

- **Prompt injection and tool misuse:** Crafted inputs such as indirect injections from web pages, files or data sources can override guardrails, trigger unauthorised tool calls (search, code exec and Robotic Process Automation [RPA]) or exfiltrate sensitive context. This is the core LLM01 class of risk in the OWASP LLM Top 10 and is amplified in agentic AI that holds broad permissions.
- **Sensitive disclosure and system prompt leakage:** Output leakage of Personally Identifiable Information (PII), internal secrets or proprietary system prompts threatens compliance and IP.
- **Excessive agency and resource exhaustion:** Over-privileged agents (LLM06) combined with unbounded consumption (LLM10) can transform a single injection into cascading action across enterprise systems and push GPU/API costs or latency into denial-of-service territory.
- **Agent drift and goal misalignment:** Agentic AI systems can deviate from intended goals, leading to behaviours that fall outside the approved boundaries. This increases the risk of unsafe autonomy.

Figure 1: AI attack surface



Model layer: Threats to the core concepts of learning and inference

Below the application layer, the model itself becomes a target. MITRE ATLAS organises real-world tactics across reconnaissance, access, staging and exfiltration, and NIST's Adversarial Machine Learning (AML) taxonomy standardises language for poisoning, evasion and privacy attacks across predictive AI and GenAI systems. Some of these attack vectors include:

- **Poisoning and backdoors:** Contaminated training data or model artifacts degrade accuracy or embed triggers that only an attacker can activate later.
- **Evasion:** Carefully perturbed inputs cause misclassification or create unsafe outputs at inference without changing the training set.
- **Model drift:** As the real-world data shifts, the model's behaviour gradually changes, rendering a drift that creates silent failures. This causes the AI systems to make decisions that no longer align with their original intended outcome.
- **Extraction and inversion:** Systematic querying can reconstruct weights or infer sensitive elements from training data, directly threatening IP and privacy.

MLOps and supply chain: Registries, pipelines and provenance

AI hinges on a complex chain of data ingestion, CI/CD, model registries and serving stacks. This interconnectedness creates a supply chain where trust is fragile, and a breach in one layer can spread quickly to others.

Figure 2: Enterprise target operating model blueprint

Malicious artifacts and vulnerable formats
Pre-trained models and datasets often arrive as serialised files (e.g. Pickle) that can execute code on load. A single compromised artifact can infect the entire training environment.

Provenance blind spots
Without a verifiable AI Bill of Materials (AIBOM) and Software Bill of Materials (SBOM), organisations lack visibility into what's inside their models, including framework versions, datasets and dependencies, which makes patching and compliance nearly impossible.

Pipeline exploits and registry abuse
CI/CD workflows and model registries are prime targets for attackers seeking to inject malicious code or swap trusted models with tampered versions. Weak identity controls and unmonitored artifact promotion amplify the risk.

Cloud supply chain attacks
Shared infrastructure introduces threats such as CloudBorne (firmware exploits) and CloudJacking (instance hijacking), bypassing application-layer defences and striking at the virtualisation layer.

Physical data centre: The core

At the foundation, the AI data centre introduces risks that traditional north-south security was not designed to handle. Massive east-west flows between GPU clusters, storage and APIs, as well as firmware and driver complexity, widen exposure.

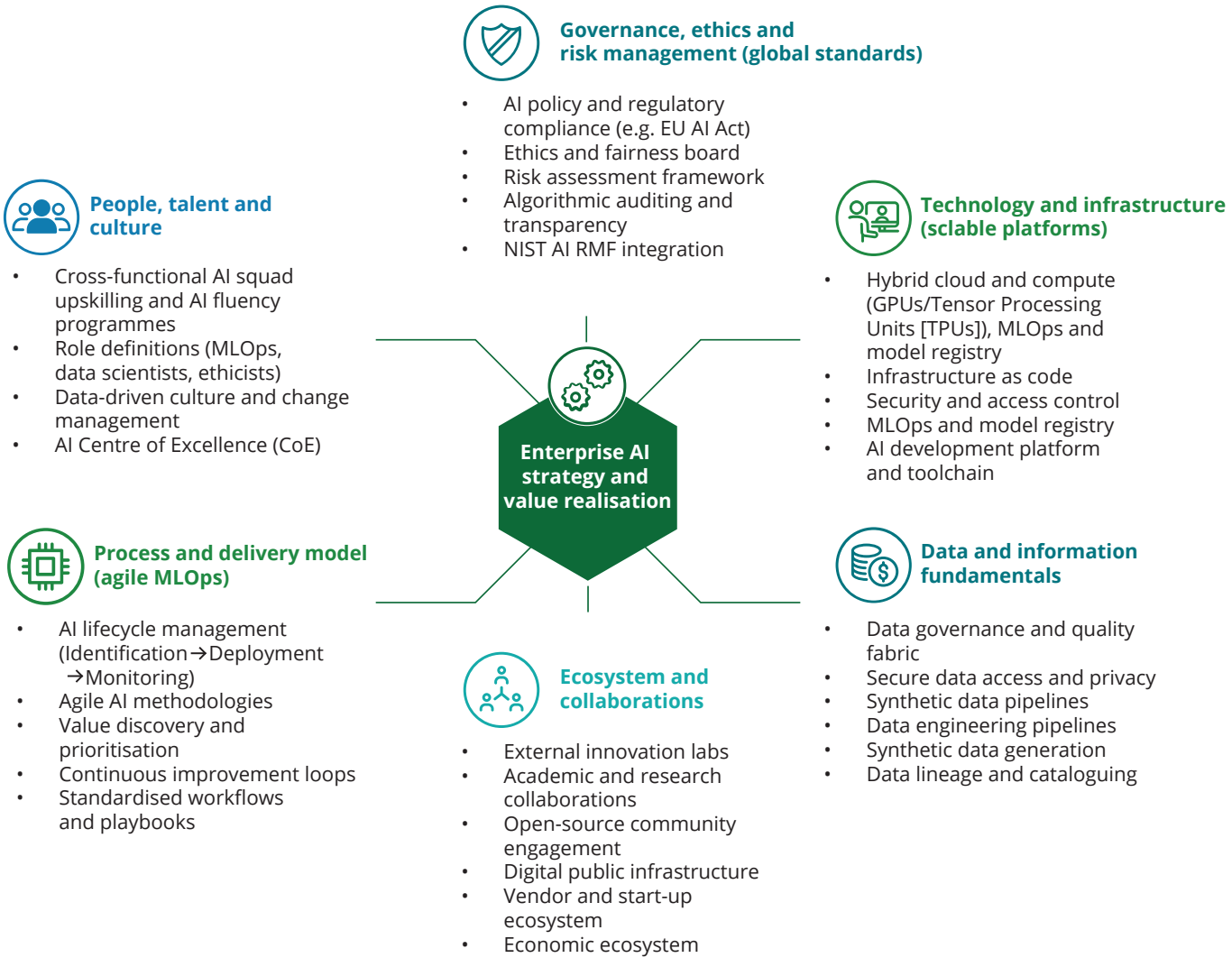
- **Firmware/driver trust:** A single exploit in GPU drivers or board firmware can enable lateral movement across high-performance clusters. Therefore, trust must be verified continuously.
- **Visibility and segmentation:** Hyperscale, low-latency east-west traffic needs inspection at speed and zero-trust segmentation to prevent propagation
- **Physical and geopolitical risk:** Supply chains, facility access and regional instability affect model availability and data sovereignty. New America outlines a six-layer framework (hardware to geopolitics) for AI data centre security.

Mapping the attack surface reveals the depth of complexity, but identifying risks alone is not enough. The next challenge lies in building a governance framework that anticipates and neutralises threats.

A unified AI governance framework for India

Effective AI security management for governments and large enterprises in India requires transitioning from ad hoc controls to a unified governance structure. It will help bridge domestic regulations such as the DPDP Act, NCAIC's AI Governance Framework and Ministry of Electronics and Information Technology's (MeitY) India AI Guidelines with global best practices such as ISO 42001, NIST AI RMF and 42005 while integrating the technical risk intelligence from frameworks such as MITRE ATLAS and OWASP LLM.

Figure 3: Enterprise target operating model blueprint



The synergy of structure and adaptability

Governance must balance structure and flexibility. ISO/IEC 42001 delivers the "how," acting as a formal, auditable AI Management System (AIMS) that enforces lifecycle controls, transparency and compliance. On the other hand, NIST AI RMF provides the "what" and "why" as a risk-adaptive framework that contextualises threats and prioritises mitigation based on impact.

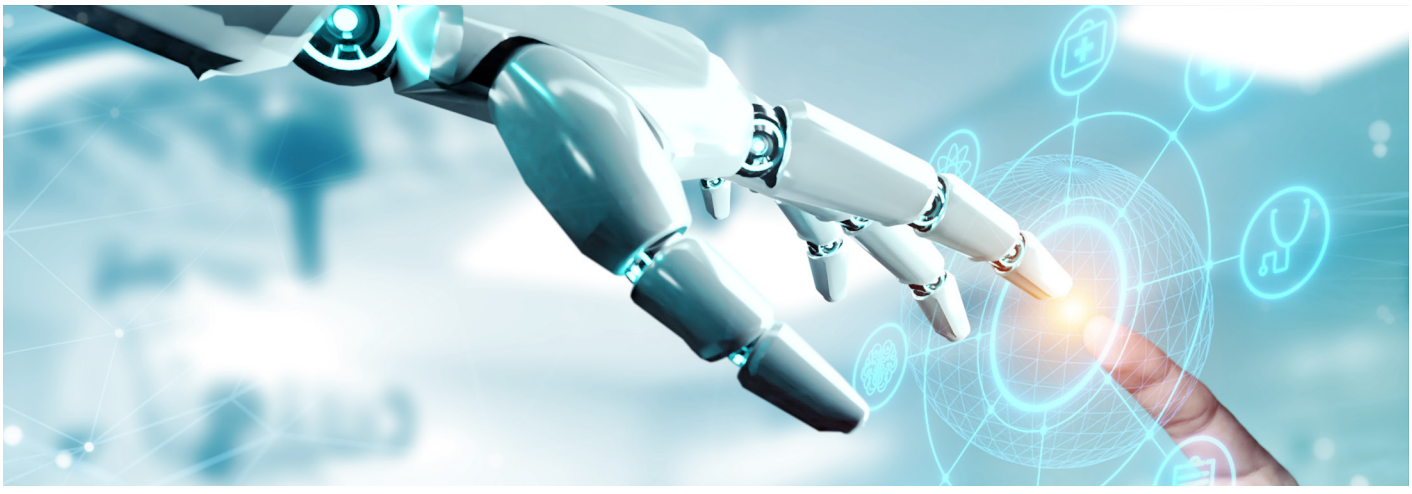
Together, they enable organisations to combine ISO's rigour with NIST's agility, ensuring both structured accountability and dynamic resilience.

ISO/IEC 42001: The auditable backbone

ISO/IEC 42001 defines how organisations establish, implement, maintain and continually improve an AI governance system. This is done while ensuring that AI innovations align with the enterprise's risk appetite, legal obligations, ethical expectations and stakeholder trust. It is a management and accountability framework designed for C-suite oversight.

For business leaders, the standard primarily offers the following strategic advantages:

1. **Trust, transparency and assurance**
   The standard compels organisations to document how data is used, how models behave, how decisions are explained and how risks are mitigated. It strengthens executive and stakeholder confidence.
2. **Structured AI risk management**
   ISO/IEC 42001 mandates systematic assessments for AI-specific risks, including bias, opacity, cybersecurity threats, drift, misuse and unintended impacts. It offers a proactive governance posture.
3. **Regulatory readiness and competitive differentiation**
   With accelerating global regulatory expectations, the standard demonstrates compliance discipline, reduces audit friction and positions the organisation as a responsible, trustworthy innovator.



The following pillars set out the core controls senior leaders should focus on to build AI security, trust and resilience:

A. Governance and accountability controls:
These controls clarify accountability, decision escalation pathways and AI alignment with organisational objectives.

Key actions
- Establish an AI governance authority with clear charters, decision rights and a cross-functional mandate.
- Define leadership responsibilities for oversight, impact assessment and ethical conduct.
- Ensure policy coherence between AI governance, cybersecurity and enterprise risk frameworks.

B. AI risk management controls:
These require structured identification, assessment and mitigation of AI-specific risks across the lifecycle,extending beyond what traditional cyber frameworks cover.

Key actions
- Implement continuous risk assessments for bias, drift, privacy leakage, adversarial attacks and misuse.
- Define risk thresholds aligned with regulatory expectations and sector-specific obligations.
- Integrate AI security risk management into enterprise risk dashboards and audit cycles.

C. Data quality, integrity and provenance controls:
Since AI quality is directly tied to data quality, the standard embeds explicit controls for lineage, validation and responsible access.

Key actions
- Ask for authenticated data sources, continuous validation and lifecycle documentation.
- Enforce data provenance tracking to ensure compliance, auditability and defensibility.
- Mandate data quality Key Performance Indicators (KPIs) tied to model performance and governance metrics.

D. Lifecycle controls for AI design, development and deployment:
The standard defines process-level controls that ensure AI systems are engineered with security, transparency and robustness from the outset.

Key actions
- Make secure-by-design the default approach.
- Ensure documentation of design decisions, assumptions and model limitations.
- Mandate testing and validation protocols for safety,robustness, explainability and performance.

E. Monitoring, performance evaluation and audit controls:
AI systems cannot be treated as "set and forget." ISO/IEC 42001 mandates continuous oversight.

Key actions
- Deploy ongoing model monitoring for drift, anomalous behaviour and security threats.
- Run structured internal audits, governance reviews and leadership updates.
- Maintain logs and evidence trails for external certification.

F. Security and information protection controls:
AI introduces new security surfaces, such as pipelines, models and data flows, that require targeted hardening.

Key actions
- Enforce controls to protect models, data, APIs and training pipelines.
- Map vulnerabilities across the AI lifecycle in training, inference, supply chain and access interfaces.
- Align with relevant security standards applicable to your business for cyber controls.

G. Ethical, transparent and societal impact controls:
AIMS requires organisations to demonstrate that AI decisions are explainable, fair and aligned with ethical principles.

Key actions
- Establish governance for fairness, non-discrimination and transparent decision-making.
- Make model behaviour and data usage explainable to regulators, customers and employees.
- Conduct AI impact assessments for high-risk use cases.

NIST AI RMF: The risk-adaptive guide

NIST AI RMF is built on a socio-technical approach, recognising that AI risks encompass ethical, legal, social and technical implications. Designed to be adaptable to organisations of all sizes and risk profiles, the framework enables them to tailor its principles according to their specific needs.

When combined, organisations can utilise the framework's flexibility to prioritise high-risk AI systems and then use the ISO 42001 structure to build a formal, auditable system governing those prioritised processes. This dual approach ensures both structured transparency and dynamic resilience against emerging threats.

Operationalising NIST AI RMF in government and enterprise

NIST AI RMF is operationalised through four core functions that guide continuous risk management throughout the AI system lifecycle: govern, map, measure and manage.

1. Govern (accountability and policy)

The function establishes the necessary structures, roles and responsibilities, ensuring that the AI risk management process aligns with organisational goals, ethical standards and legal obligations. It also helps define policies on transparency, accountability and fairness, keeping AI use within legal and ethical limits.

2. Map (contextualisation)

The function requires organisations to identify the AI system's context, scope and affected stakeholders. This step is vital for adoption in government departments in India, where the impact of AI systems on diverse demographics must be carefully understood to prevent the deployment of biased algorithms in critical public services.

3. Measure (testing and validation)

The function focuses on establishing metrics and procedures to assess and track the trustworthiness of an AI system, including its validity, reliability, safety, security and fairness. At this stage, technical defences are rigorously tested and documented.
The key requirement under this function is the explicit mandate for adversarial testing. Measure 2.7 requires organisations to use red-team exercises to actively test the system under adversarial or stress conditions, evaluate its response, assess failure modes or determine if it can return to normal function after an unexpected adverse event. This requirement for documented, active red-teaming establishes a clear regulatory necessity for sophisticated security testing solutions capable of simulating MITRE ATLAS and OWASP LLM attacks.

4. Manage (prioritisation and response)

The function involves prioritising identified risks, implementing mitigation actions, monitoring the system and ensuring robust response plans are in place to address failures or adverse events.

ISO 42005: Mandating the impact assessment

ISO/IEC 42005 (AI System Impact Assessment [AISA]) is the crucial companion standard to ISO 42001, providing a structured process for evaluating the societal, ethical and practical impacts of AI systems.

The requirement to perform an AISA is increasingly common in emerging global regulations, highlighting its importance for Indian entities seeking international compliance. ISO 42005 mandates the following:

- **Structured assessment:** Organisations must establish a consistent approach for performing and documenting impact assessments, tailoring them to their specific context and risk appetite.
- **High-risk trigger identification:** The standard mandates organisations to define specific thresholds and triggers to determine when a full impact assessment is necessary, focusing particularly on high-risk AI systems.
- **Continuous review:** The AISA is a continuous process requiring monitoring and review over time to ensure alignment between AI governance and real-world outcomes.

By adopting ISO 42005, Indian enterprises and government agencies gain a defensible structure for making informed, ethical decisions about their AI systems, especially those deployed in sensitive areas where potential bias or social harm is a concern. The following table summarises the governance synergy achieved through adopting these integrated frameworks:

Table 1: AI governance synergy: Mapping global standards to Indian imperatives

Framework	Governance focus	Mechanism	Strategic value to Indian organisations
ISO/IEC 42001	Management system structure	AIMS, annex A controls	Provides the auditable structure required for compliance with the DPDP Act, which helps address resource gaps (A.4) and maintain data quality (A.7)
NIST AI RMF	Risk-adaptive operation	Govern, map, measure, manage functions	Offers a flexible approach for prioritising efforts (risk-based) and implementing continuous threat monitoring tailored to the evolving Indian AI use cases
ISO/IEC 42005	Impact assessment	AISA (assessment process), defining risk thresholds	Helps meet ethical requirements (transparency and non-discrimination) by proactively assessing social and ethical impacts in a pluralistic society

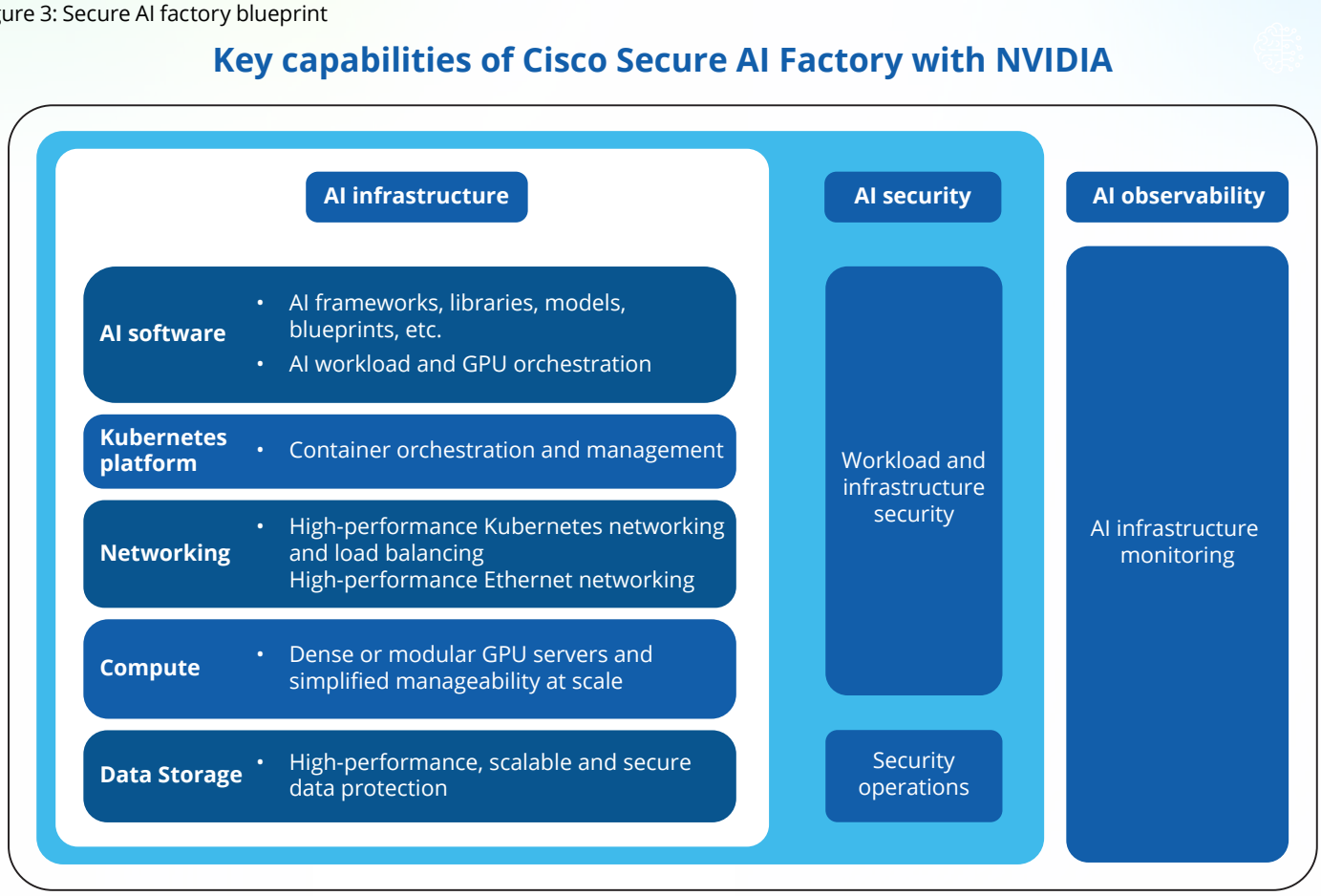
Governance provides the blueprint, but it is not sufficient to prevent breaches. The real test of resilience lies in the physical and logical backbone of AI, i.e. the data centre.

When compute becomes the new currency of intelligence, how can the vault be secured?

Securing the AI data centre: Zero Trust and physical integrity

India's AI infrastructure is expanding rapidly, with critical workloads such as national security applications and DPDP-regulated financial services. This calls for a stronger approach to data centre security. Traditional perimeter-based models are inadequate for dynamic, high-performance GPU clusters. The defence strategy must be rooted in Zero Trust Architecture (ZTA) and specialised hardening of the hardware and network fabric.

Figure 3: Secure AI factory blueprint



Core ZTA principles in AI deployment

1. **Continuous verification:** Every access request from a user, device or AI workload must be authenticated, authorised and continuously validated based on real-time activity and risk assessment. This includes dynamically modifying access privileges when abnormal behaviour is detected.

2. **Micro-segmentation:** This principle uses granular network segmentation to restrict access to only the specific resources needed by an application. In AI clusters, micro-segmentation is vital for isolating sensitive assets. Training workloads running in high-performance GPU clusters must be segmented away from latency-optimised inference services. Furthermore, access to proprietary training datasets and model outputs must be restricted only to authorised resources, drastically reducing the blast radius in the event of a breach.

3. **Least privilege access:** Users, services and devices are granted only the minimum access necessary to perform their specific tasks. This minimises the risk of exploitation in case a credential or API key is compromised

Hardening GPU cluster integrity and data isolation

High-performance AI workloads require specialised techniques to ensure data isolation, model integrity and compliance with strict data-residency requirements under the DPDP Act.

Environmental and data hygiene

For the most sensitive models or those handling regulated data, leaning towards **single-tenant bare-metal environments** offers superior control over Operating System (OS) hardening, firmware versions and network segmentation compared to multi-tenant or shared cloud architectures.

Within all environments, strict data hygiene practices are required to prevent LLM02 (sensitive information disclosure):

- **GPU memory clearing:** In shared container environments (e.g. Kubernetes), GPU memory must be zeroed out after use to ensure residual data from one job cannot be accessed by subsequent workloads.
- **Log sanitisation:** Organisations must avoid logging raw inputs, prediction results or internal errors that could inadvertently leak sensitive content. Defined log retention and rotation policies must be implemented to minimise exposure.

Supply chain integrity and firmware security

Given the inherent vulnerabilities in the MLOps supply chain, securing the hardware foundation is mandatory. Organisations must acknowledge that a firmware exploit in a shared GPU driver could allow an attacker to achieve lateral movement across the entire cluster. To mitigate this risk, security strategies must include the proactive verification of hardware trust and component firmware before deployment, countering the risk of inheriting compromised infrastructure from third-party providers.

Securing high-speed interconnects in AI data centres (Ethernet and InfiniBand)

Modern AI data centres rely on ultra-low-latency, high-bandwidth interconnects such as InfiniBand with RDMA or high-performance Ethernet fabrics with RoCEv2 (RDMA over converged Ethernet) to enable large-scale AI training and inference. While these technologies accelerate data movement between GPUs and storage systems, they also introduce unique security challenges. Since RDMA and similar mechanisms bypass the host CPU and OS, traditional security monitoring and enforcement models become ineffective. The historical assumption of a trusted HPC environment no longer applies in multi-tenant, cloud-scale AI infrastructures.

Hardware-enforced network security

To secure these high-speed fabrics without compromising performance, security must be embedded at the hardware level and enforced by network adapters and switch Application-Specific Integrated Circuits (ASICs) for both backend (inter-GPU to GPU) and frontend (GPU nodes to the data centre for inference) networks.

Ethernet and RDMA security enhancements

For Ethernet-based AI fabrics using RoCEv2 (RDMA over Converged Ethernet), security is addressed through Virtual Routing and Forwarding (VRF), Virtual Local Area Network (VLAN) segmentation, Media Access Control Security (MACsec) encryption and identity-based access controls at the Network Interface Card (NIC) and switch level. Hardware-based isolation and cryptographic enforcement help maintain Zero Trust principles across shared environments.

Zero Trust for AI infrastructure

Data centre security architects should prioritise accelerated networking platforms that natively integrate distributed, hardware-level security. This approach ensures Zero Trust security without sacrificing performance, protecting sensitive data and IP residing in GPU memory and high-speed storage systems.

Securing AI through our unified and layered defence

Securing the AI ecosystem requires a cohesive operating model that fuses prescriptive governance (DPDP, MeitY and NCAIC) with threat-informed technical defences (OWASP LLM, MITRE ATLAS and NIST AML). The goal is end-to-end visibility, validation and real-time enforcement without sacrificing the speed and scale of high-performance AI workloads. Deloitte provides the unified controls framework and leads gap assessment and assurance, while Cisco delivers the runtime protection, access governance and multi-cloud visibility. Together, these capabilities help apply best practices across the AI lifecycle.

Bridging governance and execution

Cisco AI Defense is engineered to address the growing attack surfaces of GenAI and agentic AI and to mitigate "shadow AI" risks through three core, integrated components that map directly to the requirements of global governance frameworks. The solution focuses on discovery, detection, protection and policy enforcement, minimising legal and reputational risks.

Cisco AI Defense and the NIST/ISO mandate

These core functions of Cisco AI Defense demonstrate compliance with the NIST AI RMF and ISO 42001 mandates:

1. AI model and application validation

This component focuses on security validation before deployment. It utilises AI-driven red teaming to perform automated vulnerability testing on models.

- **Operationalising the NIST measure function (M 2.7):** The NIST AI RMF mandates the use of red teaming exercises to actively test the system under stress conditions and assess failure modes. Algorithmic red teaming automates this process, using advanced techniques such as the Tree of Attacks with Pruning (TAP) to rapidly simulate billions of malicious interactions and uncover weaknesses before real-world exploitation. This rigorous, automated testing directly satisfies the NIST mandate for measuring robustness and resilience.
- **Mitigation:** This addresses poisoning attacks (AML), vulnerable pre-trained models and weak provenance in the MLOps supply chain.

2. AI runtime protection (enterprise AI security and governance platform)

AI runtime protection applies continuous validation and enforcement mechanisms to protect AI applications in production, using the network infrastructure. This component is based on an AI firewall concept that monitors and analyses inputs and outputs in real-time.

Mitigating application-layer threats (OWASP LLM top 10):

- **Prompt injection (LLM01):** The AI firewall intercepts API traffic, examines user inputs for malicious prompts and prevents the model from executing unintended commands. If a user attempts a jailbreak, Cisco AI Defense detects the attempt and ensures the Large Language Model (LLM) responds with a predefined guardrail message.
- **Sensitive data disclosure (LLM02/LLM07):** By examining model outputs, the firewall can stop the misstep before sensitive data is leaked, enforcing Data Loss Prevention (DLP) policies embedded at the network level.
- **Supply chain vulnerabilities (LLM03):** As AI deployment becomes more accessible and efficient, third-party AI assets introduce risks, which create opportunities for code execution, sensitive data exfiltration and model inversion. Runtime guardrails protect against these scenarios by providing bi-directional filtering and inspection, acting as a Model Context Protocol (MCP) gateway for agents, ensuring secure interactions with enterprise AI applications.
- **Operationalising the NIST manage function:** This provides the real-time defence layer necessary for prioritising and addressing threats detected in production environments, fulfilling the manage function's requirement for continuous monitoring and response.

3. AI access management and cloud visibility

These components provide the necessary oversight and policy enforcement across the enterprise, managing who can interact with AI resources and tracking AI usage across multi-cloud environments.

- **Governing shadow AI:** AI access management identifies and monitors AI-enabled applications, assessing the risk severity associated with unauthorised (shadow AI) tools. It enforces granular policies to restrict access to unsanctioned AI, preventing the outflow of sensitive corporate data through insecure third-party applications.
- **ISO 42001 (A.4 resources):** AI cloud visibility provides deep insights and telemetry across cloud environments to track AI workload usage and data exchanges. This telemetry supports the auditable documentation required under ISO 42001 control A.4 for identifying and managing computing resources that are crucial to the AI system's lifecycle.

The following table provides a detailed cross-mapping of the integrated defence strategy:

Table 2: Mapping Cisco AI Defense to global compliance and threat frameworks

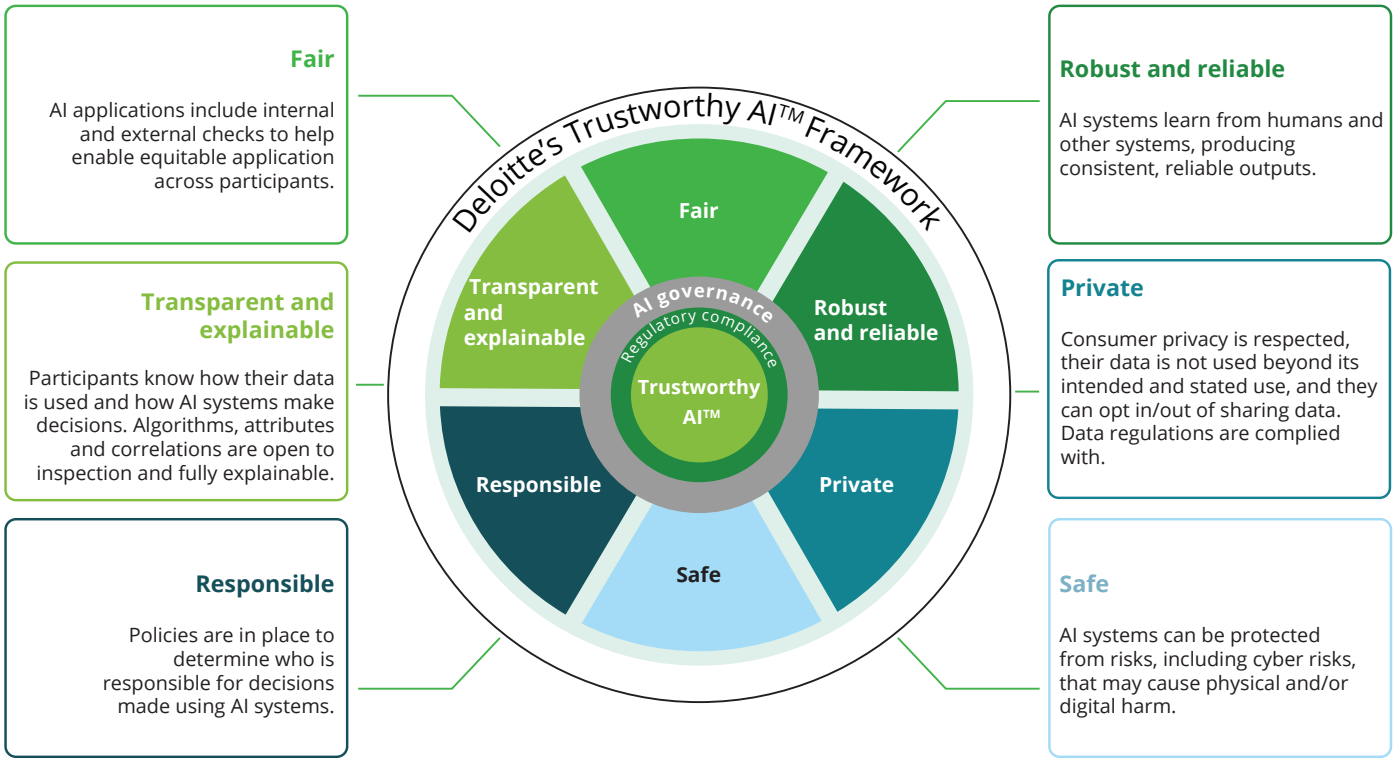
Governance framework/threat vector	Key requirement/risk	Cisco AI Defense component	Mitigation mechanism
NIST AI RMF: Measure (M 2.7)	Mandated algorithmic red teaming and stress testing	AI model and application validation	Automated, rapid simulation of adversarial attacks (e.g. TAP) to find pre-deployment model vulnerabilities
OWASP LLM01: Prompt injection	Subversion of model intent via malicious input	AI runtime protection (AI firewall)	Real-time traffic interception and analysis of inputs to detect and block malicious prompts before they reach the LLM
OWASP LLM02/07: Data leakage	Disclosure of sensitive training data or system prompts	AI runtime protection (AI firewall)	Enforcement of DLP policies on model outputs to stop sensitive data exfiltration in real time
MITRE ATLAS: Exfiltration/AML poisoning	Stealing model IP or corrupting training data integrity	AI model and application validation, AI cloud visibility	Automated testing to find backdoor injections; telemetry to track unauthorised data exchanges
ISO 42001 (AIMS) and DPDP Act	Auditable policy enforcement and governance for shadow AI	AI access management	Enforcement of granular organisational policies and access restriction to unsanctioned AI tools, mitigating unauthorised data sharing

Bringing governance to life with Deloitte's Trustworthy AI™ Framework

Deloitte's Trustworthy AI™ Framework embeds responsibility, resilience and compliance into AI programmes using a structured approach. It aligns with India's regulatory requirements and global standards, enabling safe and responsible AI use while supporting robust risk management.

The framework focuses on the following key principles:

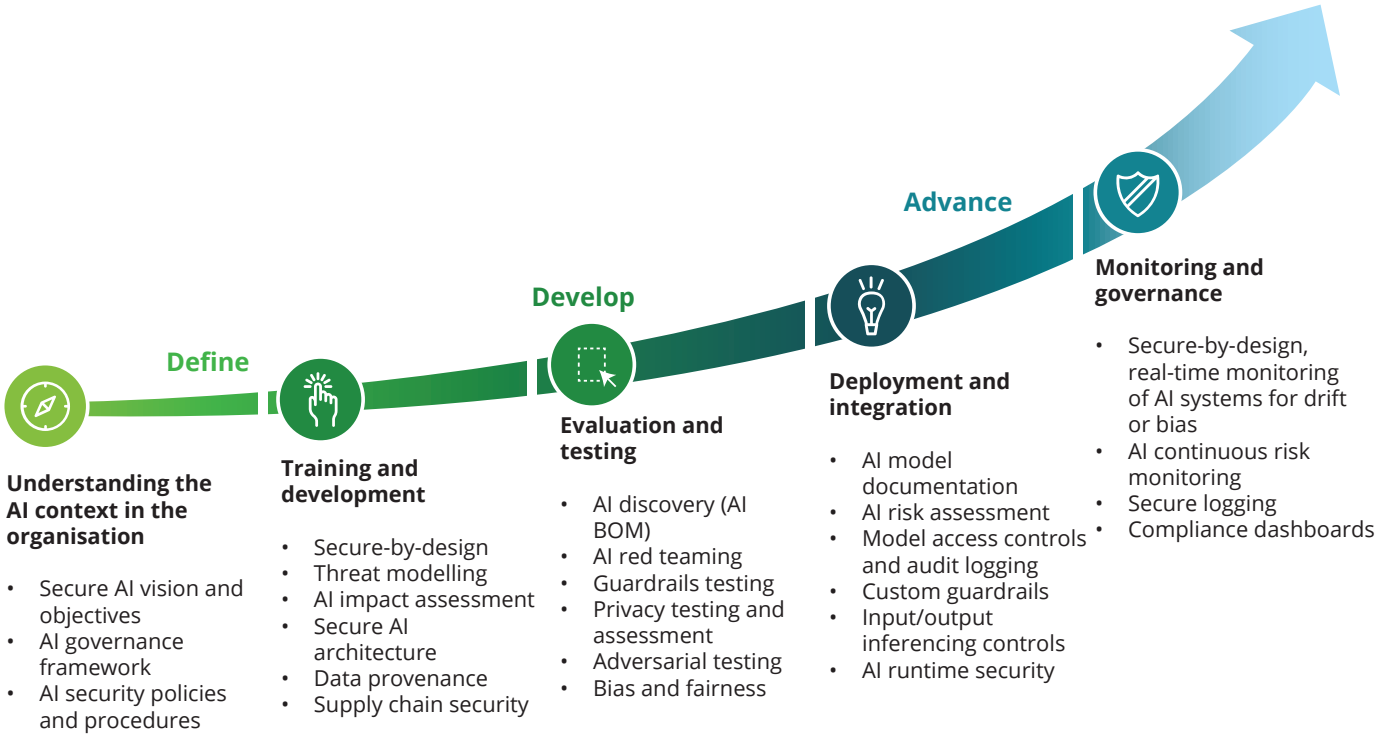
Trustworthy AI™ framework is an effective approach to managing AI quality and risk, which can be integrated into existing enterprise risk management efforts.



True resilience emerges when these principles converge with real-time technical enforcement. The next step is a synchronised approach in which Deloitte's governance expertise and Cisco's AI defense capabilities unite to deliver layered security across the entire AI lifecycle.

A phased roadmap for secure AI adoption in India

Securing AI is an ongoing journey. The scale of the required transformation, coupled with the talent constraint in India's cybersecurity sector, requires a phased, resource-optimised roadmap for adopting integrated AI security and governance.



Phase I: Governance foundation, visibility and readiness

- **AI governance and readiness assessment:** Conduct an AI gap and maturity assessment aligned with ISO 42001, NIST AI RMF and domestic regulations, including DPDP Act and NCAIC's India AI Governance Framework, to identify governance gaps across data, models and operations.
- **AIMS set up:** Initiate the process for ISO 42001 certification readiness. Define clear governance structures, roles and responsibilities (governance function), focusing on documenting critical resources and expertise.
- **Contextual mapping:** Use the NIST AI RMF Map function to identify and contextualise all existing AI systems, prioritising those that qualify as high-risk or Significant Data Fiduciaries (SDFs) under the DPDP Act.
- **Shadow AI discovery and control:** Deploy AI access management to gain immediate visibility over all third-party and unsanctioned AI applications used internally, preventing initial data leakage risks immediately.

Phase II: Validation, assurance and infrastructure hardening

- **AI assurance and validation:** Establish AI model and application validation for high-risk AI systems. Integrate algorithmic red teaming and adversarial testing into the MLOps pipelines and produce audit-ready evidence.
- **Supply chain risk assessment:** Assess training pipelines, model provenance and third-party dependencies to identify supply chain risk exposures across software, models and tooling.
- **Zero Trust deployment:** Implement Zero Trust micro-segmentation across GPU clusters. Simultaneously, work with data centre providers to enforce hardware-based security controls (Global Unique Identifiers [GUIDs], M/P-Keys) on high-speed RDMA/InfiniBand fabrics to secure the physical backbone.

Phase III: Runtime protection and continuous management

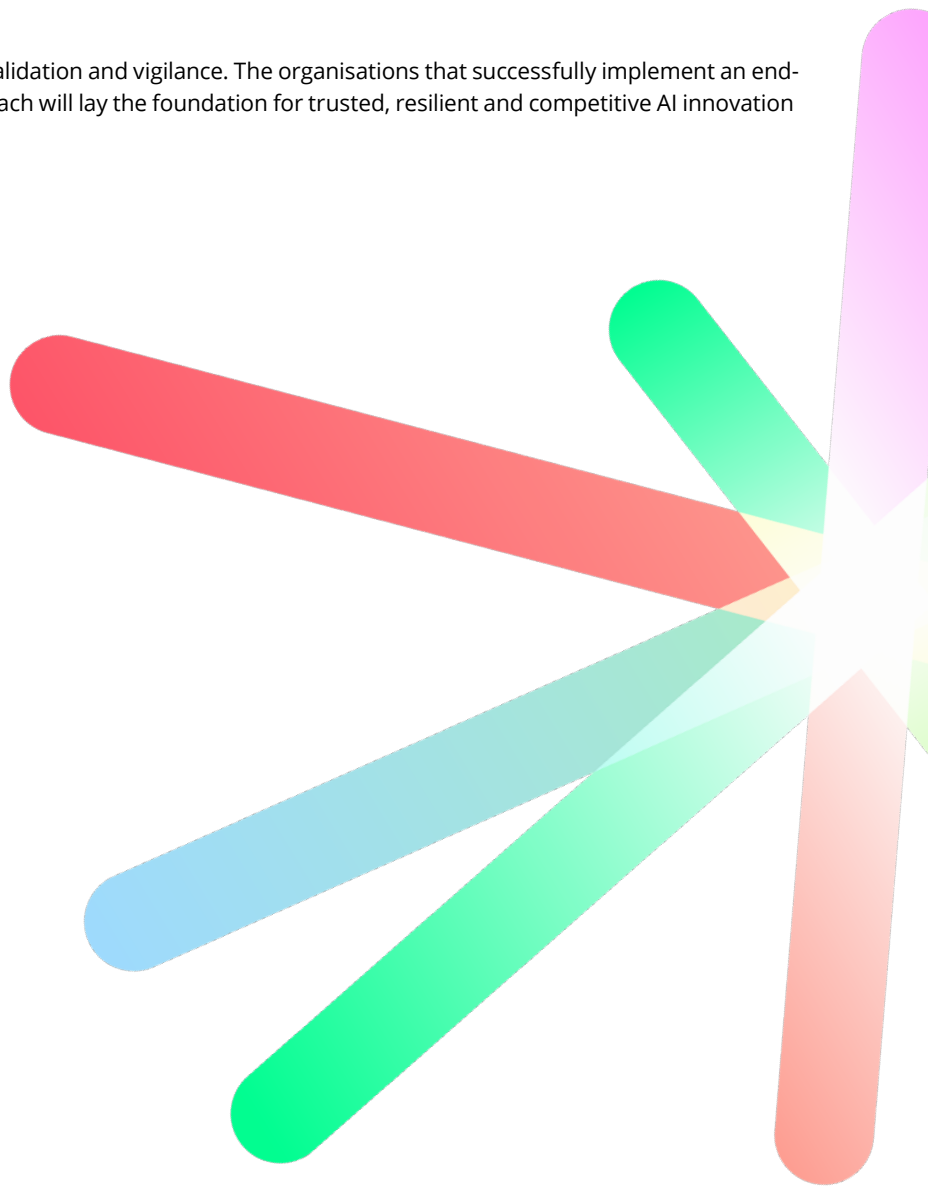
- **Full runtime defence**: Deploy AI Runtime Protection (AI Firewall) across all production AI applications to intercept and mitigate threats such as prompt injection and data exfiltration in real time.
- **Continuous monitoring:** Establish AI-specific monitoring metrics, incident response playbooks and governance review cycles to adapt based on real-world outcomes and societal impacts. This ensures the defence posture remains relevant and compliant with evolving expectations for ethical AI usage in India.
- **Ongoing impact and ethics assessment:** Conduct periodic impact assessments in alignment with ISO/IEC 42005 to evaluate societal impact, bias and fairness outcomes in coherence with India's responsible AI principles.

Conclusion: The path to sovereign AI trust

India stands at an inflexion point where its accelerated adoption of AI must be secured by integrated, high-assurance governance and technology. The sheer scale of AI deployment across critical sectors, coupled with the regulatory pressure of the DPDP Act and the strategic need for sovereign data control, elevates AI security to a matter of national resilience.

The inherent risks exposed by the immaturity paradox in enterprise governance, the specific threats identified by OWASP LLM and MITRE ATLAS, and the supply chain vulnerabilities within MLOps and data centre hardware cannot be addressed by traditional cybersecurity measures.

Securing India's AI horizon is a journey of continuous validation and vigilance. The organisations that successfully implement an end-to-end, standards-based and technology-driven approach will lay the foundation for trusted, resilient and competitive AI innovation on the global stage.



References

01. Transforming India with AI - PIB
02. ISO 42001 Standard for AI Governance and Risk Management | Deloitte US
03. Cybersecurity Skill Gap: Rising Threats Demand Talent - Deccan Herald
04. Why India Is Mandating Local Storage of AI Models Under DPDP Act: Data Sovereignty, Cybersecurity & Cross-Border Data Protection Explained
05. Localization: India's Tryst with Data Sovereignty
06. India pushes AI model localization to prevent data leaks - Tech in Asia,
07. AI Ethics in Public Policy: Case Studies & Challenges | ISPP
08. Artificial Intelligence and Ethics: An Indian Perspective for a Trusted Digital Future - NeGD,
09. AI Watch: Global regulatory tracker - India | White & Case LLP
10. Data Centre industry in India - Wikipedia
11. Securing RDMA for High-Performance Datacenter Storage Systems
12. OWASP Top 10 for LLM Applications & Generative AI: Key Updates for 2025
13. Cisco AI Defense: Securing the Safe Transformation of Enterprise AI Applications
14. MITRE ATLAS Framework
15. Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations - NIST Technical Series Publications
16. LLM03:2025 Supply Chain - OWASP Gen AI Security Project
17. Researchers Identify Over 20 Supply Chain Vulnerabilities in MLOps Platforms
18. Securing LLMs Starts at the Hardware & Firmware Layer
19. Integrating the NIST AI RMF and ISO 42001
20. ISO 42001 Annex A Controls Explained – ISMS online
21. NIST AI Risk Management Framework
22. AI governance unpacked: ISO 42001 & NIST AI RMF explained
23. How To Align with the NIST AI RMF: Step-by-Step Playbook
24. Framing the Risk Management Framework: Actionable Instructions by NIST in The "Measure" Section of the AI RMF
25. Elements of the NIST AI RMF: What you need to know
26. ISO/IEC 42005:2025 – A New Blueprint for Legal and Commercial Leaders Navigating AI Risk and Governance
27. What is ISO 42005? The AI Impact Assessment Standard you need to know
28. Zero Trust in the Age of AI: Securing Cloud Environments Against Evolving Threats - ISACA
29. Securing AI Workloads in Kubernetes
30. AI Cybersecurity Solutions for your Business - NVIDIA
31. Security considerations when hosting AI models on GPU

32. Top 5 Best Practices for Securing GPU-Accelerated AI Cloud Environments,
33. InfiniBand Multilayered Security Protects Data Centers and AI Workloads | NVIDIA Technical Blog
34. Securing AI with Cisco AI Defense
35. Research Note: Cisco AI Defense
36. Cisco Tackles AI Security with New AI Defense App - Aragon Research
37. What is Algorithmic AI Red Teaming? - Robust Intelligence,
38. Defending the Network Against AI Threats: Cisco AI Defense
39. Cisco AI Defense: A New Solution for Securing GenAI in Enterprise
40. Breaking Down AI Security Barriers with Cisco AI Defense

About Deloitte

Deloitte is one of the world's largest and most diversified professional services organisations, providing assurance and advisory, tax, management consulting, and enterprise risk management services through more than 345,374 professionals in more than 150 countries. Our organisation includes a unique portfolio of competencies integrated in one industry-leading organisation. Deloitte Touche Tohmatsu India LLP (DTTI LLP) also referred as Deloitte India is a member firm in India that provides non-audit consulting services. Our experienced professionals deliver seamless, consistent services wherever our clients operate.

In India, Deloitte is spread across 12 cities with over 12,000 professionals, who are proficient at delivering the right combination of local insight and international expertise to our clientele drawn from across industry segments. Deloitte is well-equipped to deliver solutions to the complex challenges faced by organisations across the public and private sectors. Our edge lies in our ability to draw upon a well- equipped global network and teaming this with customised services at a local office.

We have been consistently recognised as leaders by Gartner in Cybersecurity, the Data and Analytics space, as well for Public Cloud Infrastructure Managed and Professional Services and Oracle Cloud Application Services.

https://www2.deloitte.com/in/en.html

About Cisco

Cisco is the worldwide technology leader that securely connects everything to make anything possible. Our purpose is to power an inclusive future for all by helping our customers reimagine their applications, power hybrid work, secure their enterprise, transform their infrastructure, and meet their sustainability goals. Cisco delivers the critical infrastructure to help organisations thrive in the AI era. By fusing networking, security, observability, and collaboration, we power how people and technology work together.

https://www.cisco.com/site/in/en/about/index.html

Connect with us

Deloitte India

Gaurav Shukla
Partner & Leader - Cyber
Deloitte South Asia
shuklagaurav@deloitte.com

Dhruv Khanna
Partner
Deloitte India
dhruvkhanna@deloitte.com

Anand Tiwari
Partner
Deloitte India
anandtiwari@deloitte.com

Ashish Sharma
Partner
Deloitte India
sashish@deloitte.com

Munjal Kamdar
Partner
Deloitte India
mkamdar@deloitte.com

Mayuran Palanisamy
Partner
Deloitte India
mayuranp@deloitte.com

Cisco India

Arun Shetty
Sr. Director & CTO,
Cisco India & South Asia
acshetty@cisco.com

Vinay Jain
Director - Cloud & AI,
Cisco India & South Asia
vinayj@cisco.com

Ashok Shivshankar
Managing Director- Alliances,
Cisco India & South Asia
ashivas@cisco.com

Ninad Katkar
Director - Security Sales,
Cisco India & South Asia
nikatkar@cisco.com

Samir Mishra
Managing Director- Sales
Cisco India & South Asia
samimish@cisco.com

Contributors

Deloitte India

Hiten Panchal

Anupa Subrahmoni

Swetha Kumar

Cisco India

Tarandeep Bagga

Bhishma Naraya Sharma

Vivek Alexander

Acknowledgements

Sonali Lingwal

Sunita Kumari

Ruchira Thakur

This co-authored whitepaper/POV applies to Cisco and Security products described in the Cisco Services Summary. The content contained herein is correct as of December 2025 and represents the status quo as of the time it was written. Cisco's security policies and systems may change going forward, as we continually improve protection for our customers

This POV is prepared by Deloitte Touche Tohmatsu India LLP (DTTILLP) and Cisco India.