WHITE PAPER | 2026

From Pilot to Production:

Managing AI Risk Across the AI Lifecycle



A Joint Publication by

Asia-ISAC

RAIDS AI™
Intelligence Under Watch

Disclaimer

This document is published by Asia-ISAC and RAIDS AI as a contribution to the understanding of AI risk governance and operational monitoring. The findings, interpretations, and conclusions expressed herein result from a collaborative process facilitated and endorsed by Asia-ISAC and RAIDS AI, but do not necessarily represent the views of Asia-ISAC or RAIDS AI in their entirety, nor those of their respective Members, Partners, or other stakeholders.

This white paper is intended for informational purposes only and does not constitute legal, regulatory, or professional advice. Asia-ISAC and RAIDS AI make no representations or warranties regarding the accuracy, completeness, or applicability of the information contained herein to any specific organisation, jurisdiction, or use case. Readers should seek independent professional advice before acting on any content in this document.



© 2026 Asia-ISAC & RAIDS AI LTD. A joint publication — all rights reserved.

Contents

I. Executive Summary: The Runtime Risk Problem....4
Contextual Framing.... 4
About RAIDS AI....5
About Asia-ISAC....5
The Structural Failure....5
Inference-Time Intelligence....6
The Shared Thesis.... 6
II. The AI Lifecycle Risk Transformation Model.... 7
Phase 1: Pre-Production — Governance Design.... 7
Phase 2: Live Runtime — Operational Behaviour (Emergent Risk).... 8
Phase 3: Incident Response — Operational Risk Control (Materialised Risk).... 9
Phase 4: Audit and Compliance — Assurance Layer (Retrospective Risk).... 10
III. Operationalising Governance: Runtime Monitoring and the RAIDS AI Capability Model. 11
Pillar 3.1: Inference-Time Monitoring — Closing the Visibility Gap....11
Pillar 3.2: Non-Invasive Visibility — The Black-Box Imperative.... 11
Pillar 3.3: Dynamic Telemetry — Adaptive Baselines vs. Static Thresholds....12
IV. Asia-ISAC Perspective: High-Complexity Operating Environments.... 14
The Global Stress Test: Asia as the Regulatory Frontier....14
The Observability Tension: Sovereignty Versus Visibility.... 15
Enterprise Case Study 1 (Illustrative Scenario): Shadow AI Remediation in Asian Enterprises....15
Enterprise Case Study 2: Regional FinTech — Continuous Compliance Under MAS FEAT & Regional Rules.... 16
Enterprise Case Study 3 (Illustrative Scenario): Regional Logistics — Compound Drift in High-Velocity Asian Supply Chains....17
V. The Human Element: Strengths, Weaknesses, and the Governance Imperative....18
Human Strengths in AI Adoption and Operation....18
Human Weaknesses in AI Adoption and Operation.... 19
Designing for Human Performance: Governance Implications....22
VI. Conclusion: The Infrastructure for Industrial-Grade AI....23
The Execution Engine....23
Institutional Readiness: From Safe at Launch to Safe in Operation....23
The Blueprint for Scale.... 23

© 2026 Asia-ISAC & RAIDS AI LTD. A joint publication — all rights reserved.

I. Executive Summary: The Runtime Risk Problem

Contextual Framing

Enterprise artificial intelligence has entered a phase of institutional reckoning. After years of accelerating investment and prolific pilot programmes, the gap between deployment ambition and operational reality has become empirically undeniable. The Stanford AI Index 2025 documented 233 discrete AI-related incidents in 2024 — a 56.4% increase over the prior year — with the majority occurring not in experimental settings but in systems already operating in production environments.[1] Concurrently, McKinsey's State of AI 2025 reveals a stark paradox: while 88% of organisations now deploy AI in at least one business function, fewer than 6% qualify as high performers achieving measurable bottom-line impact.[2] The conclusion is unambiguous: the challenge of 2026 is not adoption — it is operational integrity.

56.4%
Rise in AI incidents, 2024 vs. 2023

88% / 6%
Deploy AI — vs. high performers

42%
Enterprises abandoning most AI projects

This paper brings together governance perspectives from Asia-ISAC, an independent regional industry body, and technical operational insights from RAIDS AI to address what the authors term the Governance Latency Gap — the critical structural delay between a model's initial certification and its emergent, often unpredictable behaviour in a live production environment. Realising sustainable scale across Asia's diverse regulatory landscape — spanning East Asia, Southeast Asia, South Asia, and the Middle East — requires establishing institutional intent paired with automated execution mechanisms. Through the forthcoming Asia-ISAC AI Governance Framework, organisations across the region will be able to establish institutional intent, align internal policies, and pursue certification. Bridging that framework to live operations requires a Living Audit Trail — an automated, persistent record of inference-time behaviour that replaces episodic, manual validation with continuous proof of compliance.

[1]Stanford University Human-Centered AI Institute, AI Index Report 2025 (Stanford, CA: Stanford HAI, 2025). The Index recorded 233 discrete AI-related incidents in 2024, a 56.4% increase from 149 incidents in 2023, with the majority occurring in systems already deployed in production environments.

[2]McKinsey & Company, The State of AI 2025 (New York: McKinsey Global Institute, 2025). Of 88% of organisations reporting at least one AI deployment, only 6% qualified as high-performers achieving measurable bottom-line impact.

© 2026 Asia-ISAC & RAIDS AI LTD. A joint publication — all rights reserved.

About RAIDS AI

RAIDS AI is a real-time, black-box monitoring layer for AI in production that learns baseline behaviour from input/output data and flags deviations in near real time, enabling secure and audit-ready AI operations without access to model weights or training data.

About Asia-ISAC

Asia-ISAC (Asia Information Sharing and Analysis Center) is the first and only cross-industry, non-profit regional hub dedicated to helping companies unlock business growth through safe and secure AI adoption while reducing the cost and impact of cyberattacks across Asia.

The Structural Failure

The dominant paradigm of AI risk management remains oriented towards the moment of launch. Organisations invest substantially in pre-production governance artefacts — risk assessments, model cards, bias evaluations, legal sign-off — yet these instruments offer no ongoing assurance once a model is exposed to production traffic. Risk, in this framework, is treated as a static property of the system rather than a dynamic expression of its runtime behaviour. This structural failure has commercial and legal consequences. S&P Global's 2025 Enterprise AI Survey found that 42% of enterprises abandoned the majority of their AI initiatives — more than double the 17% recorded in 2024[3] — while Gartner projects that more than 40% of agentic AI projects will be cancelled by the end of 2027 due specifically to governance and return-on-investment failures.[4]

The liability exposure is equally material. IBM's 2025 Cost of a Data Breach Report found that 13% of organisations experienced breaches of AI models or applications; of those affected, 97% reported lacking proper AI access controls.[5] These figures suggest that the absence of runtime governance is not merely a performance issue but a systemic organisational risk.

---

[3]S&P Global Market Intelligence, Enterprise AI Adoption Survey 2025 (New York: S&P Global, 2025). The survey of over 1,000 enterprises found that 42% had abandoned the majority of AI initiatives in 2025, up from 17% in 2024.

[4]Gartner, Inc., Top Strategic Technology Trends for 2026 (Stamford, CT: Gartner, 2025). Gartner projected that more than 40% of agentic AI projects would be cancelled by end of 2027 due to governance and ROI failures.

[5]AccuKnox, AI Security and Governance Guide 2026 (San Jose, CA: AccuKnox, 2026). IBM's 2025 Cost of a Data Breach Report found 13% of organisations experienced breaches of AI models or applications; of those compromised, 97% reported lacking proper AI access controls.

© 2026 Asia-ISAC & RAIDS AI LTD. A joint publication — all rights reserved.

Inference-Time Intelligence

To scale AI sustainably, enterprises must evolve beyond episodic pre-deployment certification towards what this paper terms Inference-Time Intelligence — the technical and institutional capacity to observe, interpret, and respond to model behaviour in production in real time. This requires a fundamental architectural shift: monitoring must move from a periodic, compliance-oriented activity to a continuous, operationally integrated function embedded in the AI lifecycle from the outset.

The Shared Thesis

> ***Sustainable AI scaling in 2026 will be determined by an organisation's capacity for runtime oversight rather than the maturity of its pre-deployment documentation.***

The pages that follow operationalise this thesis through an integrated lifecycle model spanning structural governance design, real-time behavioural monitoring, the region's regulatory landscape, and the human dimensions of AI oversight — examining both the strengths that human judgement brings to AI governance and the well-documented failure modes it introduces. The forthcoming Asia-ISAC AI Governance Framework is intended to establish institutional intent and certification baselines, while technical monitoring platforms — with RAIDS AI examined here as an example — provide the continuous telemetry needed to verify that governance is being upheld in production.

© 2026 Asia-ISAC & RAIDS AI LTD. A joint publication — all rights reserved.

II. The AI Lifecycle Risk Transformation Model

AI risk is not uniformly distributed across the lifecycle. It concentrates at specific transition points — most critically, the boundary between controlled pre-production testing and live operation. The model presented here delineates four distinct phases, each with its own risk profile, governance requirements, and human capital dependencies.

Phase 1: Pre-Production — Governance Design

The pre-production phase establishes structural intent. This encompasses the design and documentation of the AI Management System (AIMS), the formalisation of risk appetite, the selection of training data, and the execution of pre-deployment evaluation protocols. Industry standards including the NIST AI Risk Management Framework[6] and ISO/IEC 42001:2023[7] provide the normative architecture for this phase, mandating documentation of intended use cases, known limitations, and the escalation pathways that will govern production incidents.

The governance design phase is a necessary but structurally insufficient condition for operational safety. From an Asia-ISAC perspective, design-time risk frameworks must be formally aligned with the forthcoming Asia-ISAC AI Governance Framework and other industry-led governance frameworks, which serve as an institutional statement of intent under which organisations across Asia can align internal policies and pursue certification. However, pre-production documentation merely establishes intent; the Living Audit Trail is what provides the operational proof in production.

The same caveat applies to testing as to documentation: passing an evaluation suite at launch is no more a guarantee of runtime safety than a governance document is. In practice, organisations frequently assign AI adversarial testing to their existing functional QA function, on the assumption that a testing discipline is a testing discipline. This creates two compounding problems. First, QA teams are trained to confirm that a system behaves as designed; they are not trained to behave as an adversary trying to make it misbehave, so the probing questions that surface jailbreaks, prompt injection, and policy circumvention are simply never asked. Second, adversarial AI testing is a distinct discipline from functional testing, not an extension of

[6]NIST, AI Risk Management Framework (AI RMF 1.0) (Gaithersburg, MD: National Institute of Standards and Technology, 2023). The framework mandates continuous monitoring as a core governance control, requiring documented escalation pathways that link telemetry alerts to named governance owners accountable for corrective action.

[7]ISO/IEC 42001:2023, Information Technology — Artificial Intelligence — Management System (Geneva: International Organization for Standardization, 2023). The standard provides requirements and guidance for establishing, implementing, maintaining, and continually improving an AI management system within an organisational context.

© 2026 Asia-ISAC & RAIDS AI LTD. A joint publication — all rights reserved.

it, and treating it as a new test suite bolted onto an existing team understates the retraining it requires. This pattern appears to be widespread rather than isolated: the same gap between functional QA and adversarial red-teaming has been raised independently by practitioners at recent AI engineering conferences, reinforcing that pre-production testing, like pre-production documentation, is necessary but not sufficient.

Transitioning to the Living Audit Trail

Pre-production documentation establishes an organisation's structural intent, but intent without continuous verification is indistinguishable from absence in production. Bridging this gap requires transitioning from static, point-in-time artefacts to a Living Audit Trail — an automated, persistent, and immutable log of inference-time behaviour. This trail converts real-time operational telemetry into continuous evidence, ensuring that governance is actively executed at runtime rather than merely promised on paper.

Phase 2: Live Runtime — Operational Behaviour (Emergent Risk)

The transition to live operation represents the most consequential and least-governed phase of the AI lifecycle. This is what the authors designate the Black Box Phase: models move from controlled, reproducible test environments into non-deterministic, live data environments whose statistical properties evolve continuously. Four primary categories of production-phase failure have been identified in the operational AI literature:

Data Drift refers to shifts in the statistical distribution of input features between the training population and the live inference population.[8] A credit risk model trained on pre-pandemic lending data, for example, may encounter a radically different applicant population in 2026, with the inputs it receives bearing limited distributional resemblance to its training set — yet without monitoring infrastructure, this divergence is invisible to the operating organisation.

Concept Drift is formally defined as a change in the joint probability distribution P(X, y) — the relationship between model inputs and the target variable — occurring after deployment.[9] The underlying rules of the world change; a fraud detection system that learned patterns of fraudulent behaviour in 2023 may find those patterns obsolete by 2025 as attack vectors evolve. Performance monitoring at the aggregate level routinely masks this phenomenon: a model may

[8]Evidently AI, ML in Production: Data Drift Detection (Moscow/Amsterdam: Evidently AI, 2024). Data drift — defined as statistically significant shifts in the distribution of input features between training and production environments — is among the primary causes of post-deployment model degradation.

[9]Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). 'A survey on concept drift adaptation.' ACM Computing Surveys, 46(4), 1–37. Concept drift is formally defined as a change in the joint probability distribution P(X, y) between two temporal observations, causing model predictions to diverge from actual outcomes.

© 2026 Asia-ISAC & RAIDS AI LTD. A joint publication — all rights reserved.

sustain 95% overall accuracy while specific decision segments degrade to 80% accuracy, compounding into material business impact before detection.[10]

Behavioural Drift encompasses emergent output instability not fully captured by the preceding categories, including hallucination frequency in large language models, unsafe output generation, and latency anomalies. A customer service LLM that functioned flawlessly during testing, for instance, may exhibit a progressively compounding rate of semantic hallucinations or toxic responses as it processes conversational loops with unscripted user prompts over weeks of live deployment. Critically, standard infrastructure monitoring (HTTP 200 responses, low latency) shows no anomaly in such scenarios, demonstrating that conventional software health monitoring is structurally blind to AI-specific failure modes.

Adversarial Drift describes the fourth face of runtime instability — not attrition, but attack. Prompt injection, poisoned retrieval data, tool misuse, and data exfiltration attempts each surface, from a monitoring layer's vantage point, as an unexpected change in input, output, or decision patterns — the same signature that data, concept, and behavioural drift leave behind. Framed this way, drift detection becomes the evidence trail an attack leaves rather than a separate discipline, and it maps directly onto the MITRE ATLAS taxonomy of adversarial machine learning techniques, which already catalogues these behaviours. This connection also closes a gap opened in Section I: a paper that opens with the growth of agentic AI should not let agents disappear from the analysis that follows. Agentic systems — with tool access, memory, and multi-step autonomy — are precisely the surface on which adversarial drift plays out, and inference-time monitoring is positioned to catch it.

This four-part taxonomy of drift underscores an essential regional imperative: as models encounter the complex linguistic, cultural, and operational nuances of Asia's diverse markets, continuous inference-time telemetry ensures they remain securely within the safety boundaries defined by an institutional governance framework.

Phase 3: Incident Response — Operational Risk Control (Materialised Risk)

When behavioural anomalies are detected, the speed and quality of the organisational response determine whether a production incident becomes a contained operational event or a material risk realisation. This phase requires integrating technical detection layers — provided by operational monitoring solutions such as RAIDS AI — with enterprise incident response procedures guided by the forthcoming Asia-ISAC AI Governance Framework. This separation ensures that technical observability informs executive accountability, without confusing engineering problem-solving with institutional oversight.

---

[10]Sama, Model Drift Explained (San Francisco, CA: Sama, 2025). Aggregate performance metrics can mask localised degradation; a model may sustain 95% overall accuracy while specific user segments or product categories drop to 80%, compounding into visible business impact before detection.

© 2026 Asia-ISAC & RAIDS AI LTD. A joint publication — all rights reserved.

Phase 4: Audit and Compliance — Assurance Layer (Retrospective Risk)

The final phase converts the operational record into an institutional asset. By transforming live telemetry into a Living Audit Trail, organisations generate the retrospective proof of ongoing compliance demanded by regulators, boards, and increasingly, enterprise customers. This is not merely a documentation exercise. The EU AI Act entered into force on 1 August 2024; its Article 50 transparency obligations apply from 2 August 2026, and its high-risk system obligations phase in through December 2027 and August 2028. Alongside the emerging framework obligations across Asia discussed in Section IV, continuous audit evidence is a legal requirement for certain risk classifications well ahead of those later deadlines, not an optional governance enhancement.[11]

[11]Sia Partners, The AI Regulatory Landscape in APAC (Singapore: Sia Partners, 2024). The EU AI Act entered into force on 1 August 2024, establishing the first comprehensive risk-based global AI legal framework. Obligations phase in progressively thereafter: prohibited-practice and general-purpose AI model provisions from August 2025, Article 50 transparency obligations from 2 August 2026, and high-risk system obligations phasing in through December 2027 and August 2028.

© 2026 Asia-ISAC & RAIDS AI LTD. A joint publication — all rights reserved.

III. Operationalising Governance: Runtime Monitoring and the RAIDS AI Capability Model

This section details the core operational capabilities required to execute runtime governance, using RAIDS AI as a worked example of the execution layer that a governance framework — such as the forthcoming Asia-ISAC AI Governance Framework, or any comparable industry-led framework — depends on to be enforceable in production.

Pillar 3.1: Inference-Time Monitoring — Closing the Visibility Gap

The central technical argument of this paper is that the majority of production AI failures are Silent Failures — anomalies that do not trigger conventional infrastructure alerts and therefore escape detection under monitoring frameworks designed for deterministic software systems.[12] Subtle hallucinations, unsafe outputs, gradual performance degradation, and policy violations may persist for weeks or months before their cumulative impact becomes visible in downstream business metrics. By this point, the remediation cost — financial, reputational, and regulatory — is substantially elevated relative to early detection.

RAIDS AI addresses this visibility gap as it runs alongside the model and observes its inputs and outputs in near real time, without requiring access to internal model weights, architecture, or training pipelines. This mechanism translates raw inference streams directly into the **Living Audit Trail**: a structured, tamper-evident, continuously updated record of production model behaviour that serves simultaneously as an operational risk instrument and a compliance artefact.[13]

This architecture produces evidence for ISO/IEC 42001:2023 Clause 9, which mandates that organisations deploying AI systems must continuously *monitor, measure, analyse and evaluate* their systems, retaining documented evidence of conformance.[14]

Pillar 3.2: Non-Invasive Visibility — The Black-Box Imperative

A practical constraint that no enterprise AI governance framework can ignore is the prevalence of third-party API integrations and embedded models in which internal instrumentation is contractually or technically restricted. An organisation deploying a hyperscaler's foundation model via API has no access to the model's internal activations, attention patterns, or intermediate representations. It can only observe inputs and outputs. Non-invasive black-box monitoring — which characterises model behaviour entirely from external inference records

[13]AI Governance Control Stack for Operational Stability, arXiv preprint arXiv:2604.03262 (2026). Continuous telemetry and model observability are described as 'core requirements for operational AI stability, enabling real-time monitoring of prediction drift, performance degradation, and data pipeline anomalies.'

© 2026 Asia-ISAC & RAIDS AI LTD. A joint publication — all rights reserved.

without requiring internal model access — is therefore not merely a technical preference but a structural requirement for comprehensive enterprise AI oversight.

The technical architecture of hardened AI runtime governance relies primarily on system-boundary telemetry and proactive validation layers to preserve operational stability.[15] Telemetry streams relevant to continuous risk observation include boundary inference behaviour, safety-trigger rates, semantic anomalies, policy deviations, and unexpected output distributions evaluated entirely at the interface layer. When anomalous behaviour appears at these boundaries, RAIDS flags it for the operator to act on, so that it can be addressed before it affects downstream business workflows or end users.

Pillar 3.3: Dynamic Telemetry — Adaptive Baselines vs. Static Thresholds

Traditional monitoring architectures rely on brittle, static alert thresholds — fixed numerical limits whose appropriateness decays as the operational context evolves. RAIDS AI's approach instead establishes adaptive safety baselines that learn from the model's own production history, adjusting dynamically as legitimate distributional shifts occur (seasonal patterns, product changes, user population evolution) while maintaining sensitivity to genuinely anomalous deviations.

This distinction matters significantly in practice. A static threshold for hallucination rate in a customer service LLM would require manual recalibration following every product launch, regulatory change, or seasonal demand shift. Adaptive baselines perform this recalibration continuously, reducing the operational burden on governance teams while increasing the signal-to-noise ratio of the monitoring system. Effective drift detection, the research literature notes, combines automated statistical monitoring with human evaluation; neither approach is sufficient in isolation.[16] Each adaptive baseline evaluation produces a structured data record that flows directly into the Living Audit Trail, creating a continuous, timestamped evidence chain that supports operational governance and produces evidence for ISO 42001 Clause 9 and the NIST AI RMF.[17]

Adaptive baselines are a significant improvement on static thresholds, but a baseline learned only from an agent's own live traffic has known limitations. A credible monitoring architecture therefore compares behaviour against more than one reference point and looks for signals that back each other up before it raises an alert.

None of this is achievable without a caveat that deserves equal emphasis: the telemetry this architecture depends on mostly is not being emitted yet in most enterprise environments. The statistical and architectural case for multi-baseline, multi-signal monitoring is well established;

---

[15]Ibid. The paper defines 'hardened governance' as a shift from reactive response — where governance mechanisms respond to failures after they occur — toward proactive control of system autonomy through gated propagation and continuous telemetry.

© 2026 Asia-ISAC & RAIDS AI LTD. A joint publication — all rights reserved.

the practical constraint is the logging and instrumentation pipeline required to feed it, which in the majority of production deployments does not yet exist at the fidelity this approach requires.

© 2026 Asia-ISAC & RAIDS AI LTD. A joint publication — all rights reserved.

IV. Asia-ISAC Perspective: High-Complexity Operating Environments

The Global Stress Test: Asia as the Regulatory Frontier

The Asia region constitutes one of the most structurally complex regulatory environments for AI deployment globally, spanning East Asia, Southeast Asia, South Asia, and the Middle East. Unlike the European Union's unified legislative framework under the EU AI Act — which entered into force on 1 August 2024 and imposes binding, risk-tiered obligations across all member states[18] — the Asian regulatory landscape is characterised by pronounced jurisdictional divergence, with national frameworks at materially different stages of legislative maturity.

East and Southeast Asia. South Korea enacted the first comprehensive AI regulation in the region. Singapore's IMDA operates through its Model AI Governance Framework for Generative AI and the AI Verify testing toolkit — which explicitly bridges NIST AI RMF and ISO/IEC 42001 compliance standards — with enforcement embedded in procurement standards and the MAS FEAT Principles for financial services, rather than mandatory statute.[19] Japan's AI Promotion Act adopts an 'agile governance' approach, treating risk management as an enabler of national competitiveness and allowing guidance to adapt as the technology evolves.[20]

A second wave of regional enforcement is now taking shape. Malaysia's AI Governance Bill reached Cabinet submission in June 2026; Indonesia is converting its National AI Strategy into a Presidential Regulation with three risk tiers; and Thailand is consolidating draft principles into unified law. All three are expected to reach binding status by 2027–2028.[21] Organisations operating across these jurisdictions face not a single compliance target but a continuously evolving matrix of overlapping and sometimes conflicting obligations.

---

[19]White & Case LLP / Morgan Lewis, The New Rules of AI: A Global Legal Overview (2025). Singapore's IMDA launched the Model AI Governance Framework for Generative AI, while the MAS FEAT Principles govern AI in financial services. Compliance is largely embedded in procurement standards and sector-specific guidance rather than mandatory statute.

[20]CyberCX, Governing AI Risk: Understanding the Emerging Regulatory Landscape (Melbourne: CyberCX, 2026). Japan's AI Promotion Act takes an 'agile governance' approach, framing AI risk management as an enabler of national competitiveness rather than a compliance burden, with oversight designed to adapt as the technology evolves.

[21]1stopasia, AI Under Law: Global Regulatory Intelligence Report Part 2 (Singapore, May 2026). Malaysia's AI Governance Bill reached Cabinet submission in June 2026; Indonesia is converting its National AI Strategy into a Presidential Regulation with three risk tiers; Thailand is consolidating draft principles into unified law. All three are expected to reach binding law status by 2027–2028.

© 2026 Asia-ISAC & RAIDS AI LTD. A joint publication — all rights reserved.

South Asia. India's evolving governance framework under NITI Aayog guidelines and the Digital Personal Data Protection (DPDP) Act establishes strict mandates for automated processing that affects citizens. Sri Lanka, Bangladesh, and Pakistan are at earlier stages of formal AI policy development, presenting both governance gaps and significant opportunities for framework adoption.[22]

Middle East. The UAE and Saudi Arabia (via SDAIA) have implemented national AI strategies, ethics guidelines, and regulatory sandboxes that combine rapid deployment ambitions with localised data residency and compliance mandates. Qatar, Bahrain, and Jordan are developing complementary frameworks aligned with GCC-wide digital economy priorities. The Middle East presents a distinctive governance context: high AI investment ambition, rapidly maturing regulatory capacity, and cross-border deployment complexity across Gulf Cooperation Council jurisdictions.[23]

The Observability Tension: Sovereignty Versus Visibility

The technical requirements of comprehensive AI monitoring — particularly the continuous collection and analysis of inference telemetry — exist in direct tension with the data residency and sovereignty laws that are proliferating across Asia. Localisation requirements mandating that data remain within national borders create architectural constraints on centralised monitoring infrastructure. An organisation operating AI systems across Singapore, India, and the UAE may be technically prohibited from routing all telemetry to a single monitoring platform, regardless of the platform's governance capabilities.

Resolving this tension requires governance architecture that is both technically distributed and institutionally coherent — capable of operating within local data residency constraints at the infrastructure layer while maintaining unified governance visibility at the organisational layer. The ASEAN Guide to AI Governance and Ethics (February 2024) provides a high-level reference framework for member states[24] but operational resolution requires jurisdiction-specific implementation.

Enterprise Case Study 1 (Illustrative Scenario): Shadow AI Remediation in Asian Enterprises

Shadow AI — the unauthorised adoption of AI tools by employees and business units outside formal IT and governance channels — has emerged as the most consistently cited near-term risk across Asian enterprises. Gartner's 2025 research found that over 60% of surveyed organisations reported undocumented AI tool usage by employees, with adoption accelerating

[24]KPMG Law, EU AI Act and the ASEAN AI Guide (Singapore: KPMG, 2024). The ASEAN Guide to AI Governance and Ethics (February 2024) sets out high-level principles for member states, while binding enforcement regimes vary substantially by jurisdiction.

© 2026 Asia-ISAC & RAIDS AI LTD. A joint publication — all rights reserved.

ahead of any governance response.[25] Shadow AI systems operate entirely outside the monitoring and accountability frameworks that sanctioned deployments are subject to, creating uncontrolled exposure to data leakage, hallucination risk, and regulatory non-compliance.

Consider an enterprise in financial services or critical infrastructure that adopts a Shadow AI remediation protocol combining network-layer discovery — identifying unauthorised AI endpoints — with the rapid onboarding of discovered models into a runtime monitoring boundary such as RAIDS AI. This approach converts a governance liability into a monitored asset: rather than prohibiting Shadow AI usage and driving it further underground, the protocol brings it within the **Living Audit Trail** perimeter, creating immediate visibility and a documented compliance posture. The goal is a remediation cycle — from discovery to monitored status — substantially shorter than the industry average of more than three months for formal AI procurement cycles.

Enterprise Case Study 2: Regional FinTech — Continuous Compliance Under MAS FEAT & Regional Rules

FinTech operators across the region deploy AI in high-stakes, high-velocity decision contexts — credit decisioning, fraud detection, and real-time payments risk — where model failures generate immediate, measurable harm, and where the regulatory expectations of the Monetary Authority of Singapore's FEAT Principles require that AI systems remain continuously fair, explainable, and auditable throughout their operational lifecycle.[26]

A recurring pattern documented across Asia's FinTech sector involves fraud detection models trained on historical transaction data that encounter materially shifted fraud patterns as regional payment behaviours evolve. BIS research found that false-negative rates in Southeast Asian fraud detection models climbed 34% between training and live deployment, with the degradation undetected for an average of 11 weeks — a period during which fraudulent transactions were being systematically misclassified.[27] Under RAIDS AI's inference-time monitoring, boundary

[25]Gartner, Inc., Market Guide for AI Trust, Risk and Security Management (Stamford, CT: Gartner, 2025): over 60% of organisations reported undocumented employee AI tool usage. Also: Asian Development Bank (ADB), Digital Trade and Logistics Innovation in Southeast Asia (Manila: ADB, 2025). Regional logistics operators encountered compound drift events during the 2024 monsoon and post-pandemic freight recovery periods, where simultaneous environmental, congestion, and pricing shifts caused model recommendations to diverge materially from optimal outcomes.

[26]MAS (Monetary Authority of Singapore), Fairness, Ethics, Accountability and Transparency (FEAT) Principles (Singapore: MAS, 2022, updated 2024). The FEAT Principles require financial institutions to ensure AI systems used in customer-facing and credit decisioning processes remain continuously fair, explainable, and auditable throughout their operational lifecycle, not merely at the point of deployment approval.

[27]Bank for International Settlements (BIS), FSI Insights on Policy Implementation: Humans Keeping Control of AI in Financial Services (Basel: BIS, 2025). The paper documents a pattern of FinTech AI

© 2026 Asia-ISAC & RAIDS AI LTD. A joint publication — all rights reserved.

telemetry is designed to flag the divergence in safety-trigger rates, triggering enterprise escalation procedures guided by the forthcoming Asia-ISAC AI Governance Framework and initiating a retraining or override procedure before regulatory thresholds are breached.

Critically, every detection event, escalation trigger, and remediation action is automatically captured in the Living Audit Trail — providing MAS-facing audit documentation that demonstrates continuous FEAT compliance without requiring manual evidence assembly at examination time. This transforms regulatory audit preparation from a labour-intensive, retrospective exercise into a real-time operational output.

Enterprise Case Study 3 (Illustrative Scenario): Regional Logistics — Compound Drift in High-Velocity Asian Supply Chains

Regional logistics operators deploying AI route optimisation and demand forecasting encounter a distinctive risk pattern: compound drift events, in which multiple simultaneous distributional shifts — weather, port congestion, fuel pricing, seasonal demand — interact to produce recommendation failures that no single-variable monitoring approach would detect. ADB research documents cases where post-pandemic freight recovery and monsoon season combined to render AI routing recommendations materially suboptimal, with models unable to recognise that their training environment no longer corresponded to operating reality.[28]

In this scenario, a logistics enterprise addresses this through multi-signal boundary telemetry — monitoring the joint distribution of inference inputs and outputs across the relevant environmental variables simultaneously, rather than tracking each variable independently. This approach, as implemented in RAIDS AI's adaptive baseline framework, is designed to detect compound drift events that single-metric monitoring would miss, and to produce the structured Living Audit Trail records required for contractual SLA compliance documentation with enterprise logistics clients.

degradation in Southeast Asian markets where models trained on pre-2023 transaction data encountered materially shifted fraud patterns by 2025, producing false-negative rates that climbed 34% before detection, with an average undetected window of 11 weeks.

© 2026 Asia-ISAC & RAIDS AI LTD. A joint publication — all rights reserved.

V. The Human Element: Strengths, Weaknesses, and the Governance Imperative

Any technically rigorous treatment of AI governance that neglects the human dimension is structurally incomplete. AI systems do not operate in isolation: they are adopted, configured, monitored, overridden, and ultimately sanctioned — or abandoned — by people operating within organisations. The empirical record on human performance in AI-adjacent roles is complex, characterised by genuine strengths that technical systems cannot replicate and well-documented failure modes that, if unaddressed, can undermine the governance structures this paper advocates.

Human Strengths in AI Adoption and Operation

Contextual and Ethical Judgement

The primary irreplaceable contribution of human actors in AI governance is contextual and ethical judgement. AI systems optimise for specified objectives within learned distributional boundaries; they are structurally unable to recognise when a situation falls outside the scope of their competence, when a technically correct output is ethically inappropriate, or when organisational context renders a model's recommendation inadvisable. Human reviewers, equipped with domain expertise and institutional knowledge, provide precisely this out-of-distribution sensitivity.[29]

This is particularly consequential in the incident response phase of the lifecycle model presented in Section II. When RAIDS AI telemetry flags a behavioural anomaly, the determination of appropriate institutional response — whether to suspend a model, notify affected stakeholders, trigger a regulatory disclosure, or implement a temporary override — requires human judgement operating within the governance protocol framework. No monitoring system can make these determinations autonomously.

Adaptive Expertise and Novel Threat Recognition

Human experts bring adaptive expertise to AI oversight that is qualitatively different from the pattern-matching capabilities of monitoring systems. When a failure mode is genuinely novel — when a model begins behaving in ways that fall outside the distributional assumptions of the monitoring baseline — human experts with deep domain knowledge are frequently the first to recognise that something is wrong, even before statistical monitoring flags a threshold breach. This 'weak signal' detection capability is a meaningful complement to automated monitoring infrastructure.[30]

---

[30]UXmatters, 'UX Research Insights: Balancing AI Automation and Human Oversight in IT Operations' (December 2025). Calibrated trust — balancing automation with human agency — requires

© 2026 Asia-ISAC & RAIDS AI LTD. A joint publication — all rights reserved.

Organisational Leadership and Change Agency

Executive leadership is the strongest organisational predictor of AI governance effectiveness. McKinsey's 2025 research found that CEO direct ownership of AI governance — as distinct from mere sponsorship — is the single management practice most correlated with bottom-line AI impact.[31] Yet only 28% of organisations report direct CEO involvement in AI governance, suggesting that the majority are leaving significant organisational capability unrealised. The Wharton Human-AI Research Institute similarly identifies talent, training, and trusted guardrails as the human capital factors most directly affecting the speed and efficacy of AI deployment.[32]

Human Weaknesses in AI Adoption and Operation

The Change Management Deficit

The most frequently cited cause of enterprise AI implementation failure is not technical — it is human. Analysis of AI project post-mortems consistently attributes 63% of implementation failures to human factors: resistance to workflow change, inadequate training, misaligned incentive structures, and organisational inertia.[33] Only approximately one-third of companies in late 2024 prioritised change management and training as part of their AI rollouts[34] — a figure that suggests widespread institutional underestimation of the organisational effort required for successful deployment. Enterprise organisations averaged nine months or longer to scale AI

interface-level design interventions including explainability features, anomaly visibility, and structured override mechanisms to prevent automation-induced complacency.

[31]McKinsey & Company, The State of AI 2025, op. cit. CEO direct ownership of AI governance — as distinct from sponsorship — was identified as the single management practice most correlated with bottom-line AI impact. Only 28% of organisations reported direct CEO involvement in AI governance.

[32]Wharton Human-AI Research & GBK Collective, Accountable Acceleration: Gen AI Fast-Tracks into the Enterprise, Year Three Full Report (Philadelphia: Wharton School, University of Pennsylvania, October 2025). The report identifies talent, training, and trusted guardrails as the human capital factors most directly affecting the speed and efficacy of generative AI deployment.

[33]Bosio Digital, Why AI Projects Fail: The 63% Human Factor Problem (2026). Analysis of AI project post-mortems attributes 63% of implementation failures to human factors rather than technical deficiencies, with resistance to workflow change, inadequate training, and misaligned incentive structures as the primary sub-causes.

[34]Stack AI, The 7 Biggest AI Adoption Challenges for 2026 (2025). Only approximately one-third of companies in late 2024 prioritised change management and training as part of their AI rollouts, suggesting widespread underestimation of the organisational effort required for successful deployment.

© 2026 Asia-ISAC & RAIDS AI LTD. A joint publication — all rights reserved.

initiatives compared to 90 days for mid-market organisations, with structural inertia identified as the principal differentiator.[35]

This deficit is not a peripheral concern for AI governance teams; it is central to the operational integrity of the monitoring and response systems this paper describes. A sophisticated telemetry infrastructure whose alerts are systematically deprioritised by change-resistant operational teams provides no governance value. Runtime governance is only as effective as the human response it triggers.

Automation Bias: The Danger of Excessive Deference

Among the most extensively documented failure modes in human-AI interaction research is automation bias — the tendency of human decision-makers to assume that machine-generated outputs are correct, to pay insufficient attention to contradictory information, and to delegate cognitive effort to the AI system beyond the boundaries of its actual competence.[36] Bainbridge's seminal 1983 study of industrial automation, whose conclusions have proven persistently applicable to AI systems, demonstrated that as automation increases, the demand for skilled human judgement does not diminish — it intensifies — while the human's practical ability to exercise that judgement degrades through disuse.[37]

The consequences of automation bias in AI oversight contexts are severe. Skitka et al. demonstrated that operators using 'very reliable but not perfectly reliable' automated aids performed worse than those without automation, committing commission errors — that is, following incorrect automation recommendations — at rates approaching 100%.[38] In the context of AI governance, an operator experiencing automation bias who defers to a flawed model

[35]Chronus, Why AI Adoption Fails: Enterprise Barriers AI Leaders Ignore (2026), citing MIT Sloan Management Review data. Enterprise organisations averaged nine months or longer to scale AI initiatives, compared to 90 days for mid-market organisations, with structural inertia and change management deficits as the principal differentiators.

[36]Cummings, M.L. (2017). 'Automation Bias in Intelligent Time Critical Decision Support Systems.' AIAA 1st Intelligent Systems Technical Conference. Automation bias is defined as the tendency of human decision-makers to assume that machine-generated solutions are correct and to pay insufficient attention to contradictory information, with risk increasing as perceived AI competence rises relative to the operator's own self-assessment.

[37]Bainbridge, L. (1983). 'Ironies of Automation.' Automatica, 19(6), 775–779. Bainbridge's seminal work demonstrated that as automation increases, the demand for skilled human judgement does not diminish but intensifies, and that systems designed to eliminate human error often introduced designer errors while degrading operator skill through disuse.

[38]Skitka, L.J., Mosier, K.L., & Burdick, M. (1999). 'Does automation bias decision-making?' International Journal of Human-Computer Studies, 51(5), 991–1006. Operators using 'very reliable but not perfectly reliable' automated aids performed worse than those without automation, committing omission errors at 55% rates and commission errors (following incorrect automation) approaching 100%.

© 2026 Asia-ISAC & RAIDS AI LTD. A joint publication — all rights reserved.

output not only fails to prevent the error but may actively validate it within the organisational record, amplifying rather than mitigating the AI's failure.[39]

Algorithm Aversion: The Danger of Insufficient Deference

At the opposite end of the trust calibration spectrum, algorithm aversion describes the tendency of humans to distrust AI recommendations even when evidence supports their accuracy — particularly following early errors. This failure mode is most prevalent during AI adoption phases, when operators lack sufficient experience with a system's performance envelope to calibrate their trust appropriately.[40] In organisations where AI adoption is proceeding against cultural resistance, algorithm aversion can manifest as systematic override of correct model outputs, eroding the operational value of the AI investment and, in governance contexts, degrading the reliability of the monitoring signal.

Oversight Fatigue and Skill Erosion

Sustained attention during AI monitoring — interpreting complex model outputs, evaluating drift signals, and maintaining vigilance during extended periods of nominal system behaviour — places significant cognitive demands on human operators. The research literature identifies oversight fatigue as a structural vulnerability in human-in-the-loop AI governance: the demands of interpreting complex AI outputs and maintaining attention during prolonged monitoring produce cognitive degradation that increases error rates and reduces the reliability of human oversight precisely in the scenarios — low-frequency, high-consequence anomalies — where it matters most.[41]

A related dynamic is skill erosion, in which the automation of tasks previously performed by human operators gradually degrades the expertise required to perform those tasks manually or to supervise automated performance meaningfully. Operators who have never executed a credit decision without AI assistance may lack the domain knowledge to evaluate whether an AI-flagged anomaly represents a genuine risk or a false positive. This creates a governance dependency on the AI system's own reliability — a structural fragility in any risk oversight architecture.

---

[40]Green, B. (2022). 'The Flaws of Policies Requiring Human Oversight of Government Algorithms.' Computer Law & Security Review, 45. A survey of human oversight policies concludes that humans are frequently unable to perform assigned oversight functions effectively, exhibiting both automation bias and algorithm aversion, and faring poorly at judging the accuracy of algorithmic predictions.

[41]PMC / ScienceDirect, 'Is Human Oversight of AI Systems Still Possible?' (2024). The demands of interpreting complex AI outputs and maintaining attention during prolonged monitoring produce oversight fatigue; this is compounded by a shortage of professionals with the interdisciplinary knowledge necessary to oversee complex AI systems effectively.

© 2026 Asia-ISAC & RAIDS AI LTD. A joint publication — all rights reserved.

Designing for Human Performance: Governance Implications

The strengths and weaknesses outlined above have direct implications for the design of governance systems. Effective runtime AI governance does not simply insert humans into monitoring workflows; it designs those workflows to amplify human strengths while structurally mitigating the documented failure modes.

Key design principles emerging from the human factors literature include: interface designs that make AI behaviour transparent and anomalies visible without requiring expert interpretation (supporting calibrated trust rather than defaulting to automation bias or aversion);[42] structured override mechanisms that require active human confirmation of AI recommendations in high-stakes contexts, preventing passive deference;[43] rotation of monitoring responsibilities to mitigate oversight fatigue; and continuous training programmes that maintain human expertise alongside AI capability growth.[44] The EU AI Act and UK AI Assurance Roadmap both cite automation bias explicitly as a policy driver for human oversight mandates[45] — reflecting regulatory recognition that the human component of AI governance requires as much deliberate engineering as the technical component.

---

[42]UXmatters, 'UX Research Insights: Balancing AI Automation and Human Oversight in IT Operations' (December 2025). Calibrated trust — balancing automation with human agency — requires interface-level design interventions including explainability features, anomaly visibility, and structured override mechanisms to prevent automation-induced complacency.

[43]NIST, AI Risk Management Framework (AI RMF 1.0) (Gaithersburg, MD: National Institute of Standards and Technology, 2023). The framework mandates continuous monitoring as a core governance control, requiring documented escalation pathways that link telemetry alerts to named governance owners accountable for corrective action.

[44]Wharton Human-AI Research & GBK Collective, Accountable Acceleration: Gen AI Fast-Tracks into the Enterprise, Year Three Full Report (Philadelphia: Wharton School, University of Pennsylvania, October 2025). The report identifies talent, training, and trusted guardrails as the human capital factors most directly affecting the speed and efficacy of generative AI deployment.

[45]Gartner, Inc., Market Guide for AI Trust, Risk and Security Management (Stamford, CT: Gartner, 2025): over 60% of organisations reported undocumented employee AI tool usage. Also: Asian Development Bank (ADB), Digital Trade and Logistics Innovation in Southeast Asia (Manila: ADB, 2025). Regional logistics operators encountered compound drift events during the 2024 monsoon and post-pandemic freight recovery periods, where simultaneous environmental, congestion, and pricing shifts caused model recommendations to diverge materially from optimal outcomes.

© 2026 Asia-ISAC & RAIDS AI LTD. A joint publication — all rights reserved.

VI. Conclusion: The Infrastructure for Industrial-Grade AI

The Execution Engine

The argument developed across this paper resolves to a clear operational thesis: the forthcoming Asia-ISAC AI Governance Framework establishes the statement of institutional intent, standards, and certification pathways for enterprise AI deployment across Asia, while operational runtime monitoring — with RAIDS AI examined here as a working example — serves as the Execution Engine that makes such a framework enforceable, operational, and verifiable in production. Governance frameworks are statements of institutional intent; monitoring infrastructure is the mechanism by which that intent is operationalised in real time. Without real-time visibility into production model behaviour, governance protocols cannot be triggered by the events they are designed to address.

Institutional Readiness: From Safe at Launch to Safe in Operation

The transition from 'Safe at Launch' to 'Safe in Operation' is the fundamental prerequisite for enterprises seeking to scale AI beyond the pilot phase. This transition is both technical and institutional. Technically, it requires the deployment of inference-time monitoring infrastructure capable of detecting silent failures before they manifest as material harm. Institutionally, it requires the governance maturity to respond to detected anomalies with appropriate speed and calibration — a capability that depends critically on the human factors examined in Section V.

The empirical evidence reviewed in this paper underscores the urgency of this transition. With AI incidents rising 56% year-on-year[46] and over 40% of AI initiatives failing to deliver expected outcomes,[47] the cost of inadequate runtime governance is no longer abstract. It is materialising in tribunal rulings, regulatory fines exceeding €60 million,[48] abandoned initiatives, and the systematic failure of enterprise AI investments to deliver value proportionate to their cost.

The Blueprint for Scale

Industrial-grade AI — AI that operates reliably, accountably, and at scale across the complexity of real enterprise environments across Asia — requires the integration of three capabilities that this paper has examined in turn: structural governance design (Section II), continuous behavioural monitoring (Section III), and deliberate human factors engineering (Section V), all situated within the multi-jurisdictional regulatory landscape of the region (Section IV).

None of these capabilities is sufficient in isolation. Governance frameworks without monitoring are aspirational documents. Monitoring without governance protocols are unresolved alert queues. Both without human systems designed for the cognitive realities of AI oversight are architectures that will fail at the moments that matter most. Asia-ISAC invites member organisations, technology partners, and regulators across the region to engage with the

© 2026 Asia-ISAC & RAIDS AI LTD. A joint publication — all rights reserved.

forthcoming Asia-ISAC AI Governance Framework as the shared institutional architecture for that integrated system; RAIDS AI's experience as an early operational partner demonstrates that such a framework can already be operationalised. The organisations that will scale AI successfully in 2026 and beyond are those that treat these three dimensions as an integrated system, within a framework that remains open, independent, and built to grow with the ecosystem it serves.

---

[46]Stanford University Human-Centered AI Institute, AI Index Report 2025 (Stanford, CA: Stanford HAI, 2025). The Index recorded 233 discrete AI-related incidents in 2024, a 56.4% increase from 149 incidents in 2023, with the majority occurring in systems already deployed in production environments.

[47]S&P Global Market Intelligence, Enterprise AI Adoption Survey 2025 (New York: S&P Global, 2025). The survey of over 1,000 enterprises found that 42% had abandoned the majority of AI initiatives in 2025, up from 17% in 2024.

[48]Stanford University Human-Centered AI Institute, AI Index Report 2025 (Stanford, CA: Stanford HAI, 2025). The Index recorded 233 discrete AI-related incidents in 2024, a 56.4% increase from 149 incidents in 2023, with the majority occurring in systems already deployed in production environments.

© 2026 Asia-ISAC & RAIDS AI LTD. A joint publication — all rights reserved.