Non-Agentic AI Govenance Singapore

GOVERNING GENERATIVE AI

From Verifiable Systems
to the Drift Problem

A Governance Framework for Enterprise
AI Deployment, Agentic Risk,
and Human Oversight

Edwin Koh Wui Kiat

NON-AGENTIC AI GOVERNANCE SINGAPORE

WISL™ No. 53 · Non-Agentic AI 2.0™ · 2026

GOVERNING GENERATIVE AI

From Verifiable Systems to the Drift Problem

A Governance Framework for Enterprise AI Deployment, Agentic Risk, and Human Oversight

---

Edwin Koh Wui Kiat · Tiger · P-LIFE 1.00™

Non-Agentic AI Governance Singapore · ACRA T260229801

kohedwin.ai · non-agentic.ai · 2026

謙虛·沉默·尊严·仁

Humility · Silence · Dignity · Benevolence

SYNOPSIS

Governing Generative AI: From Verifiable Systems to the Drift Problem is a practitioner's governance framework for enterprise AI deployment. It integrates two convergent bodies of analysis: the technical architecture required to move AI systems from closed-book hallucination risk to open-book verifiable performance; and the governance architecture required to prevent the silent erosion of human authority as these systems become increasingly autonomous.

The publication proceeds in five parts. Part I establishes the strategic imperative — why static LLMs are insufficient for high-stakes enterprise deployment and what verifiable architectures require. Part II covers retrieval, reasoning, and agentic workflow design. Part III presents the taxonomy of systemic risks, including the central challenge of agentic drift. Part IV compares present-day software-layer mitigation with structural prevention approaches. Part V addresses enterprise governance alignment, environmental sustainability, and intellectual property obligations.

The publication is grounded in the following source materials:

- Enterprise AI Governance Protocol —CET946, NTU, 2026
- Cai, C. J., et al. (2023) — The Lancet Digital Health
- EU AI Act, Singapore Model AI Governance Framework, WHO Guidelines

This publication is a Non-Agentic AI Governance Singapore Publication. It is available at kohedwin.ai.

PART I — THE STRATEGIC IMPERATIVE

1.1 The Limits of Static LLMs

The enterprise AI landscape has reached a critical inflection point. Traditional Large Language Models operate as probability engines in a closed-book environment, relying solely on internal parameterised memory. This architecture is fundamentally insufficient for high-stakes enterprise deployment. It yields opaque outputs prone to hallucination, lacks source traceability, and operates on knowledge that is outdated the moment training concludes.

Strategic deployment now requires the transition to open-book verifiable systems. This shift is not merely technical. It is a governance imperative: decisions grounded in external, verifiable, cited data are auditable; decisions generated from internal parametric memory are not.

Dimension	Closed-Book (Static LLM)	Open-Book (Verifiable System)
Data Currency	Fixed at training cutoff	Dynamically updated via external sources
Source Traceability	None — knowledge is opaque	High — generates citations from source documents
Hallucination Risk	High — invents facts under uncertainty	Low — grounded in retrieved, verified content
Governance Auditability	Not auditable	Auditable to source document level

1.2 Temperature as Legal Liability

From a governance perspective, parameters such as Temperature are not merely technical settings — they are levers of legal liability. A high Temperature in a legal-assistant RAG system, for instance, could lead to hallucinated precedents, resulting in professional negligence and systemic inconsistency.

Calibrating these engines is the first step toward securing the enterprise risk surface. Governance must mandate appropriate temperature ranges for deployment contexts, particularly in regulated industries.

PART II — ARCHITECTURE: RETRIEVAL, REASONING, AND AGENTS

2.1 Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation serves as the primary mechanism for transforming LLMs from simple text predictors into interpretable business tools. The standard RAG pipeline follows three stages:

- **Indexing and Chunking:** Enterprise data is split into semantic chunks, converted into vector embeddings, and stored in specialised databases.
- **Retrieval:** Dense Retrieval (e.g., DPR) matches semantic intent through neural networks, outperforming Sparse Retrieval (e.g., BM25) which relies on keyword frequency alone.
- **Generation:** The system synthesises retrieved chunks with the original query to produce a grounded, cited response.

Advanced optimisations include pre-retrieval query expansion (ensuring the engine interprets complex queries correctly) and post-retrieval re-ranking (filtering the most relevant information to prevent information overload at the generation stage). For static, high-frequency internal reference sets, Cache-Augmented Generation (CAG) provides tighter factual grounding than standard RAG.

2.2 Structured Reasoning

Retrieval alone is insufficient for complex problem-solving. To handle multi-step STEM, coding, and logical tasks, LLMs must operate in a thinking mode characterised by increased inference-time compute. Three key methodologies govern this:

- **Chain-of-Thought (CoT):** The model explicitly states step-by-step logic, providing an auditable trail of internal processing and improving explainability.
- **Search and Select (Self-Consistency):** Multiple diverse reasoning paths are generated and a majority vote is taken across the results. The Tree of Thoughts framework branches into multiple potential steps and uses reward models to prune unsuccessful paths.
- **Reflect and Refine:** The model observes its own mistakes through internal heuristics or external environments (such as a code executor) and refines subsequent attempts iteratively.

Process-Supervised Reward Models (PRMs) verify the logic at each individual step of the reasoning process, ensuring integrity throughout. However, this higher compute consumption increases operational costs and requires strict cost-benefit analysis before deployment.

2.3 Agentic Workflow Design

The next evolution in enterprise AI is the shift from passive chatbots to autonomous executors. AI agents are entities capable of sensing their environment, planning complex workflows, and executing actions to fulfil high-level goals. Every agentic framework must comprise four core components:

Component	Function	Governance Implication
Controller	The LLM brain — central planner and decision-maker	Reasoning chain must be auditable
Memory	Short and long-term storage for observations and past actions	Data retention and access controls required
Tool Set	APIs, search engines, external models	Each tool expands the risk surface
Environment	Contextual surroundings providing feedback after action	Production environments carry vastly higher risk than sandboxes

The ReAct (Reason + Act) framework governs agent operation: the agent generates a reasoning trace, executes an action in the environment, and uses the resulting observation to inform its next step. The environment is where enterprise risk manifests. An agent operating within a production database possesses a vastly higher risk profile than one in a restricted sandbox.

PART III — THE THREAT LANDSCAPE

3.1 Taxonomy of Systemic Risks

AI failures are often statistically plausible yet factually disastrous. A rigorous taxonomy of risks is essential for strategic risk management.

Risk Category	Description	Mitigation Priority
Hallucination	Model invents facts when uncertain; opaque and untraceable	High — addressed by RAG/CAG
Algorithmic Bias	Systematic errors from unrepresentative training data; demographic and cultural stereotyping	High — continuous auditing
Disinformation / Deepfakes	Hyper-realistic synthetic media erodes trust in authentic content (Liar's Dividend)	High — content provenance tools
Prompt Injection	Deceptive user instructions manipulate model logic	Critical — system-level guardrails
Data Poisoning	Corruption of training datasets to introduce backdoors	Critical — supply chain integrity
Shadow AI	Unauthorised employee use of generative tools; IP leakage risk	High — policy and monitoring
Slopsquatting	LLM invents fictitious code libraries; malicious actors weaponise the hallucinated names	Critical — code review and validation
Agentic Drift	Silent, gradual misalignment of agent from its objectives	Critical — see Part IV
Intellectual Complacency	Over-reliance on AI reduces human critical thinking and spatial memory	Medium — workforce development

3.2 Agentic Drift: The Silent Danger

Agentic drift is the governance challenge that distinguishes autonomous AI from all preceding software. Unlike traditional systems, agents do not crash when they fail. They degrade quietly — continuing to operate while providing subtly incorrect or non-compliant outputs, with no visible crash, no error log, no alarm.

Two definitions help characterise the problem from different governance perspectives:

Framework	Definition of Agentic Drift
Industry / Technical	The gradual misalignment of an AI agent’s behaviour from its intended objectives due to model updates, data distribution shifts, or context changes.
Constitutional / Sovereignty	The gradual and silent transfer of decision-making authority from human operators to an AI system, resulting in the loss of meaningful human control over consequential actions.

The industry definition frames drift as a performance problem. The constitutional definition frames it as a sovereignty problem. Both are accurate. Together they explain why drift is both technically dangerous and ethically consequential.

Statistical evidence confirms the severity. A 2023 study in The Lancet Digital Health found automation bias in 73% of clinical AI deployment studies reviewed, with human reviewers overriding AI recommendations in fewer than 12% of cases even when given the option. In healthcare, banking, and any regulated domain, this represents a fundamental accountability gap: the AI is not legally responsible for decisions, yet the human is not genuinely making them.

3.3 The Readiness Gap

83% of organisations plan to deploy agentic AI systems. Only 29% consider themselves operationally ready. This gap is not a technology problem. It is a governance problem. Organisations are deploying agents faster than they are building the governance structures to manage them.

PART IV — RESPONDING TO DRIFT: PRESENT VS FUTURE

4.1 The Present Response: Software-Layer Mitigation

The current industry standard for managing agentic drift is a layered set of operational practices. The enterprise prescribes the following mandatory controls:

- **Continuous Monitoring:** Real-time auditing of every step in the ReAct loop.
- **Behavioural Baselines:** Establishing normal operational boundaries to detect subtle deviations from intended behaviour.
- **Policy-as-Code Guardrails:** Limiting agents to permitted action sets.
- **Regression Testing:** Full regression testing of tools and models before deployment to prevent drift recurrence.
- **Version Control:** Rollback to known-good states when drift is detected.
- **Autonomy Tier Classification:** Matching the level of oversight to the risk profile of each agent.

These controls are necessary. They are also insufficient on their own for the highest-risk deployment environments. Every control listed depends on consistent human action: humans must act on monitoring alerts, humans must decide to initiate rollbacks, humans must enforce the baselines they established. In environments where alert fatigue and workflow pressure are constant — healthcare, banking, critical infrastructure — this dependency is the structural gap.

4.2 The Structural Gap

The structural gap is this: software-layer governance responds to drift. It does not prevent it. The guardrails can be softened under operational pressure. The baselines require human judgement to interpret. The rollback requires human initiative to trigger.

In the interval between when drift begins and when it is detected and acted upon, consequential decisions are being made — by a system that has already departed from its authorised mandate. In life-critical environments, that interval is the danger zone.

4.3 The Structural Prevention Approach

A structurally different approach addresses drift at the architectural level, before deployment, rather than at the operational level, after it occurs. Three principles define this approach:

Principle	Description	Mechanism
Offer-Only Logic	The AI generates recommendations but cannot self-execute. Every output requires explicit human authorisation before it propagates to action.	Architectural constraint on output type
Hardware-Enforced Deliberation	A mandatory timing gate inserted between AI output and human action. The gate does not	FPGA-based timing circuit with OTP fuse

Principle	Description	Mechanism
	evaluate. It waits. No software override exists.	
Constitutional Failure Mode	System defaults to inaction on failure, not degraded operation. A broken gate stops the system rather than permitting drift to continue.	Inaction as the safe state

The key distinction from software-layer mitigation: structural prevention does not depend on human consistency under pressure. The hardware gate does not tire. The inaction default does not negotiate. The audit trail does not require human initiative to maintain.

4.4 Comparative Summary

Dimension	Software-Layer Response	Structural Prevention
When it acts	After drift is detected	Before autonomous action is possible
Human dependency	High — requires consistent human response	Low — enforced regardless of human action
Bypass risk	Guardrails can be overridden under pressure	Hardware gate — no software bypass path
Failure mode	Degraded operation continues	System halts — inaction is the safe state
Accountability	Post-hoc audit trails	Immutable real-time ledger
Drift response	Detect, characterise, remediate	Certain categories architecturally prevented

PART V — ENTERPRISE GOVERNANCE FRAMEWORK

5.1 Global Regulatory Alignment

Enterprise AI governance must treat international frameworks not as suggestions but as operational constraints that dictate market access and regulatory standing. Three frameworks are primary:

- **EU AI Act:** Classifies AI systems used in high-risk contexts (healthcare, critical infrastructure, law enforcement) as high-risk, requiring human oversight, transparency, and conformity assessment.
- **Singapore Model AI Governance Framework (IMDA MGF):** Focuses on incident reporting, human-in-the-loop requirements, and content provenance tools including digital watermarks.
- **WHO Guidelines on AI for Health:** Emphasises transparency, accountability, and human oversight in clinical AI applications.

Because international frameworks currently focus on policy and ex-post compliance, organisations must supplement these with real-time enforcement to ensure safety at the moment of decision — not only in retrospective audit.

5.2 The 4-Stage Lifecycle Directive

A mandatory four-stage lifecycle governs all AI deployments:

Stage	Action	Governance Requirement
Map	Red-teaming: identify all potential impact surfaces and risk vectors before deployment	Documented risk register; stakeholder sign-off
Measure	Quantify frequency and severity of identified risks through rigorous benchmarking	Baseline metrics established pre-deployment
Mitigate	Mandate system-level meta-prompts, safety filters, and constrained user interfaces	Technical controls documented and tested
Manage	Implement long-term oversight, real-time behavioural enforcement, mandatory incident reporting	Continuous monitoring; version control; rollback procedures

5.3 Shared Accountability and Content Provenance

Total AI governance requires shared accountability across developers, deployers, and users. Alignment with global standards requires commitment to content provenance: digital watermarking of AI-generated media serves as a strategic defence against the Liar's Dividend — the erosion of trust in authentic media caused by the proliferation of hyper-realistic deepfakes.

5.4 Environmental Sustainability

AI deployment carries significant hidden costs. Large-scale model inference demands substantial energy and water for cooling, contributing to carbon emissions and inviting regulatory scrutiny. Enterprise governance mandates:

- Consumption Monitoring: Continuous tracking of data centre energy and water usage.
- Compute Optimisation: Implementing inference-time compute optimisation to avoid over-resourcing low-complexity tasks.
- Provider Selection: Preference for hardware and cloud providers with transparent carbon reporting.

5.5 Intellectual Property Obligations

Current legal landscapes require human authorship for copyright and patent protection. Purely AI-generated outputs remain ineligible for these protections. Two obligations follow:

- **Human-in-the-Loop Logs:** Employees must maintain detailed logs of human contributions and iterative refinements for any AI-assisted creative or inventive work.
- **IP Audits:** Periodic reviews of AI-assisted outputs to ensure they meet criteria for human authorship and protectability.

CONCLUSION

Governing Generative AI requires the seamless integration of three pillars: technical grounding (RAG and CAG to eliminate hallucination), structured reasoning (CoT, Self-Consistency, Tree of Thoughts, PRMs to ensure logic integrity), and rigorous governance (lifecycle management, regulatory alignment, shared accountability).

At the intersection of these three pillars sits the central governance challenge of autonomous AI: agentic drift. Unlike all other AI risks, drift does not announce itself. It accumulates silently within systems that continue to function. Outputs appear reasonable. No alarm sounds. The human approves, reflexively. And the authority gap widens.

The present response to drift — continuous monitoring, behavioural baselines, version control — is necessary but structurally dependent on consistent human action under variable conditions. In high-stakes environments, this dependency is the gap between governance policy and governance reality.

The structural prevention approach addresses this gap at the architectural level: offer-only logic, hardware-enforced deliberation, inaction as the failure mode. These mechanisms do not respond to drift. They remove the conditions under which certain categories of drift can occur.

Achieving Verifiable AI — AI that is both powerful and demonstrably accountable — requires deploying systems that are transparent by design, auditable by architecture, and constitutionally incapable of usurping the human authority they are built to serve.

仁義禮智信 · 止於至善

Benevolence · Righteousness · Propriety · Wisdom · Trust

謙虛·沉默·尊严·仁

REFERENCES

Cai, C. J., et al. (2023). Automation bias in AI-assisted clinical decision-making. *The Lancet Digital Health*, 5(3), e123–e131.

European Parliament. (2024). *EU AI Act*. Official Journal of the European Union.

IBM. (2026). *Agentic drift: The hidden risk in AI production deployments*. https://www.ibm.com

Infocomm Media Development Authority (IMDA). (2020). *Model AI governance framework* (2nd ed.). Singapore. https://www.imda.gov.sg

Kyndryl. (2026). *Preventing agentic AI drift: Governance and monitoring strategies*. https://www.kyndryl.com

National Institute of Standards and Technology (NIST). (2023). *AI Risk Management Framework (AI RMF 1.0)*. U.S. Department of Commerce.

Non-Agentic AI Governance Singapore. (2026). *Enterprise AI Governance Protocol: Module 4 Strategic Summary*. CET946, Nanyang Technological University.

World Health Organization (WHO). (2021). *Ethics and governance of artificial intelligence for health*. Geneva: WHO.
