AnimusLab

Comments on Draft 'Guidance on Regulatory Principles for Model Risk Management, 2026'

Submitted to: The Chief General Manager, Operational Risk Group, Department of Regulation, Reserve Bank of India

Submitted by: **Tanishq Dasari**, Lead Researcher & Founder, AnimusLab

Contact: tan@animuslab.dev | animuslab.dev

Date: July 2026

Reference: Draft 'Guidance on Regulatory Principles for Model Risk Management, 2026' (Press Release 2026-2027/528, June 24, 2026)

AnimusLab is an independent research organisation based in India, developing open-source governance infrastructure for autonomous AI systems in regulated financial markets. Our primary technical output, Anchor, is an open-source governance engine (*pip install anchor-audit*) that enforces a cryptographically signed rule set across static code analysis and runtime AI decision interception, with rule mappings to the RBI FREE-AI recommendations among other frameworks. We welcome the opportunity to comment on this draft Guidance and structure our comments against the specific paragraphs where we believe a technical implementation perspective adds value.

On Paragraph 21 (Model Inventory) and Paragraph 9 (MRMF Scope, including AI/ML)

Paragraph 21 requires REs to maintain an inventory of models such that 'no model is used, relied upon, or deployed unless it is part of inventory.' We submit that an equivalent discipline is needed one layer up: at the level of the governance rules themselves, not only the models they govern.

In practice, an RE's model risk policies, AI/ML controls, and validation criteria are frequently built with reference to external frameworks - NIST AI RMF, OWASP LLM Top 10, FINOS AI Governance Framework, and this Guidance itself once finalised. These external frameworks are revised over time. We have observed, in building Anchor's rule-mapping engine, that there is currently no standard mechanism by which an RE can demonstrate that its internal governance rules remain current against the external frameworks they were originally built to satisfy. A rule set built against NIST AI RMF version X can silently fall out of alignment with version X+1 without the RE being aware, in the same way an undocumented model would fall outside the inventory discipline this Guidance rightly requires.

We have built and released an open-source tool, Canon (v0.1.0, Apache 2.0, github.com/AnimusLab/Canon), that addresses this specific gap. It monitors external governance sources for rule changes, produces a structured evidence record of any change detected, and routes that evidence to a human approval gate before any internal rule set is updated - no automated change occurs without individual, attributed sign-off. We raise this because we believe the Guidance's inventory discipline in Paragraph 21, and its requirement in Paragraph 9 that the MRMF be comprehensive, would benefit from an explicit acknowledgement that governance rule sets themselves require the same inventory and currency discipline as the models they govern.

On Paragraph 41 (Change Management - Record of Changes and Approvals)

Paragraph 41 requires REs to 'maintain a comprehensive record / log of changes and versioning, and approvals.' We support this requirement and wish to highlight a specific technical property that, in our experience, distinguishes a record that satisfies supervisory examination from one that does not: tamper-evidence.

A change log that is itself editable after the fact - even where access is restricted to authorised personnel - does not allow an RE to demonstrate to a supervisor, with certainty, that the log was not altered following a change. We recommend the Guidance specify that the change and approval record required under Paragraph 41 be cryptographically tamper-evident: for example, through hash-chaining, such that each entry's hash incorporates the hash of the entry before it, making any retroactive alteration to a historical entry detectable. This is a well-understood technique, already used in append-only audit log design, and does not require disclosure of the underlying model or data to be verifiable - an RE can demonstrate the integrity of its change record without exposing proprietary model details to a third party.

On Paragraph 56 (Models with Dynamic or Automatic Updates)

Paragraph 56 requires 'strict justifications for enabling automatic updates' and 'more stringent and frequent monitoring' for models with dynamic or automatic update behaviour. We submit one structural principle that we believe strengthens this requirement: the update mechanism itself should be incapable of bypassing human review, by construction, rather than by policy alone.

In our implementation experience, a policy stating that automatic updates require justification is necessary but not sufficient if the underlying system technically permits an update to take effect without that justification being recorded first. We recommend REs be expected to demonstrate that their dynamic-update pathways are architected such that no model or rule update can take effect without a corresponding, individually attributed approval record being written first - not as a parallel compliance step, but as a precondition the system enforces. This distinction - approval as a technical precondition for the update path itself, rather than a separate documentation exercise - is, in our view, what gives Paragraph 56's 'strict justification' requirement teeth.

On Paragraph 57 (Enhanced Documentation - Traceability, Reproducibility, Auditability)

Paragraph 57 requires enhanced documentation of AI models 'to enable traceability, reproducibility, and auditability.' We recommend the Guidance specify, or the final version clarify through guidance notes, the minimum technical properties such documentation should possess to be considered auditable in practice rather than in form. Based on enforcement precedent we are aware of in other jurisdictions - notably the U.S. CFPB's 2024 action against a major financial institution, which found that AI-assisted decisions could not be explained at the level of the individual decision despite the existence of model-level documentation - we suggest that documentation satisfying Paragraph 57 should, for material AI/ML models under Paragraph 52, include at minimum: a machine-readable reason or basis for each individual model-driven decision (not only aggregate model documentation), a record of which model and policy version was in effect at the time of that decision, and a tamper-evident timestamp. Aggregate documentation of model behaviour, while necessary, does not by itself allow an RE to answer a supervisor's question about a specific decision.

On Paragraph 60 (Human Oversight - Kill-Switch Arrangements)

Paragraph 60(ii) requires 'override, suspension, or deactivation mechanisms, including kill-switch arrangements.' We welcome this requirement and offer one observation from our research on AI governance failure modes in financial services: the effectiveness of a kill-switch is contingent on the RE's ability to detect, in time, that deactivation is warranted.

Our analysis of three well-documented financial AI/algorithmic failures - including a case where a deprecated, unmonitored code path executed without authorisation and caused losses exceeding $460 million within 45 minutes - indicates that the technical capacity to halt a system is necessary but insufficient without a corresponding mechanism to detect, at the moment a problem begins, that halting is required. We recommend the Guidance's expectations under Paragraph 60 be read together with Paragraph 36's requirement that model outputs be 'replicated and stable in production,' such that REs are expected to demonstrate not only that a kill-switch exists, but that monitoring granular enough to trigger it in a timely manner - ideally at the level of individual decisions rather than aggregate performance review - is in place for models in scope of Paragraph 52's risk tiering.

On Paragraph 52 (AI Model Risk Tiering - Reliance and Autonomy)

Paragraph 52 requires REs, when tiering AI models, to 'consider the extent of reliance and the level of autonomy placed on the model outputs for decision-making.' We submit a related but distinct concept that we believe merits explicit recognition in the final Guidance: the distinction between a model producing an inaccurate output, and a system taking an action it was never authorised to take, regardless of the accuracy of that output.

These are different risk categories with different controls. A model risk tiering framework built solely around output accuracy and autonomy may not adequately capture the risk that arises when an AI/ML system - particularly one with tool-calling or multi-step autonomous capability, increasingly relevant as REs adopt agentic AI architectures - executes an action outside its intended authorisation boundary even when its underlying output was technically correct. We recommend the final Guidance, or subsequent AI-specific guidance referenced in Paragraph 2, explicitly distinguish execution authority risk (whether an action was permitted to occur) from model output risk (whether the output was accurate), as the controls appropriate to each differ materially.

Summary and Availability

The observations above are drawn from our experience building Anchor, an open-source governance engine with rule mappings to the FREE-AI recommendations underlying this Guidance, and Canon, an open-source tool addressing the governance-rule-currency problem described under Paragraph 21 above. Both are available under the Apache 2.0 license at github.com/AnimusLab. We would be glad to make a governance assessment of an RE's AI-adjacent codebase available for illustrative purposes, on a no-cost, read-only basis, should that be useful to the Department's consideration of implementation guidance following finalisation of this Guidance.

We welcome any questions the Department may have regarding the above at **tan@animuslab.dev**.

Respectfully submitted,

Tanishq Dasari
AnimusLab