

Dr. Anandkumar Prakasam

Chennai, India

root.anand@gmail.com · +91 97906 36324

22 July 2026

The Chief General Manager

Operational Risk Group, Department of Regulation

Reserve Bank of India, Central Office

Mumbai

Subject: Comments on the draft Guidance on Regulatory Principles for Model Risk Management, 2026

Dear Madam or Sir,

I write in response to the invitation for public comments on the draft Guidance on Regulatory Principles for Model Risk Management, 2026, released for public comments on 24 June 2026. This submission is made in my personal capacity as an independent researcher and practitioner in AI governance. It represents no client, employer or commercial mandate.

I find the draft well-conceived and, in several respects, ahead of comparable instruments internationally. Its insistence that the regulated entity remains accountable irrespective of model origin at paragraphs 8 and 45, its treatment of human oversight as an enforceable control rather than a formality in Chapter V, and its recognition that materiality rather than technical labelling determines what counts as a model at paragraph 7(3), are all sound and important choices. My comments are offered in that same constructive spirit. Each identifies a place where the Guidance can be made more operational and verifiable without adding compliance burden, and several propose ready definitions or artifacts that the regulated entity community could adopt.

The comments cluster around six themes. Theme A concerns making human oversight measurable rather than merely present. Theme B concerns defining the kill-switch as a process with a defined safe state rather than a single button. Theme C concerns making explainability thresholds testable and recognising behavioural assurance alongside adversarial testing. Theme D concerns strengthening third-party and inherited-model accountability through a standard disclosure. Theme E concerns requiring tamper-evident evidence so that the duty to demonstrate means more than the duty to assert. Theme F concerns reading this Guidance together with the Digital Personal Data Protection Act, 2023 for the large number of models that process personal data.

I would be glad to assist the Reserve Bank further on any of the points raised here, including through the Department of Regulation or any working group that may be constituted for the AI specific requirements contemplated in paragraph 2 of the draft.

Yours faithfully,

Dr. Anandkumar Prakasam

Comments

A. Human oversight should be measured, not merely present, paragraphs 60 to 63

Observation. The draft rightly requires human-in-command arrangements and that personnel be able to effectively challenge, override or escalate at paragraph 62, and that oversight be periodically reviewed and strengthened at paragraph 63. As drafted, however, a regulated entity could satisfy the letter of these clauses by placing a human next to a decision without any evidence that the human is exercising genuine judgment. In practice a reviewer who approves nearly everything provides the appearance of oversight while functioning as what I have termed a liability sponge. Automation bias, which the draft correctly flags at paragraph 61, is not a character flaw of individual staff, it is the predictable output of interfaces that reward agreement and measure only throughput.

Suggestion. Add to paragraph 63 an expectation that regulated entities maintain quantitative oversight-effectiveness indicators, reviewed at least annually by the risk management committee, such as an override or disagreement rate with unusually low rates triggering review, a blind-acceptance indicator capturing approvals that show no evidence of genuine engagement such as sub-threshold review time, and review latency and decision-reversal rates on human-approved cases. This converts the phrase reviewed and strengthened from a qualitative undertaking into something an internal auditor and a Reserve Bank inspector can actually examine. I would be glad to share a fuller indicator set if useful.

Suggestion. In paragraph 62, consider adding that the cost or friction of recording a dissent or invoking the human-assistance or override path should not be materially higher than the cost of approval, since otherwise the oversight mechanism is structurally biased toward rubber-stamping even where competence exists.

B. The kill-switch should be defined as a process with a defined safe state, paragraph 60(ii)

Observation. Paragraph 60(ii) requires override, suspension or deactivation mechanisms, including kill-switch arrangements. This is welcome and important. As worded, however, it risks being read as a single control, a button, rather than a process. A deactivation that halts a model without a defined fallback can convert a model-risk event into an operational-continuity event, in effect an outage. This intersects directly with paragraph 43 on business continuity and fallback mechanisms, but the link is not made explicit.

Suggestion. Clarify paragraph 60(ii), or add an explanation, to require that override, suspension and deactivation arrangements specify in advance the trigger conditions and the authority empowered to invoke them, the defined safe state the system transitions into such as queue-and-hold, abstain, or route-to-human in the manner of paragraph 43, and a tamper-evident record of the invocation. A kill-switch without a designed degraded mode is, in effect, an unplanned outage.

Suggestion. For AI models with dynamic or automatic updates at paragraph 56, consider requiring that the deactivation authority be able to revert to the last validated and approved model state, not merely stop the current one. A graduated narrowing of authority, rather than a binary on or off, materially reduces continuity risk.

C. Make explainability thresholds testable, and recognise behavioural assurance, paragraphs 54, 54(1) and 55

Observation. Paragraph 54(1)(i) requires regulated entities to define the explainability and transparency thresholds for all AI models. Without guidance on how thresholds are expressed, entities may set them as unfalsifiable narrative statements. Separately, the structured challenge processes at paragraph 55, including red-teaming or equivalent testing, are the right instinct, but in current industry practice red-teaming overwhelmingly means adversarial testing such as jailbreaks and prompt injection, which is necessary but does not catch the most common real-world failures of customer-facing models, where no rule is broken and no attacker is present. The Air Canada chatbot that committed the airline to a non-existent refund policy is the canonical example. It was a collision between two legitimate objectives, being helpful and respecting institutional authority, that adversarial testing is structurally blind to.

Suggestion. In paragraph 54(1)(i), clarify that explainability thresholds should be expressed in verifiable terms, for instance that the model must surface the principal factors behind a material decision in a form a reviewer or an affected consumer can act upon, and tie the threshold to the decision's materiality tier, which the draft already establishes at paragraphs 17 to 20 and 52. Where full explainability is not achievable, the enhanced controls at paragraph 54(1)(ii) are well judged, and I would suggest listing output-corroboration before reliance expressly among them.

Suggestion. In paragraph 55, consider broadening red-teaming or equivalent to expressly include behavioural or normative testing, that is, testing how the deployed system resolves conflicts between competing legitimate objectives such as helpfulness against authority, safety against user intent, and global policy against local law, and not only its resistance to malicious attack. This is exactly the class of failure most relevant to the customer-facing and generative models at paragraph 59. I have catalogued fourteen such recurring boundaries and a measurement method, which I would be happy to contribute to any standard-setting effort.

D. Strengthen inherited-model accountability with a standard disclosure, paragraphs 45 to 48, 51 and 53

Observation. The draft holds the regulated entity fully accountable for third-party models at paragraph 45 and requires independent validation notwithstanding any validation provided by the third-party provider at paragraph 46(i). This is correct and necessary. Paragraphs 47 to 48, however, place the burden of due diligence on the entity while giving it limited leverage to obtain the information needed, particularly for foundational and frontier models where the provider discloses little, which is the paragraph 51 scenario.

Suggestion. Consider recommending, even if not mandating, a standard inherited-risk disclosure that entities request from model providers as part of the paragraph 48 contractual documentation, capturing the provider's residual-risk register, the conditions under which the model's assurances hold, and the foreseeable-reliance limitations. A standardised artifact would make paragraph 47 due diligence comparable across providers, give entities collective leverage to obtain the paragraph 48 minimum documentation, and create a clean audit trail for the risk-committee oversight of third-party models. I have developed such a template, which I call a Model Risk Inheritance Statement (MRIS), and would gladly share it.

Suggestion. In paragraph 53, consider explicitly naming provider-driven model updates as a change-management trigger, cross-referencing paragraphs 38 to 42 and 56, since a silent capability change pushed by a foundational-model provider can alter behaviour without any action by the entity.

E. Require tamper-evident evidence so that demonstrate is verifiable, paragraphs 21 to 24, 33, 41, 57 and 63

Observation. Several clauses turn on the entity being able to demonstrate what was done, the inventory at paragraphs 21 to 22, the ten-year retention of decommissioned models at paragraph 23, validation reports at paragraph 33, change and version logs at paragraph 41, AI traceability, reproducibility and auditability at paragraph 57, and the oversight record at paragraph 63. The draft does not, however, address the integrity of these records. Logs held in mutable databases can be altered, backdated or selectively reconstructed after an incident, precisely when their integrity matters most. The reproducibility goal at paragraph 57, in particular, is difficult to honour years later unless the relevant state was sealed at the time.

Suggestion. Consider adding, in Chapter III or Chapter VI, an expectation that records material to model governance, at a minimum validation outcomes, change logs, AI traceability records and the human-oversight and override record, be maintained in a manner that is tamper-evident and independently verifiable, for instance cryptographically sealed and trusted-timestamped at the time of the event and retained for the applicable period. This raises the standard from we kept a log to we can prove what was known and decided at a given time, which is what supervisory examination, grievance redress at paragraph 25, and any subsequent dispute actually require. The technical means are mature, and I have built a working reference implementation that produces such sealed, independently re-verifiable records, so the property can be evidenced today. The Guidance need not prescribe the technique, since the property of verifiable integrity is what matters.

F. Read this Guidance together with the Digital Personal Data Protection Act, paragraphs 5, 8, 25, 28 and 45

Observation. A large share of the models this Guidance governs, a credit model, a fraud engine, an onboarding or a collections assistant, are at the same moment an activity of processing personal data under the Digital Personal Data Protection Act, 2023. The two regimes are not neighbouring subjects that happen to touch, they are two readings of the same system. This matters for a practical reason. The Guidance, read on its own, is proved at a periodic supervisory inspection, on the supervisor's initiative, with no reversed burden and no outside claimant. The data-protection regime, by contrast, reverses the burden of proof onto the data fiduciary at section 6(10), makes the compliance duty non-delegable at section 8(1) so that it cannot be contracted away to a vendor, requires accuracy and completeness where personal data drives a decision that affects a person at section 8(3), and arms the regulator with the production powers of a civil court at section 28(7). For a personal-data model, the same governance acts this Guidance asks for, the inventory, the validation report, the monitoring log and the oversight record, are therefore separately required, and separately provable, by law that already applies.

Suggestion. Consider acknowledging this intersection expressly, for example in a note to paragraph 5 on reading the Guidance alongside other applicable law, or at paragraph 28 on development-data governance, so that a regulated entity treats the governance record for a personal-data model as serving both regimes at once. A single tamper-evident model governance record, of the kind described in Theme E, would then discharge the supervisory expectation under this Guidance and stand ready as the fiduciary's proof of compliance under section 6(10) of the Digital Personal Data Protection Act, 2023, without the entity having to build the same evidence twice. Framing the two regimes as one evidentiary duty, rather than two, is the most economical path to compliance for the entities this Guidance reaches, and it

lowers the risk that a model judged adequate at a periodic inspection is later found wanting when the reversed burden of the data-protection regime is applied to the same decision.

Closing

The draft Guidance arrives at the right moment and takes the right shape. Its accountability rule for models of any origin, its treatment of human oversight as an enforceable control, and its materiality-based definition of a model are the choices that will decide whether the framework holds up in practice, and the draft gets each of them right. The six themes above pursue a single idea. A framework of this kind succeeds when a regulated entity can show, and an examiner can verify, that the controls it describes actually operated on the decisions that reached customers. The changes I suggest move the Guidance a short distance further in that direction, from a standard that is documented to one that is demonstrable, at little added burden to regulated entities, since the records in question can be produced by the systems that already execute the decisions. I hope these suggestions are useful, and I remain available to assist the Reserve Bank as the AI specific requirements contemplated in paragraph 2 are developed.

Respectfully submitted,

Dr. Anandkumar Prakasam

PhD (Computer Applications), LLB (Hons) | root.anand@gmail.com | 9790636324 | Chennai.