NIBM WORKING PAPER SERIES

AI-Enabled Cyber Risks in Banking: Institutional Preparedness, Governance and the Regulatory Response

Alka Vaidya
Deepankar Roy
Pramod C. Mane

Working Paper
(WP65/2026)

NATIONAL INSTITUTE OF BANK MANAGEMENT
Pune, Maharashtra, 411048
INDIA
July 2026

Citation Guideline:

Vaidya, Alka; Deepankar Roy and Pramod C. Mane (2026), "AI-Enabled Cyber Risks in Banking: Institutional Preparedness, Governance and the Regulatory Response". NIBM Working Paper Series WP 65/July.

https://www.nibmindia.org/static/working_paper/NIBM_WP65_AVDRPM.pdf

AI-Enabled Cyber Risks in Banking: Institutional Preparedness, Governance and the Regulatory Response

Alka Vaidya, Deepankar Roy, and Pramod C. Mane

NIBM Working Paper No. 65
July 2026

ABSTRACT

The rapid emergence of frontier Artificial Intelligence (AI) systems capable of discovering and chaining software vulnerabilities at machine speed has introduced a qualitatively new category of cyber risk for the banking sector. In response, India's central bank has directed regulated entities to complete a board-approved gap assessment of AI-related cybersecurity risk and to formulate a timebound action plan (RBI, 2026). This paper examines what stands between that regulatory check and institutional readiness in the Indian banking sector, and sets out the policy response required to close the distance.

The principal constraint facing Indian banks is not awareness or willingness but capability. Banks lack direct access to, or working experience with, the frontier AI tools they are now expected to defend against; visibility into the AI and software components embedded in their own technology stacks remains incomplete; compute and graphics processing unit (GPU) infrastructure required for AI-enabled defence is scarce at both the institutional and national level; and cyber-security governance structures have not yet been extended to cover the full range of AI-related exposure, from vendor-embedded AI to informally adopted "Shadow AI[1]" tools (Puthal D et al., 2025). The paper sets out the implications of these constraints for regulatory policy, institutional governance, and sector-level coordination, and concludes that the central policy task is not to mandate AI adoption for its own sake, but to build the shared infrastructure, disclosure standards and governance discipline that allow banks to respond to AI-enabled threats in a measured, evidence-based and coordinated manner (Luca C, 2025).

Alka Vaidya (Corresponding Author)
alka@nibmindia.org

Deepankar Roy
d_roy@nibmindia.org

Pramod C. Mane
pramod.mane@nibmindia.org

[1] Shadow AI: is the use of unauthorized or unverified artificial intelligence tools, applications, or models by employees for work-related tasks.

AI-Enabled Cyber Risks in Banking: Institutional Preparedness, Governance and the Regulatory Response

1. Introduction

Banking supervision has periodically had to absorb new categories of technology risk into its existing cyber-security architecture internet banking, mobile payments, cloud outsourcing, and now artificial intelligence. What distinguishes the present moment is the speed at which the underlying technology is moving relative to the institutional capacity to understand it. Over the period of 2025 and into 2026, frontier AI models demonstrated the ability to scan large codebases for vulnerabilities, including zero-day flaws that had remained undetected in widely used software for years, and to chain multiple lower-severity weaknesses into a single viable path of exploitation(CERT-IN, 2026). These capabilities did not remain confined to research settings; they entered public and policy discourse rapidly enough that government and central banks, including Indian finance ministry, began referring to them explicitly in their communications with the banking sector.(PTI, 2026)

India's central bank directive (RBI, 2026) asking banks and other regulated entities to complete a board-approved gap assessment of AI-related cyber risk, and to formulate a timebound action plan, should be read in this context. It is not a conventional compliance circular prescribing a fixed set of controls. It is, in effect, a request that each institution examines itself because the regulator does not yet possess, and cannot yet prescribe, a single template for a risk whose contours are still being established. This places banks in the position of having to assess their exposure to a class of frontier AI capability that almost none of them have directly observed, tested, or operated against.

This paper assesses the gap between that regulatory expectation and the operational reality of Indian banking institutions today. Section 2 sets out the substantive risk landscape and the institutional response it demands, organised thematically rather than as a record of any single discussion. Section 3 draws out the policy implications at three levels: the regulator, the individual institution, and the banking sector collectively. Section 4 concludes.

2. The AI-Enabled Risk Landscape and the Institutional Response

Indian banks are being asked to defend against a category of threats most have not directly encountered, while simultaneously being expected to build the institutional capability to manage it. This section sets out both halves of that problem together: the substantive risks AI introduces into the banking threat landscape, and the practical and strategic response such risks demand of institutions. The two cannot be separated cleanly, because the shape of an adequate response depends entirely on the specific nature of the risk it is meant to address.

2.1 An Unfamiliar Threat, Assessed at a Distance

The most basic difficulty Indian banks face is that they are being asked to prepare for a capability they have not themselves seen in operation. Frontier AI tool such as Mythos was originally tested within a closed user group and later Anthropic announced Project Glasswing, a collaborative cybersecurity initiative to a closed cohort of launch partners (Anthropic 2026). Subsequently, Anthropic extended the project to 15 more countries and 150 organisations providing critical infrastructure related services, however the names had not been disclosed. (Economic Times, 2026). This closed nature of frontier models like Mythos is likely to affect red-teaming[2] exercise of banks that would uncover novel attack paths. The reported capabilities of such tools rapid, deep vulnerability scanning, and the chaining of discrete weaknesses into a usable exploit path are significant enough to warrant serious institutional attention, but the absence of direct exposure makes it genuinely difficult to translate that attention into precise controls. An institution cannot calibrate its defences against a threat whose mechanics it can only infer second-hand.

This is a capability gap rather than a question of institutional reluctance. Indian banks have, on the whole, been willing to strengthen their cyber posture in response to regulatory guidance; what they lack is a reliable way to validate vendor claims about AI-related defences, to design realistic AI-based red-teaming exercises, or to test whether existing security operations centre (SOC), security information and event management (SIEM), vulnerability assessment and penetration testing (VAPT) and patch-management practices hold up against AI-accelerated attack patterns, when the attacking technology itself remains opaque to them. A controlled national or sectoral testing facility one that allows banks to study exploit chains and build defensive playbooks without exposing sensitive data or compromising confidentiality would address a structural gap that no individual bank can close through its own initiative.

2.2 From Vulnerability Discovery to Exploitable Chains

AI systems capable of scanning code at scale change the nature of vulnerability management in a specific, technical way: they do not merely find more vulnerabilities, they find combinations of vulnerabilities that function together. A flaw that would normally be classified as medium severity, and therefore queued well behind higher-priority items, can become critical the moment it forms one link in a viable chain toward privilege escalation or lateral movement. Conventional vulnerability management practice built around severity scores, asset criticality and exposure was not designed with this kind of compounding risk in mind, and banks already receiving large volumes of findings from routine scanning, audits and VAPT need a different basis for deciding what to fix first.

The appropriate shift is toward exploitability-based prioritisation: identifying which vulnerabilities are actually reachable from outside the perimeter, which can be chained with other known weaknesses, and which create a realistic path to system compromise, rather

[2] Red teaming is a proactive, adversarial exercise where a group of ethical hackers or experts simulates real-world attacks against an organization's systems.

than ranking findings primarily by abstract severity. AI-based red-teaming is one route toward this kind of analysis, but it remains an immature discipline without settled standards for what counts as adequate AI-based testing; vendor and consulting-firm claims in this space should be treated as a starting point for proof-of-concept evaluation rather than as a substitute for independent validation. A further complication is that the output of effective AI-based red-teaming, a documented exploit chain is itself highly sensitive information, and needs handling discipline at least as rigorous as the vulnerability it describes, particularly in the window before a patch or compensating control exists.

2.3 Patch Discipline Under Compressed Timelines

AI-enabled vulnerability discovery shortens the window of time between the discovery of a vulnerability and its potential exploitation, putting direct pressure on patch turnaround times. But banking infrastructures are not amenable to just patch faster as a general answer. Patches must be tested, certified and checked for stability before being put in place; an untested patch may itself cause an operational incident, reduce the availability of an application or disturb services that face the customers. Some patches do not come into effect until after a system reboot which interacts with disaster-recovery arrangements and uptime commitments in ways that cannot simply be ignored.

A further constraint is that banks frequently cannot patch the underlying issue themselves. Where a zero-day is identified in proprietary software, the fix depends on the original equipment manufacturer (OEM) releasing a patch, and the bank's interim defence rests on compensating controls, network-level restriction, monitoring and virtual patching. Virtual patching, in turn, is only as good as the telemetry and attack-pattern knowledge behind it; a generic perimeter rule is unlikely to substitute for precise understanding of what specifically needs to be blocked. This argues for tighter integration between vulnerability management, SOC operations, application owners, OEMs and perimeter security, since none of these functions can compress the patch-to-exploit window on its own.

2.4 Detection Capability and the Limits of Automation

Existing SOC, SIEM and security orchestration, automation and response (SOAR) capability was largely built to detect human-paced attack behaviour, and now it would need a careful re-examination against AI-enabled patterns: adaptive payload generation, automated reconnaissance, vulnerability chaining executed in rapid succession, AI-assisted social engineering, and unusual inside-out connectivity that conventional rule sets may not flag. One promising design places AI agents between the SIEM and SOAR layers, using existing telemetry to accelerate Level 1 and Level 2 triage, but this only works with careful design, ongoing validation, and a human-in-the-loop checkpoint retained for escalation and final decision-making.

The more basic discipline that is necessary here is to approach AI-generated discoveries as inputs that need to be confirmed, not as conclusions that can be acted upon directly. AI output might be partial, contextually poor or just plain inaccurate. A bank’s operational decisions patch priority, incident response, regulatory reporting should not be

made on unverified AI data. This is to argue against the unchecked spread of agents and bots in the institution. A more defensible paradigm is a small number of narrowly scoped agents, each created for a specific task vulnerability assessment, uniform resource locator (URL)-change detection, telemetry review, SOC triage, rather than broad, loosely regulated automation that is harder to validate and harder to audit.

2.5 Talent, Compute and the Limits of Going It Alone

Two structural constraints sit beneath nearly every institutional response discussed so far: people and infrastructure. Building AI-enabled cyber defence requires understanding not only conventional cybersecurity but AI models, agent design, prompt-based systems, telemetry pipelines, data governance and model validation. Low-code and no-code agent-building tools lower some of this barrier, but they do not remove the need for skilled personnel who can define the right use case, supply the right institutional context, validate outputs, and integrate agents into existing controls.

Compute is the second constraint, and it is not primarily a cost problem. GPU availability is genuinely scarce for individual banks and, to a meaningful extent, for the country as a whole, which pushes institutions toward cloud-based AI services and the confidentiality, data-residency, log-integration and compliance questions that come with them. No individual bank is in a position to build frontier-scale AI capability independently, and the more realistic path is coordinated capability-sharing: a shared facility, supported by an appropriate sectoral institution, for AI-enabled testing, red-teaming, research, training and threat intelligence, designed from the outset with confidentiality, data segregation, governance and procurement clarity built in. The case for this kind of coordination is strengthened by a simple fact: AI-enabled threats rarely respect institutional boundaries. A vulnerability in a common vendor product, an open-source library, or an AI-generated code template is latent risk across every bank that uses it, not a risk contained within one.

2.6 Supply-Chain Opacity and the Governance Gap

A bank may have strong internal controls, but its exposure can still increase via a weak vendor, fintech partner or cloud provider. The bigger problem is visibility: banks often don't know what open source libraries, AI-generated code or third-party components are embedded inside the products they've licensed. We've seen this pattern in previous supply-chain incidents not knowing precisely where a vulnerable component, such as Log4j, was embedded was the core operational problem then, and it's happening again here, but in a more complex way, because AI-generated code can be reused across many products built in the same way, spreading a single flaw across institutions that have no reason to suspect a common point of failure.[3]

[3] Cyber Safety Review Board (2022), *Review of the December 2021 Log4j Event*, National Institute of Standards and Technology, 11 July 2022.

Software Bills of Materials (SBOMs)[4] are there to fill exactly this gap, but the equivalent for AI components, a "AI-BOM", and for application programming interfaces (APIs) are immature, with vendors already claiming solutions far ahead of any agreed standard for what such an inventory should contain. Cloud and multi-tenancy configurations introduce a correlated weakness: The computing needs of AI are nudging banks toward cloud infrastructure, but legacy cloud-hosted products have not always been designed with current standards for tenant isolation, and some still segregate customer data only at the field level of a shared table rather than true schema-level separation, a design choice that, if exploited, can impact multiple institutions through a single breach. This is not something that can be solved by a bank's internal controls alone, it requires vendors and OEMs to disclose what their products actually contain, and a settled standard for what that disclosure should look like.

2.7 Governance That Extends Beyond Customer-Facing AI

AI governance at a bank is too often restricted to customer-facing or business-use models – chatbots, credit underwriting tools, analytics platforms with AI used in cybersecurity itself, AI embedded inside vendor solutions, and AI adopted informally by individual departments outside of the same scrutiny. This is the governance blind spot that makes "shadow AI" possible: tools that are in active use across an institution that were never inventoried, never risk-assessed, and never brought under the same control framework as sanctioned systems.

Multiple, overlapping advisories from regulators, government bodies and industry associations only compound the problem, as institutions end up consolidating broadly similar requirements expressed in different language into their own control libraries and action trackers useful internal discipline, but not a substitute for converting advisory language into specific, implementable controls covering AI inventory, model and agent testing, shadow AI detection, AI-enabled SOC capability, third-party AI tool use and data leakage through AI systems. A truly effective governance architecture views all of these as a single AI risk surface, not a set of exceptions to be managed in isolation.

2.8 A Disciplined, Resource-Conscious Strategy

The right institutional posture toward AI-enabled cyber defence is not maximal adoption of AI tools, but a focused, economical and context-aware strategy that strengthens existing cyber-resilience capability rather than displacing it. Given that no individual bank can match the compute scale of a frontier AI platform, the more realistic objective is identifying a small set of cyber-risk use cases vulnerability assessment, URL scanning, change monitoring, threat modelling, attack-surface review, SOC triage, exploit-path analysis where AI adds clearly measurable value, and resisting the temptation to apply it everywhere. Token and compute consumption are real, recurring costs, and an unfocused deployment of AI can become expensive without making an institution materially more secure.

---

[4] National Telecommunications and Information Administration (2021), *The Minimum Elements for a Software Bill of Materials (SBOM)*, U.S. Department of Commerce, 12 July 2021.

Within that focused scope, the more durable design choice is a platform built to host small, purpose-specific AI agents for vulnerability scanning, URL-change detection, architecture review, threat modelling, API testing, control validation, SOC alert triage rather than general-purpose tools that search broadly and return generic output. These agents are only as useful as the institutional context they are given: structured repositories such as retrieval-augmented generation (RAG) libraries, decision libraries, vulnerability knowledge bases, architecture and threat-modelling templates, control libraries, and historical incident and telemetry records give an agent the specificity that distinguishes a genuinely useful output from a generic one. Building this capability is not purely a technical exercise either; cybersecurity, technology, risk, compliance and business-domain expertise all have a role in defining agent tasks, designing effective prompts, and validating what the agents produce.

An AI inventory is the foundational control beneath all of this. AI already enters a bank through customer-facing models, credit-risk tools, employee productivity software, vendor-embedded systems, cybersecurity platforms, and informal departmental use, and an institution that cannot enumerate where AI is actually operating cannot govern it. At the same time, the capability that makes tools such as Mythos concerning rapid identification of vulnerability chains is the same capability that makes them potentially valuable on the defensive side: a bank receiving thousands of discrete vulnerability findings needs exactly this kind of chain-level analysis to know which combination of weaknesses constitutes a genuine path of exploitation. The strategic task is to intentionally capture that defensive value, while keeping human judgement in the loop for anything that touches production systems, patching priority, incident response or regulatory reporting, and while keeping a tight rein on token usage and data exposure wherever cloud-based AI platforms are involved since vulnerability details, source code, customer data and internal telemetry sent to a third-party platform constitute their own separate exposure, apart from whatever risk the platform is being used to address.

3. Policy Implications

We believe the risk landscape and the institutional response demands carry distinct implications for three layers of decision-making, namely, the regulator, the individual institution, and the banking sector acting collectively. Instead of a calculated response, treating these as a single, undifferentiated set of action items runs the danger of creating an unfocused, resource-draining one. The following implications are arranged according to the degree of action that is necessary.

3.1 For the Regulator

1. Realistic sequence of the mandate. A board-approved gap assessment is a reasonable first step, but only if banks can assess a risk they have not directly observed." The regulatory entity in conjunction with the government and sectoral institutions, should define the criteria on which banks are to assess frontier-AI related exposure, without direct access to such tools, whether through indicative threat scenarios, a sectoral testing facility, or a phased approach that separates immediately actionable controls from longer-horizon capability building.

2. Expand expectations for disclosure on AI and supply chain visibility. A big part of the practical difficulty that banks face is the lack of visibility into the software and AI components embedded in vendor products. Regulatory guidance should drive convergence on SBOM-equivalent disclosure for AI components (an "AI-BOM") as a supervisory expectation for both banks and the vendors and OEMs that serve them, recognising that the standards for such disclosure are still maturing and may need to be developed jointly with industry.

3. Think of AI governance as an overlap with current cyber-security frameworks rather than an add-on. The current cyber-security baseline is strong, but not sufficient for AI-specific exposure. Rather than issuing a parallel AI-risk framework, the regulator should consider extending the existing master-direction structure with AI-specific annexes covering AI inventory, shadow AI, model and agent validation, and AI-enabled SOC detection mirroring how cloud-specific controls were earlier layered onto the existing baseline.

4. Examine the case for shared sectoral infrastructure. Compute and GPU scarcity is a structural constraint that no individual bank, and arguably no individual regulator, can resolve through guidance alone. The regulatory agencies, working with bodies such as the National Institute of Bank Management (NIBM), the Institute for Development and Research in Banking Technology (IDRBT) and the Indian Banks' Association, should assess the feasibility of shared or coordinated AI-testing and red-teaming infrastructure, with confidentiality, data segregation and governance designed in from the outset rather than retrofitted.

3.2 For Individual Institutions

1. Build an AI inventory before building AI defences. An institution cannot govern what it cannot see. A comprehensive inventory covering customer-facing models, vendor-embedded AI, employee productivity tools, and informally adopted "shadow AI" should be treated as a precondition for, not a parallel track to, AI-enabled cyber defence.

2. Shift vulnerability management to an exploitability-based priority. In a situation where AI-enabled chaining can turn a medium-severity finding into a critical attack path, severity scoring is no longer an adequate basis for patching priority. Institutions should focus on the capacity to evaluate not only severity, but reachability and chainability.

3. Use focused, task-specific AI agents instead of wide automation. In the end, frugal, well-defined agents for vulnerability scanning, URL-change detection or SOC triage provide more dependable, auditable value than general-purpose AI deployment, and are more defensible from a token-cost and data-governance perspective.

4. Make humans accountable for the outputs of AI systems. Where AI outputs have implications for patching priority, incident response, or regulatory reporting,

institutions should retain explicit human validation stages and resist the temptation to accept AI-generated discoveries as self-certifying, no matter how complex the underlying model. Apply data-governance discipline to AI tool usage itself, not only to its outputs. Vulnerability details, source code, and customer data shared with cloud-based AI platforms constitute a distinct exposure separate from the vulnerabilities such tools may find. Token-usage discipline and data-handling controls should be treated as core cyber-risk controls, not as a separate cost-management exercise.

3.3 For the Banking Sector Collectively

1. Coordinate on shared threats rather than compete on shared infrastructure. Where banks rely on common vendor products, open-source libraries, or AI-generated code templates, a vulnerability in one is latent risk in many. Sector-level information-sharing on such common exposures would reduce duplicated discovery effort without requiring any bank to disclose its own internal posture.

2. Pool scarce technical capability rather than duplicate it. Talent and compute constraints affect nearly every institution identically. A shared or institutionally coordinated facility for AI-enabled testing, red-teaming and training with confidentiality and segregation safeguards would allow smaller institutions in particular to access capability they could not build independently.

3. Build a shared evidence base through structured, anonymised research. Because no single bank can see across the sector, applied research of this kind has a distinct role in surfacing patterns such as common vendor exposure or recurring governance gaps that individual institutions are not well placed to identify on their own.

4. Conclusion

India's central bank directive asking banks to assess their AI-related cyber-security gaps has set a genuine institutional task in motion, but the evidence assembled in this paper suggests that the binding constraint on Indian banks is not motivation but capability. Banks are being asked to assess their exposure to frontier AI tools they have not directly seen, to inventory AI and software components that their own vendors will not always disclose, to build computational capability that scarce GPU access makes difficult to scale, and to extend governance structures that have historically treated cyber security as an institution-specific, competitively guarded function.

In many respects, the only credible method to begin regulating a risk whose technical ceiling is still moving is a board-approved gap assessment that begins with an honest institutional self-examination, rather than a uniform prescribed protocol. This paper proposes that the subsequent phase of policy response should focus on the structural gaps that individual banks are unable to resolve independently. These gaps include the need for a sectoral mechanism to coordinate on threats that, by their very nature, do not respect institutional boundaries, shared access to the type of testing capability that would enable

banks to comprehend frontier-AI behaviour without each institution having to acquire it independently, and visibility into AI and software supply chains.

The overarching lesson is that of disciplined proportionality. The reality of AI-enabled cyber risk is that it is both real and evolving rapidly. However, the appropriate response is not to engage in panic-driven procurement or adopt AI to the fullest. The patient work of developing AI inventories, exploitability-based vulnerability management, parsimonious and well-governed AI agents, and human-validated decision processes is layered onto, rather than replacing, the sound cyber-security fundamentals. Institutions and regulators that internalise this distinction are likely to be better equipped to manage the AI-enabled threat environment as it continues to evolve, in addition to meeting the immediate June deadline.

References

Anthropic (2026), "Project Glasswing Securing critical software for the AI era", 2026, https://www.anthropic.com/glasswing

CERT-In Advisory (2026), CIAD-2026-0020, "Defending Against Frontier AI Driven Cyber Risks", April 2026,https://www.cert-in.org.in/s2cMainServlet?VLCODE=CIAD-2026-0020&pageid=PUBVLNOTES02

Economic Times (2026), "Indian cyber, telecom, banking & finance firms get access to Mythos; IT left out", June 2026, https://economictimes.indiatimes.com/tech/artificial-intelligence/indian-cyber-telecom-banking-finance-firms-get-access-to-mythos-it-left-out/articleshow/131491369.cms?from=mdr&utm_source=contentofinterest&utm_medium=text&utm_campaign=cppst

Luca, C. (2025). Federated Learning Models for Privacy-Preserving Fraud Analytics Across Global Banking Networks. IEEE Access, 11, 114882-114897.

PTI (2026), "Nirmala Sitharaman urges bankers to brace for AI threats amid concerns over Anthropic's Mythos", April 2026, https://www.ptinews.com/story/business/amid-mythos-concerns-fm-asks-financial-sector-to-be-exceptionally-vigilant-on-cybersecurity/3601846

Puthal, D., Mishra, A. K., Mohanty, S. P., Longo, A., & Yeun, C. Y. (2025). Shadow AI: Cyber security implications, opportunities and challenges in the unseen frontier. *SN Computer Science*, *6*(5), 405.

RBI (2026), RBI Advisory No 6/2026, "AI Accelerated Cyber Threats and Related Safeguards", April 2026, NON-PUBLIC in nature.