Office of the
Comptroller of the Currency

OCC Supervisory Framework on Bank Use of SUPER INTELLIGENCE

November 2026

Contents

1. Executive Summary **1**
1.1 Overarching Principle 2
1.2 Six Pillars of the Framework 2
2. Framework Components **3**
2.1 Defined Purpose and Permitted Use 3
2.2 Risk Classification and Criticality Assessment 3
2.3 Human-in-the-Loop and Human-on-the-Loop Controls 4
2.4 Governance & Accountability 5
2.5 Model Risk Management Extensions 5
2.6 Data Governance 6
2.7 Privacy 6
2.8 Compliance 7
2.9 Third-Party & Vendor Risk 7
2.10 Operational Resilience 8
2.11 Cybersecurity 9
2.12 Performance Monitoring 9
2.13 Testing 10
2.14 Documentation & Auditability 11
2.15 Internal Audit & Independent Challenge 11
2.16 Reporting & Transparency 12
Summary Table: Component and Illustrative SI-Specific Risk 13
3. Illustrative Controls Framework and SI/Agent Risk Tiering **14**
3.1 Illustrative Risk-Tiering Methodology 14
3.2 Illustrative Risk Tier Definitions 15
Appendix A: Lexicon **16**

1. Executive Summary

This document provides bank examiners with a supervisory framework[1] for evaluating banks' use of Super Intelligence (SI)[2] in normal-course business activities, including, but not limited to, credit decisioning, payments, fraud detection, customer service, operations, risk management and market activities.[3] SI and agentic SI can facilitate banks serving customers faster and more consistently, defend against cyber threats, improve the detection of fraud and financial crimes, and automate manual operations allowing skilled staff to focus on higher value tasks. Used with adequate controls, SI and agentic SI can strengthen risk management, lower costs, and expand access to financial services.

Examiners may use the framework to identify SI-specific considerations that supplement existing model risk management, governance, technology, compliance, and operational-risk practices and to tailor supervisory work to the nature and potential consequences of a bank's SI use cases. Examiners should not treat SI as a separate and distinct risk category, and banks are not required to define a separate SI risk appetite. SI use can introduce, amplify or modify risk in existing risk categories, such as operational, compliance and credit. Those risks, including SI's contribution to them, should be identified, measured, monitored and controlled according to the bank's risk management framework.

The intention is for this examiner framework to be updated regularly. SI and agentic SI are novel technologies which are experiencing rapid evolution. As the technology continues to evolve so will the risks and associated controls. Banks and other market participants continue to deploy SI across new use cases not previously contemplated. This framework should therefore be treated as a living document which is expected to change as technology, use cases, and supervisory understanding change and mature. Further, each bank is different and may present specific risks and issues. Examiners will likely encounter various levels of SI or agentic SI sophistication and should apply judgement, in conjunction with consideration of the bank's size, complexity, and risk profile.

The document presents sixteen framework components. The components describe how SI may affect established control areas and identify illustrative considerations for applying risk-based supervisory judgment. In addition, an illustrative control risk tiering methodology and a technical lexicon are provided to supplement the framework.

---

[1] Model risk management principles described in OCC Bulletin 2026-13 apply to non-generative and non-agentic SI models. As such, the supervisory framework described herein is more applicable to generative and agentic SI.

[2] "Super Intelligence" refers to the technologies and systems previously encompassed by the term "Artificial Intelligence" as defined in 15 USC 9401(3) as well as the promise, potential, and rapidly advancing capabilities of those technologies. See Exec. Order No. 14434, 91 Fed. Reg. 63129 (Oct. 2, 2026).

[3] Consistent with Executive Order 14434, the OCC supports the reasonable adoption of SI and is committed to providing a model of best practices to its supervised institutions. SI tools can contribute to efficiencies in governmental services, including bank supervision, and reduce undue burdens imposed on supervised institutions. Accordingly, this framework was prepared with the assistance of SI. The content has been reviewed, edited, and approved by a human.

1.1 Overarching Principle

A bank's control intensity could increase with the potential consequences of an SI agent's failure or misuse. Higher consequence use cases may warrant greater restrictions on autonomy, stronger controls, more direct human oversight, more independent validation, and more robust resilience measures. Risk classification may be based on potential consequences, transfer of agency, and the degree of human oversight, rather than the technical sophistication of the SI used. For example, a simple rules-based agent authorized to initiate transactions could present greater risk than a sophisticated model limited to internal drafting assistance. Because SI risk can affect multiple risk categories, the framework may help identify transmission channels and link SI risk management to enterprise functions, including risk identification and risk appetite while not inferring any requirement to aggregate SI risk.

1.2 Six Pillars of the Framework

At a high level, the sixteen components are grouped into six pillars. Each pillar identifies a principal supervisory consideration with the components that primarily address it in parentheses:

1. **Purpose:** Defines the actions the SI system or agent is permitted to execute. (Defined Purpose and Permitted Use)
2. **Risk:** Assesses the consequences if the SI system or agent's output or action is not fit for purpose. (Risk Classification and Criticality Assessment)
3. **Control:** Identifies controls that may prevent unauthorized, inappropriate, or unsafe actions of the SI system or agent. (Human-in-the-Loop and Human-on-the-Loop Controls; Model Risk Management; Data Governance; Privacy; Compliance; Cybersecurity)
4. **Oversight:** Establishes who monitors the SI system or agent and who is accountable for its use. (Governance & Accountability; Human-in-the-Loop and Human-on-the-Loop Controls; Performance Monitoring; Reporting & Transparency)
5. **Resilience:** Addresses the bank's ability to detect, contain, recover from, and remediate SI system or agent failures while continuing critical business processes. (Operational Resilience; Cybersecurity)
6. **Assurance:** Provides evidence that SI system or agent controls continue to operate effectively. (Testing; Documentation & Auditability; Third-Party & Vendor Risk; Internal Audit & Independent Challenge)

2. Framework Components

Each component describes how SI could affect the relevant control area, identifies key concepts, and provides illustrative considerations. The framework is intended to support examiners' risk-based judgment and should be considered in a manner proportionate to the nature, consequences, autonomy, and complexity of the bank's SI use cases.

2.1 Defined Purpose and Permitted Use

Effective SI risk management begins with a clearly defined and documented purpose. A purpose statement may establish the authorized scope of an SI system or agent and provide a basis for identifying scope drift, unauthorized use, and control failures.

Key Concepts: Before deployment and after a material change, a bank may document what an SI system or agent is designed to do, what it is not designed to do, who or what may invoke it, the data and systems it is permitted to access, the actions it could take, and the circumstances that may require human review or suspension. The bank's inventory may link these boundaries to ownership, provenance, autonomy, monitoring, and continuity arrangements.

Examples of Purpose and Permitted-Use Considerations

Authorized users and use cases: The bank may identify the people, systems, and business processes permitted to invoke the SI system or agent. For instance, a use case approved for internal research may warrant additional review before it becomes customer-facing or decision-relevant.

Data, systems, and actions: Authorized data sources, connected systems, tools, accounts, and actions may be stated in terms that can be tested rather than as a general capability description.

Decision boundaries and escalation: The purpose statement may distinguish actions the system can take autonomously from actions that call for review, approval, abstention, or escalation.

Inventory and change management: An inventory may capture the use case, owner, provider, permissions, deployment status, monitoring, and fail-over arrangements and may be updated when those characteristics change.

Scope drift: Activity outside the approved purpose may be treated as a control failure even when the resulting output or action appears reasonable.

2.2 Risk Classification and Criticality Assessment

Not every SI use case presents the same risk. A consequence-based assessment may help a bank apply more intensive controls to consequential uses without imposing the same control structure on lower-risk activities.

Key Concepts: A bank may assess each use case before deployment and when material changes occur. Relevant factors could include decision consequence, reversibility, customer and financial impact, autonomy, data sensitivity, explainability, provenance, permission scope, and interdependencies. Classification may be based on the consequences of failure and transfer of agency.

Examples of Risk-Classification Considerations

Decision consequence and reversibility: A system that takes direct action or produces an outcome that is difficult to reverse may warrant more intensive controls than a system that provides nonbinding internal assistance.

Autonomy and permissions: The length of time a system is permitted to operate without a meaningful human checkpoint and the breadth of its data, tool, and account access may materially affect risk.

Customer, financial, compliance, and operational impact: The assessment may consider the magnitude, breadth, and duration of potential harm to customers, the bank, counterparties, markets, or critical operations.

Data, explainability, and provenance: Sensitive data, limited traceability, or opaque third-party models could increase uncertainty and reduce the bank's ability to oversee the use case.

Aggregate exposure: Shared providers, data, permissions, infrastructure, and control weaknesses could create concentration or correlated risk across otherwise separate use cases. A highly autonomous payment agent may present greater SI risk than an SI agent only used for internal drafting.

2.3 Human-in-the-Loop and Human-on-the-Loop Controls

Human oversight is most effective when the bank defines the nature of the involvement and can demonstrate that the control operates as intended.

Key Concepts: Human-in-the-loop generally refers to review and approval before an output or action is used or executed. Human-on-the-loop generally refers to active monitoring with authority and practical ability to intervene. The level of involvement may be calibrated to the potential consequences, autonomy, reversibility, and operating volume of the use case, among other factors.

Examples of Human-Oversight Considerations

Competency and authority: The assigned reviewer may need sufficient expertise, information, and authority to evaluate, reject, modify, pause, or halt the system's output or action.

Meaningful review: Review triggers, evidence, and procedures may be defined so that approval is more than a nominal click. High volumes or bulk approval could reduce review quality and contribute to alert fatigue.

Span of control and monitoring: A bank may evaluate how many agents, accounts, processes, alerts, or actions one person can credibly monitor and the time available to intervene.

Override, logging, and escalation: Override authority, reasons, patterns, and escalation conditions may be logged and reviewed. A misfire, hallucination, or control breach may justify temporarily increasing human involvement.

Independence: Where appropriate, the reviewer may be separate from the team that developed or owns the system.

2.4 Governance & Accountability

Effective governance establishes accountability for SI risk and supports the operation of the other framework components. Rapid deployment, technical complexity, and distributed development or ownership could make accountability more difficult to identify.

Key Concepts: The board and senior management may establish an SI risk appetite, identify accountable executives and use-case owners, assign responsibilities and receive information on material SI use. Governance may be supported by a complete inventory and approval authority at meaningful lifecycle gates.

Examples of Governance and Accountability Considerations

Risk appetite and autonomy exposure: Risk appetite could address autonomy, permitted actions, customer and financial impact, legal and operational exposure, and concentration in addition to model-performance measures.

Approval authority: Roles and decision rights may be defined for development or procurement, risk classification, testing, deployment, material change, exception approval, suspension, and retirement.

Reviewer fluency and information access: Board members, senior management, independent risk management, validation, and audit may need technical knowledge and access proportionate to the bank's use cases.

Third-party ownership: Although development or hosting could be distributed across vendors, accountability for managing the bank's risk generally remains with the bank.

Reporting and challenge: Board and Management reporting may present material use cases, changes, exceptions, incidents, customer harm, concentrations, and remediation in a form that supports timely challenge and action.

2.5 Model Risk Management Extensions

Some SI systems may not conform to more traditional assumptions of model stability, reproducible outputs, a single output type, or directly traceable decision logic.

Key Concepts: Certain Model Risk Management (MRM) principles may be instructive for development, use, validation, governance, and ongoing monitoring. However, the scope, frequency, and independence of those activities may be adapted to the use case's potential consequences, rate of change, non-determinism, and reliance on third-party or continuously updated models.

Examples of SI-Specific MRM Considerations

Continuous change and drift: Adaptive models, frequent retraining, prompt changes, or vendor updates could alter behavior between periodic reviews and may justify more frequent monitoring or re-validation.

Explainability limits: Explainability tools may help identify factors associated with an output, but they do not reveal the SI's reasoning and provide a verified source. Examiners may evaluate whether the method and resulting explanation are appropriate for the model, use case, and applicable obligations.

Non-determinism and reproducibility: Repeated testing, defined samples, distributional measures, and consistency thresholds may provide more useful evidence than a single successful run.

Comparative testing: Traditional champion-challenger methods may not fit open-ended generative outputs. Alternative benchmarks, reference sets, expert review, or controlled comparisons may be considered.

2.6 Data Governance

The reliability of an SI system depends on the quality, lineage, representativeness, and permitted use of the data used to train, fine-tune, and operate it. Large data volumes, third-party sources, and synthetic data could make these attributes more difficult to assess.

Key Concepts: Data governance may address data quality, lineage, representativeness, access, retention, and use throughout the SI lifecycle. The bank may consider both data provided to the system and new data created through prompts, outputs, embeddings, logs, feedback, or synthetic generation.

Examples of SI Data-Governance Considerations

Data lineage: The bank may assess the origin, licensing, transformations, and limitations of training, fine-tuning, source, and production data, including data embedded in third-party foundation models.

Drift and representativeness: Monitoring may identify changes in input populations or data distributions before they materially degrade outcomes.

Synthetic data: Synthetic data may be subject to quality, approval, representativeness, and bias review because deficiencies can be difficult to identify and may be amplified through training.

Permitted sources and access: Technical controls may limit the system to approved data sources and prevent access to information outside its purpose or authorization.

Data architecture: Governance and data models could evolve as SI produces new data types and dependencies or changes how existing data is created, combined, or retained.

2.7 Privacy

SI can create privacy risks beyond conventional data-handling concerns. Models may retain or reproduce training data, infer sensitive attributes, and use bank or customer information in vendor-hosted environments that are not fully transparent to the bank.

Key Concepts: Privacy considerations may include applicable law, permissible use, data minimization, access, retention, deletion, notice, consent, vendor use, and cross-border processing. These considerations may be mapped to each use case and revisited when the model, provider, data, or purpose changes.

Examples of SI Privacy Considerations

Model memorization and inference: A model might reproduce specific training examples or infer sensitive characteristics that were not expressly provided, creating risks distinct from a conventional data breach.

Prompt and output retention: Prompts, outputs, embeddings, and logs could become repositories of sensitive information and may warrant access controls and risk-based retention periods.

Vendor data use: Contracts and technical configurations may clarify whether bank data can be used to train or improve models made available to other customers.

Cross-border processing: The location of data storage and processing could affect data-residency, transfer, notice, consent, and third-party obligations.

2.8 Compliance

Compliance risk could increase when an SI system produces inaccurate or inconsistent outcomes, is used outside its intended purpose, or is not supported by controls that address applicable laws and regulations.

Key Concepts: SI does not change the bank's underlying legal obligations and has the potential to facilitate a bank's compliance with them. However, SI's complexity, opacity, speed, and capacity to act may make it more difficult to identify how a requirement applies, explain an outcome, detect noncompliance, or prevent use outside approved boundaries. A bank may consider compliance through design, development, procurement, deployment, monitoring, change management, and retirement.

Examples of Applicable Compliance-Related Issues

Bank Secrecy Act and sanctions compliance: Banks may use SI in transaction monitoring, risk assessment, case management, reporting, and sanctions screening. Inaccurate, incomplete, or unrepresentative data, or poorly calibrated thresholds could increase false negatives, false positives, or other compliance risks.

Consumer compliance: SI used in customer service, marketing, servicing, collections, disclosures, or other consumer-facing activities could create consumer compliance risk when executing tasks, especially if outputs are materially inaccurate, misleading, or inconsistent with product terms. Banks may consider whether data, model design, testing, monitoring, and human overrides support compliance with the consumer rules and regulations.

Hallucinations: Hallucinations concerning fees, rates, account status, or policies could contribute to legal or customer-harm exposure. Banks may consider controls that limit automated representations, require verification in higher-impact contexts, and support prompt correction and remediation.

2.9 Third-Party & Vendor Risk

Banks frequently obtain SI capabilities from third parties rather than develop them internally. This reliance has the potential to introduce model, operational, data, legal, concentration, and resilience risks that are not fully addressed by conventional service-level and information-security reviews.

Key Concepts: Third-party risk management for SI systems and agentic SI may supplement existing interagency third-party risk management guidance. Additional considerations should be given to the examples below where the provider's SI can affect the bank's decisions, customers or data.

Examples of SI Third-Party and Vendor-Risk Considerations

Provider agent visibility: Banks may define the scope of authority, identity, permitted tool use, data use and escalation limits for provider agents that interact with bank systems and data. The bank may also monitor and potentially limit what provider agents do without impacting the relationship with the provider.

SI incidents: The bank may choose to define a specific SI incident vocabulary and notification timelines to capture events such as hallucination, misfire, drift or manipulation.

Liability: The bank should consider how contracts assign responsibility for model related errors or losses including customer harm for which the bank may seek clarity, indemnification, caps or loss responsibility.

Change and exit: Existing vendors may add SI features, functionality or change models connected to products already in use by the bank. The bank may require notice and transparency of new SI functionality and may reassess the relationship or retest the model before or when the changes are made.

2.10 Operational Resilience

Operational resilience for SI includes both loss of service and incorrect operation. An agentic system could remain available while taking inappropriate or harmful actions, requiring the bank to detect, contain, recover, remediate harm, and continue the underlying business process.

Key Concepts: For more consequential or autonomous use cases, a bank may document and test detection, containment, suspension, continuity, recovery, escalation, and remediation arrangements. The approach may address internal SI systems, fallback processes, and external SI agents that interact through bank channels.

Examples of SI Operational-Resilience Considerations

Misfire detection and containment: Monitoring, observability tools, control agents, deterministic rules, circuit breakers, rate limits, least-privilege permissions, and kill switches could help identify and contain an agent that is operating but acting incorrectly.

Business continuity and fail-over: The bank may test whether the business process can continue through trained personnel, a narrower fallback agent, or another approved process while the primary system is restricted or unavailable.

Recovery and post-incident review: Rollback, data correction, restoration of decision support, root-cause analysis, model-risk review, and controlled return to service could be incorporated into recovery exercises.

Customer remediation: When a misfire affects customers, the bank may establish a separate workstream to identify the complete affected population, assess individual harm, provide notice or correction, and independently verify completion.

External SI agent interactions: Customer, counterparty, fintech, or agentic-commerce systems that the bank does not control could present channel risk. Identity, provenance, delegation, authentication, monitoring, velocity limits, dispute handling, suspension, and revocation may be calibrated to the action requested.

Stress and systemic effects: High-speed agent activity could affect liquidity, fraud losses, service availability, or market behavior. Scenario analysis may consider correlated or cascading activity across customers, agents, and providers.

2.11 Cybersecurity

SI systems may be targets of cyberattacks and tools that increase the speed, scale, or sophistication of attacks. Prompts, model inputs, training and fine-tuning pipelines, connected tools, and non-human identities create attack surfaces that may not be fully addressed by controls designed for conventional code and network perimeters.

Key Concepts: Cybersecurity controls may extend to the SI lifecycle and supporting infrastructure, including identity and access management, data and model integrity, secure development, monitoring, incident response, and recovery. The breadth of agent permissions and tool access may be treated as part of the attack surface.

Examples of SI Cybersecurity Considerations

Prompt injection and jailbreaking: Malicious or misleading inputs can alter system behavior, bypass restrictions, or cause an agent to misuse connected tools without a conventional network breach.

Data and model poisoning: Compromised training, fine-tuning, verified source, or feedback data could embed undesirable behavior that is difficult to detect after deployment.

Model extraction and data leakage: Repeated queries or crafted inputs could be used to infer model behavior, recover sensitive training information, or expose confidential data.

Tool use and excessive agency: Broad credentials or permissions could allow a compromised or manipulated agent to cause fraud, data exposure, or operational disruption at machine speed.

Non-human identities and permissible use: Agent identities, credentials, logs, and data use may be governed in a manner comparable to privileged users and aligned with acceptable-use, privacy, and incident-reporting practices.

2.12 Performance Monitoring

Pre-deployment testing evaluates performance in a controlled environment. Ongoing monitoring evaluates production behavior as data, usage patterns, prompts, permissions, providers, and underlying models change after deployment.

Key Concepts: Metrics, thresholds, reporting, and escalation may be calibrated to the use case's potential consequences, autonomy, expected variability, and rate of change. Monitoring and in-line real time observability may address both average performance and low-frequency events that could cause material harm.

Examples of SI Performance-Monitoring Considerations

Accuracy and output quality: Metrics may include accuracy, precision and recall, task completion, hallucination, abstention, complaint, override, exception, and remediation rates, as applicable to the use case.

Population and input drift: Changes in the population or input distribution could provide an early indication that performance is degrading before outcome metrics show material deterioration.

Behavioral consistency: For non-deterministic systems, repeated or similar inputs may be sampled to assess whether output variability remains within accepted ranges.

Vendor and model changes: Monitoring may identify upstream version changes, prompt or configuration changes, fine-tuning, data changes, or new tools and permissions that alter behavior without a traditional model release.

Feedback loops and escalation: The bank may monitor whether model outputs influence future training data and may define thresholds that trigger review, re-validation, reduced autonomy, or suspension.

2.13 Testing

Testing is central to managing risk from the use of SI. The potential lack of a verified source, non-determinism, continual redevelopment, and high volume of inputs and outputs may require methods that go beyond confirming a single expected result. Testing can contribute to system development, use-case approval, independent validation, and ongoing monitoring.

Key Concepts: The appropriate range, rigor, frequency, and independence of testing may depend on the materiality of the use case, the potential consequences of failure, and the limitations of the SI system. Lower materiality uses may warrant limited testing, while consequential or highly autonomous uses may warrant broader testing of both performance and the design and operation of controls throughout the SI lifecycle.

Testing Across the SI Lifecycle

Design and development: Testing could evaluate requirements, data quality, architecture, permissions, known limitations, expected operating conditions, and whether proposed controls address identified risks before the use case is finalized.

Pre-deployment: Functional, comparative, scenario, stress, adversarial, compliance, resilience, human-oversight, and acceptance testing could help determine whether the system is fit for its approved purpose and operating environment.

Deployment and use: Production sampling, monitoring, incident analysis, change testing, and re-validation could identify drift, new failure modes, or control weaknesses that were not apparent before launch.

Examples of SI-Specific Testing Approaches

Subjective reviews: When an SI system produces outputs without a corresponding verified source, expert review can help assess quality, completeness, appropriateness, and consistency.[4] Reviewer qualifications, procedures, sampling, response rater consistency, and the number of observations may be calibrated to the undesirable-outcome rate the bank is attempting to detect.

Adversarial testing: Misleading, malicious, or edge-case inputs may be used to expose hidden vulnerabilities, unsafe behavior, boundary failures, or weaknesses in prompt, data, and tool controls.

Red teaming: Red teaming could extend adversarial testing by evaluating whether the bank can detect, contain, respond to, and recover from realistic attacks or misuse. Independent or external expertise may be useful when internal capability is limited.

[4] Some SI systems may produce outputs that can be compared objectively, such as a default prediction. In those circumstances, traditional testing methods and performance metrics may be appropriate.

LLM-as-a-judge: One model may be used to compare outputs, score them against a reference answer, or support sampling. The bank may consider the judge model's validation, independence, position or verbosity bias, correlated failure modes, and the need for human review.

Hallucination Detection and Mitigation

Illustrative approaches: Generative SI could produce fluent but fabricated output. Testing could measure hallucination under representative and adversarial prompts, assess verified sources, evaluate confidence and abstention behavior, define categories requiring human verification, and sample live outputs after deployment. The bank may establish responses for outputs that cannot be verified and may reassess those responses when prompts, context windows, data, or model versions change.

2.14 Documentation & Auditability

Documentation enables individuals other than the system's developers to understand, evaluate, and audit its design and operation. SI could complicate documentation because outputs are probabilistic, decision paths may not be directly traceable, and relatively small changes may materially alter behavior.

Key Concepts: Documentation may be proportionate to the use case and may address purpose, ownership, design, data, providers, versions, prompts, permissions, assumptions, testing, validation, monitoring, decisions, approvals, incidents, and changes. Records may support reconstruction of material outputs or actions even when the model cannot produce a rule-based rationale.

Examples of SI Documentation and Auditability Considerations

Decision traceability: Inputs, outputs, context, retrieved sources, tool calls, approvals, overrides, and actions may provide evidence needed to reconstruct a material decision or event.

Version and change control: Model, prompt, fine-tuning, data, configuration, provider, tool, and permission changes may be logged and linked to testing and approval evidence.

Retention and access: Records may be retained for periods appropriate to legal, supervisory, audit, customer-remediation, and operational needs while protecting sensitive information.

Proportionality: Documentation may be scaled to complexity and consequence without creating incentives to select a less suitable system solely because it is easier to document.

2.15 Internal Audit & Independent Challenge

Validation assesses whether an SI system operates as designed. Independent internal audit could provide additional assurance regarding governance, validation, control design, implementation, and the operation of the broader SI risk-management framework.

Key Concepts: Audit scope, cadence, skills, access, and reporting may be established through the audit framework and scaled to the bank's SI use cases. Effective challenge may depend on independence from both the group that develops or owns the system and the group that validates or monitors it.

Examples of Internal-Audit and Independent-Challenge Considerations

Independence: Reporting lines, staffing, compensation, and prior involvement may be considered when evaluating whether audit can challenge the groups responsible for SI system development and ownership and SI validation and monitoring objectively.

Access: Audit may need access to model and system documentation, data lineage, validation results, vendor contracts, monitoring, incidents, overrides, remediation, and underlying evidence rather than only management summaries.

Cadence and coverage: More consequential, rapidly changing, or highly autonomous use cases may warrant more frequent or substantive coverage than lower risk uses.

Control agents and backstops: When a control agent, fallback agent, circuit breaker, or human process serves as a backstop, audit may assess whether the backstop is independently governed, tested, and effective.

Reporting and closure: Findings may be reported to the audit committee or board, and closure may be supported by independent evidence that the underlying condition has been corrected.

2.16 Reporting & Transparency

Effective oversight depends on timely and sufficiently detailed information about a bank's material SI use cases, changes, incidents, failures, and customer harm. Reporting practices may need to keep pace with rapid changes in systems, providers, permissions, and operating conditions.

Key Concepts: Internal reporting may be risk based and may include material use cases, accountable owners, significant changes, control exceptions, incidents, customer remediation, concentrations, and assurance results. Various notifications required by laws and regulations, as well as contracts, could be applicable to SI use cases.

Examples of Reporting and Transparency Considerations

Incident thresholds: The bank may define internal escalation and notification thresholds based on severity, duration, customer impact, financial exposure, legal requirements, and potential systemic effects.

Use-case and change reporting: Reviewing a bank's reporting may help supervisors understand consequential or highly autonomous use cases, material changes in purpose or permissions, and reliance on significant providers.

Information quality: Reports may be reconciled to the SI inventory, incident records, validation results, audit findings, and board reporting so that material differences are identified.

Summary Table: Component and Illustrative SI-Specific Risk

Ref #	Component	Illustrative SI-Specific Risk
2.1	**Defined Purpose and Permitted Use**	Scope drift or unauthorized action without an explicit, testable purpose boundary
2.2	**Risk Classification and Criticality Assessment**	Control intensity that does not reflect consequence, autonomy, reversibility, or aggregate exposure
2.3	**Human-in-the-Loop and Human-on-the-Loop Controls**	A named human reviewer without practical authority, competency, capacity, or a meaningful review standard
2.4	**Governance & Accountability**	Diffuse ownership or insufficient technical fluency limiting effective oversight and challenge
2.5	**Model Risk Management Extensions**	Point-in-time validation that does not address non-determinism, continuous change, or new output types
2.6	**Data Governance**	Unclear lineage, representativeness, permissions, or quality in third-party, synthetic, or rapidly changing data
2.7	**Privacy**	Memorization, inference, retention, or vendor use of sensitive information beyond approved purposes
2.8	**Compliance**	SI design or use that makes legal requirements, decision reasons, or customer-facing representations difficult to control or explain
2.9	**Third-Party & Vendor Risk**	Opaque provider chains, weak contractual rights, switching constraints, or concentration in common models and infrastructure
2.10	**Operational Resilience**	A system that remains available while acting incorrectly, with no tested containment, continuity, recovery, or remediation path
2.11	**Cybersecurity**	Prompt, data, model, tool, or identity attacks that bypass controls designed for conventional systems
2.12	**Performance Monitoring**	Production behavior drifting from the validated baseline without timely detection or escalation
2.13	**Testing**	Testing that confirms expected behavior but does not search for unknown, adversarial, or hallucination-related failure modes
2.14	**Documentation & Auditability**	Inability to reconstruct material decisions, changes, tool calls, approvals, or system behavior
2.15	**Internal Audit & Independent Challenge**	Insufficiently independent or technically capable assurance over development / ownership and validation / monitoring activities
2.16	**Reporting & Transparency**	Material use cases, changes, incidents, or customer harm not reported completely or timely

3. Illustrative Controls Framework and SI/Agent Risk Tiering

This section describes how controls may be scaled; a practical risk-tiering methodology, applied separately to SI models and to agentic systems as risk increases. The goal is a graduated control model: light-touch for low-risk, low-autonomy use, materially heavier for anything approaching customer-impacting, irreversible, or highly autonomous action.

3.1 Illustrative Risk-Tiering Methodology

Every SI and agentic use case may be scored against a common set of risk factors, independent of business line, so that a comparable risk gets a comparable control regardless of where it originates in the bank.

Decision consequence: Does the system inform a human decision, make a decision directly, or take an action directly (e.g., moving funds, updating a customer record, market interaction)? Direct action carries materially higher risk than decision support.

Reversibility: Can an erroneous output or action be caught and undone before harm occurs, or is it effectively irreversible once executed (e.g., a payment sent, a disclosure made, an account closed)?

Customer/financial impact: Magnitude of potential harm to a customer, the markets, or the bank if the system is wrong/makes an error (e.g., a single transaction versus a portfolio-level decision, a low-dollar versus high-dollar threshold).

Autonomy level: Degree of human involvement in the loop, human-in-the-loop (approves each action), human-on-the-loop (monitors and can intervene), or fully autonomous (acts without real-time human checkpoint).

Data sensitivity: Whether the system touches regulated, confidential, or personally identifiable data, and whether inputs or outputs could constitute a privacy or consumer law exposure.

Explainability: Whether the system's output can be traced to a rationale that is sufficient to meet legal, regulatory, or customer-disclosure obligations (e.g., adverse action notices).

Provenance: Whether the system is built in-house, fine-tuned from a third-party base model, or consumed as a fully external vendor service; each origin carries a different visibility and control profile.

These factors combine into a “composite tier,” illustrated below. As an example, a simple internal productivity tool with no customer or financial data exposure sits at Tier 4, whereas a fully autonomous, customer-facing payment agent sits at Tier 1, regardless of how well it performs in testing.

3.2 Illustrative Risk Tier Definitions

Illustrative Composite Tier	Illustrative Profile Characteristics	Illustrative Use Case
Tier 1: Critical	Direct action, low/no reversibility, high financial or customer impact, autonomous or partially autonomous	Autonomous payment agent; real-time credit decisioning; fraud-hold release; market access
Tier 2: High	Direct action or high-weight decision support, human-on-the-loop, moderate-to-high impact	SI-assisted underwriting; agent-initiated customer communications; AML alert disposition; code assistance for internal code or data access changes
Tier 3: Moderate	Decision support with mandatory human-in-the-loop, limited data sensitivity, reversible	Internal document summarization for credit memos; agent-assisted research
Tier 4: Low	No customer/financial data exposure, no direct action, fully human-controlled use	Internal productivity tools; code assistance; general employee Q&A

Appendix A: Lexicon

This should be treated as a living Lexicon which is expected to change as technology, use cases and supervisory understanding change and mature. Definitions for items marked with an asterisk [*] are from the AI Lexicon workstream of the Artificial Intelligence Executive Oversight Group (AIEOG). For consistency with this framework, references to "artificial intelligence" and "AI" in the source definitions have been changed to "Super Intelligence" and "SI."

Adversarial SI:* Techniques and attacks used to manipulate SI systems, causing them to make incorrect or unintended predictions or decisions. These techniques exploit vulnerabilities in SI models, often by subtly altering input data, training data, or model interactions to manipulate the SI system.

Agentic SI:* A category of SI systems capable of independently making decisions, interacting with their environment, and optimizing processes without direct human intervention.

Benchmarking:* An alternative prediction or approach used to compare a model's inputs and outputs with estimates from alternative internal or external data or models.

Bias:* A systematic distortion, as opposed to random error, that reduces the representativeness or accuracy of an SI system's outputs or performance for its intended purposes and operating conditions. Bias may be introduced inadvertently or purposely and may also emerge as the SI system is used in an application; this could arise when the data used to develop or operate the system are not representative of the intended population or operating conditions. Common sources/subcategories or bias include statistical/computational, systemic, and human bias (not exhaustive).

Black box:* The nature of some SI techniques whereby the inferential operations are complex, hidden, or otherwise opaque to their developers and end users in terms of providing an understanding of how classifications, recommendations, or actions are generated and what overall performance will be

Data lineage:* The history of the processing of a data element, which could include point-to-point data flows and the actions performed on the data element.

Data poisoning:* An attack that corrupts or contaminates training data to compromise an SI system's performance.

Data quality/validity:* The usefulness, accuracy, and correctness of data for its application.

Deterministic (algorithm/model):* An algorithm/model that, given the same inputs, always produces the same outputs.

Documentation:* The collection of records that describe an SI system's purpose and intended use, key design choices, training and operational data characteristics and provenance, testing and evaluation results, limitations, and version history, maintained to support transparency, oversight, and accountability across the SI lifecycle**.**

Explainability:* Property of an SI system that enables a given human audience to comprehend the reasons for the system's behavior; the ability to understand an SI system's output and decision given certain inputs.

Foundation models:* Large machine learning models trained on vast amounts of raw and unlabeled data through unsupervised learning that can be adapted and applied to versatile

downstream tasks. Large language models are common subsets of foundation models and underpin many generative SI applications in the financial sector.

Generative SI:* The class of SI that emulates the structure and characteristics of input data to generate derived synthetic content, including images, video, audio, text, and other digital content.

Guardrails:* Layered safeguards to prevent access to bad information and behavior in an SI system. These may encompass policies, technical controls, and monitoring mechanisms, and could exist at the data, model, application, and infrastructure levels. These safeguards aim to ensure generative SI systems behave ethically, safely, and within organizational or regulatory boundaries by filtering training data, aligning model behavior, and enforcing post-deployment controls.

Hallucination:* A phenomenon when SI produces output that is erroneous or flawed but is still in the form of a convincing narrative or presentation. Generative SI can still produce flawed information even if underlying data is free of defects**.**

Human-in-the-Loop (HITL):* A risk-control approach for SI where a human is integrated within the SI's decision-making process.

Human-on-the-Loop (HOTL): An oversight and control model in which an SI Agent or system is operating independently and a qualified human is overseeing, overriding and/or halting actions based on predefined parameters.

Interpretability:* Transparency into the inner workings of SI output in the context of their designed functional purposes, which helps users gain deeper insights into the functionality and trustworthiness of the system and its outputs.

Large language model:* A subset of machine learning that uses algorithms trained on large amounts of data through self-supervised machine learning to recognize patterns and respond to user requests in natural language.

Machine learning:* An SI learning method that enables computational systems to learn patterns, make predictions, and optimize decisions from large amounts of data without being explicitly programmed for each task. Machine learning encompasses supervised, unsupervised, and reinforcement learning paradigms, serving as the technical foundation for data-driven intelligence and automation.

Model integrity:* The process of protecting a model against improper information modification or destruction and ensuring information non-repudiation and authenticity.

Model risk:* The potential for adverse consequences from decisions based on incorrect or misused model outputs and reports. Model risk can be from individual models and be in the aggregate. Aggregate model risk is affected by interaction and dependencies among models; reliance on common assumptions, data, or methodologies; and any other factors that could adversely affect several models and their outputs.

Output validation: A systematic process for verifying and confirming that SI-system outputs meet specified requirements, accuracy standards, and quality criteria before downstream use.

Override:* An output or input that is ignored, altered, rejected, or reversed.

Performance monitoring:* Ongoing activities that confirm an SI system is implemented appropriately, used as intended, and continues to perform as intended over time.

Performance threshold:* A particular value or range of values of a performance measure or diagnostic that determines acceptance or rejection of a model's performance.

Prompt:* Natural-language text describing the task that an SI system should perform.

Prompt injection:* An attack on an SI system that exploits how an application combines untrusted input with a prompt written by a higher-trust party, such as the application designer, so the system follows the untrusted instructions.

Retrieval augmented generation (RAG):* A type of generative SI system in which a model is paired with a separate information retrieval system (or "knowledge base"). Based on a user query, the RAG system identifies relevant information within the knowledge base and provides it to the generative SI model in context for the model to use in formulating its response. RAG systems allow the internal knowledge of a generative SI model to be modified without the need for retraining.

Service level agreement (SLA):* Contractually binding terms, often incorporated into a broader services contract, between a service provider and a customer that specify the services to be delivered and the measurable performance and service-quality commitments, such as availability and response times. SLAs also typically define each party's responsibilities and provisions for monitoring/reporting, issue resolution, and remedies if service levels are not met.

Service Provider Concentration (Financial Institution):* The extent to which a financial institution relies on a service provider, directly or indirectly, to support the financial institution's activities, particularly critical activities.

Service Provider Concentration (Financial Sector):* The extent to which financial institutions rely on a service provider, directly or indirectly, to support financial institutions' activities, particularly critical activities.

Service Provider Concentration Risk (Financial Institution):* The potential for disruption or degradation at a service provider(s) to threaten the ability of a financial institution to continue performing the financial institution's activities, particularly critical activities, or cause the financial institution to suffer significant adverse effects.

Service Provider Concentration Risk (Financial Sector):* The potential for disruption or degradation at a service provider(s) to threaten the ability of financial institutions to continue performing their activities, particularly critical activities, or cause the financial institutions to suffer significant adverse effects, with the potential for systemic impact to the financial sector.

SI Agent:* A system that autonomously perceives its environment, decides what to do, and takes actions to achieve its goals.

SI drift/decay:* The tendency for an SI model's performance to degrade over time when deployed in a real-world setting with differing conditions from those present in training and testing.

SI governance:* The set of organizational policies, rules, frameworks, roles, and oversight processes that direct how SI is adopted, developed, deployed, and monitored within the organization, with the objective of ensuring SI-related risks are identified, managed, and monitored across the SI lifecycle.

SI lifecycle:* The set of phases an SI system goes through. These are plan and design, collect and process data, build and use model, verify and validate, deploy and use, and operate and monitor. These phases are often iterative, and not necessarily sequential.

SI model:* A component of an information system that implements SI technology and uses computational, statistical, or machine-learning techniques to produce outputs from a given set of inputs.

SI model/system exploitation: Adversarial actions that exploit vulnerabilities in an SI model or system to force misperformance against intended objectives, disrupt access to outputs or functionality, or enable unauthorized access to restricted or proprietary information.

SI risk assessment:* A risk-management process for identifying, estimating, and prioritizing risks arising from the operation and use of an SI system, incorporating threat and vulnerability analyses and considering mitigations provided by controls planned or in place**.**

SI system:* The term 'super intelligence system'

(A) means any data system, software, application, tool, or utility that operates in whole or in part using dynamic or static machine learning algorithms or other forms of super intelligence, whether

(i) the data system, software, application, tool, or utility is established primarily for the purpose of researching, developing, or implementing super intelligence technology; or

(ii) super intelligence capability is integrated into another system or agency business process, operational activity, or technology system; and

(B) does not include any common commercial product within which super intelligence is embedded, such as a word processor or map navigation system.

SI use case:* A specific scenario in which SI is designed, developed, procured, or used to achieve a particular objective, such as delivering a product or service, enhancing decision-making, or providing a defined benefit.

SI use case inventory:* A maintained repository or listing of an organization's SI use cases, intended to support governance, transparency, and risk management by documenting where and how SI is designed, developed, procured, or used, and the purpose and outputs associated with those uses.

Super Intelligence (SI): A machine-based system that, for a given set of human-defined objectives, can make predictions, recommendations, or decisions that influence real or virtual environments. SI systems use machine- and human-based inputs to perceive environments, abstract those perceptions into models through automated analysis, and use model inference to formulate options for information or action.

Synthetic data:* Data that has been generated using a purpose built mathematical model or algorithm, that is statistically realistic but artificial, that can be used for activities like model development and training.

Text/word embedding:* A numerical vector representation of text that machine learning and super intelligence systems use to work with meaning in text, such as comparing similarity between pieces of text.

Third-party SI risk:* Risk arising when an organization relies on another entity to develop, provide, host, operate, or support SI systems or key SI components such as models, data, and related infrastructure.

Training data:* A subset of input-data samples used to train a machine-learning model.

Validation:* Confirmation, through objective evidence, that an SI system or model meets requirements for a specific intended use or application and achieves its intended use in its intended operational environment.

Version control:* Systematic practice of tracking, managing, and documenting changes to SI assets throughout their development and deployment lifecycle.