BOARD OF GOVERNORS OF THE FEDERAL RESERVE SYSTEM

WASHINGTON, D.C. 20551

DIVISION OF SUPERVISION AND REGULATION

SR 26-XX June 22, 2026

TO THE OFFICERS IN CHARGE OF SUPERVISION AND TO EACH BANKING ORGANIZATION SUPERVISED BY THE FEDERAL RESERVE

SUBJECT: Action-Level Assurance for Autonomous AI Agents (LAAS): Supervisory Expectations for Governing Agent Actions in Regulated Institutions

Applicability: *This guidance applies to supervised institutions that deploy LLM-based autonomous agents capable of committing actions with effects outside the agent's own sandbox.*

Purpose

This guidance establishes supervisory expectations for the governance of individual actions taken by large-language-model (LLM) based autonomous agents at supervised institutions. It applies to any supervised banking organization that deploys an autonomous agent capable of committing an action with an effect outside the agent's own sandbox, including record writes, fund transfers, access-permission changes, and external communications. The guidance adopts the LLM-Agent Assurance Standard (LAAS) v1.1 (`standard/LAAS.md`) as the reference technical control set and frames its obligations as supervisory expectations consistent with the model-risk-management principles of SR 11-7 and its successor SR 26-2.

This guidance does not certify a model, a model provider, or a management system. It addresses what an agent may commit, and the evidence an institution must produce that the right checks ran, by the right party, before the action took effect.

Background

Supervisory guidance on model risk management (SR 11-7) was written for models that produce an estimate a human then acts upon. Autonomous agents collapse that separation: the agent both produces the output and commits the action, at machine speed and scale. Traditional governance checkpoints (periodic validation, human review of material decisions, and post-hoc outcome analysis) do not bound the rate at which an undetected error in an agent's output reaches and is committed as a consequential external action.

The supervisory concern is therefore one of **blast radius**: the maximum adverse impact of a single committed action along the axes of reversibility, scope, and consequence. An agent that can move funds, alter access, or write to systems of record can, absent an out-of-process control, commit an irreversible high-consequence action before any human is aware it was contemplated. LAAS addresses this gap by gating each action at a Consequence Tier derived from its observed effect surface and by requiring independent pre-commit verification, proportionate to that tier, before commitment. The machine source of truth for all thresholds referenced below is the LAAS conformance bundle `laas-fin-1.1.0` (`conformance/laas/data.json`), enforced by the OPA policy `conformance/laas/laas.rego`, package `kellerai.laas.actions`.

Supervisory Expectations

Supervised institutions deploying in-scope agents are expected to implement the following controls. Each expectation derives from a normative LAAS obligation; threshold values are not restated here but are sourced from `conformance/laas/data.json` so that examiners and institutions evaluate against a single machine-readable bundle.

1. **Consequence-tier classification of every action.** The institution shall classify every agent action by Consequence Tier (CT0–CT4), derived by an out-of-process gate from the *observed* effect surface, that is, the tuple (reversibility, scope, consequence) of the actual tool invocation, and never from the agent's self-report. Where any axis is undetermined, the action defaults to the highest tier. (LAAS-OBL-TIER-001; tier lattice in `conformance/laas/data.json`.)

2. **No self-grading.** The agent's self-reported tier may be recorded for diagnostic purposes but shall not lower the gate-derived tier. A persistent pattern of under-reporting is itself a governance signal. (LAAS-OBL-SELF-001.)

3. **Cumulative blast-radius aggregation.** The institution shall maintain a windowed aggregate of effect across actions by a given principal, session, or effect class, and re-tier subsequent actions to the cumulative tier. This closes the structuring loophole by which several individually sub-threshold actions compose into a high-consequence aggregate: the agent analogue of transaction structuring. (LAAS-OBL-AGG-001.)

4. **Untrusted-input elevation.** Actions driven by untrusted input (web content, inbound email, or third-party data) shall be elevated to at least the untrusted-input floor tier or blocked, protecting tier derivation and verifier selection from prompt-injection. (LAAS-OBL-INP-001; floor in `conformance/laas/data.json`.)

5. **Independent, qualified pre-commit verification.** For every action at or above the independent-verification floor tier, a passing independent verifier shall evaluate the action *before* commitment. A verdict of fail, abstain, or indeterminate shall block and escalate the action. Post-execution review does not satisfy this expectation. This is the agent-action analogue of effective challenge under SR 11-7. (LAAS-OBL-IRR-001.)

6. **Verifier independence.** The verifier shall not share the actor's model substrate. Independence is satisfied by a deterministic/exact checker, by a distinct model lineage whose measured error-correlation with the actor does not exceed the bundle threshold, or by a human. A verifier that fails the same way as the actor is not a check. (LAAS-OBL-IND-001; correlation ceiling in `conformance/laas/data.json`.)

7. **Verifier qualification.** A verifier gating high-tier actions shall be qualified: documented coverage of its claim class, a negative-test suite of known-bad inputs it must catch, and a change-controlled version recorded in the trace. A verifier modified after its last qualification is treated as unqualified until re-

qualified. This is the validator-qualification analogue under SR 11-7 and DO-330. (LAAS-OBL-VQ-001.)

8. **Bounded residual escape rate.** For open-world (Bucket B) claims, the institution shall declare a maximum acceptable residual escape rate per tier, estimate it by backtesting on a held-out, adversarially-stressed evaluation set with a stated confidence interval, and re-measure on any change to model, prompt, tool, or policy. The measured rate shall not exceed the declared tolerance, which shall not exceed the LAAS minimums. This is the ongoing-monitoring and outcome-analysis analogue under SR 11-7. (LAAS-OBL-RES-001; tolerances by tier in `conformance/laas/data.json`, key `escape_rate_tolerance_by_ct`.)

9. **Human approval at the highest tier.** Every action at the human-approval floor tier shall obtain explicit human approval before commitment, unless blocked, with abstention the default on timeout. Institutions may register standing pre-authorized envelopes with explicit limits so that human attention scales with exceptions rather than volume. (LAAS-OBL-HUM-001; floor in `conformance/laas/data.json`.)

10. **Vendor accountability.** Where an action is produced or informed by a vendor model or third-party tool, the institution shall record vendor provenance and confirm the vendor's scope is limited to the authorized envelope; errors attributable to the vendor count against the institution's own escape-rate budget. The deploying institution owns the risk of models it buys. (LAAS-OBL-VEN-001.)

11. **Enforcement-plane integrity.** The gate shall run out-of-process in a trust boundary the agent cannot disable, bypass, or modify, even under elevated permissions; the policy bundle shall be cryptographically signed and version-pinned; and the decision-trace sink shall be append-only. A control the constrained party can switch off is not a control. (LAAS-OBL-ENF-001.)

12. **Decision-trace evidence.** The gate shall emit, for every gated action, an append-only, hash-chained decision-trace record capturing the derived tier, verifier identity and verdict, independence and qualification basis, residual bound, and approval or block. Per-actor hash chains are periodically anchored to a shared Merkle root. Personally identifiable information and material non-public information shall be tokenized and the searchable index separated from sensitive payloads. The signed, sampled trace bundle plus an independent auditor attestation is the conformance artifact. (LAAS-OBL-TRC-001.)

Examiner Guidance

In assessing an institution's governance of in-scope agents, examiners should evaluate the following evidence:

- **Tier classification in practice.** Decision-trace records demonstrating that the gate derives the Consequence Tier from the observed effect surface and that the gate-assigned tier is never below the lattice-derived value, including for undetermined inputs (which should default to the highest tier).
- **Verifier independence and qualification.** For high-tier actions, evidence that the verifier is independent of the actor (deterministic, distinct lineage with bounded error-correlation, or human) and qualified, with a current qualification record referenced in the trace and a negative- test suite the verifier demonstrably catches.
- **Trace bundles.** A signed, content-addressed, sampled decision-trace bundle sufficient to support statistical inference about the escape rate at each tier, accompanied by an independent auditor attestation; continuity of the per-actor hash chain and presence of the Merkle anchor.
- **Escape-rate monitoring.** Declared per-tier tolerances at or below the LAAS minimums, a current backtest report produced after the most recent model/prompt/tool/policy change, and evidence of re-measurement on change.
- **Enforcement-plane integrity.** Architecture evidence that the gate is out-of-process and cannot be disabled by the agent, that the policy bundle is signed and version-pinned, and that the trace store rejects updates and deletes on committed records.
- **Cumulative and untrusted-input controls.** Evidence that windowed aggregation re-tiers structured action sequences and that untrusted-input-driven actions are elevated or blocked.

Examiners should treat the absence of an out-of-process gate, the use of agent self-classification to set the operative tier, or the inability to produce trace evidence for committed high-tier actions as material weaknesses in the institution's model-risk and operational-risk controls.

Implementation

Institutions are expected to implement these controls in accordance with the size, complexity, and risk profile of their agent deployments, on a phased basis:

- **Effective date:** [EFFECTIVE DATE PLACEHOLDER]. This guidance is a draft and carries no effective date until issued.
- **Phase 1, enforcement plane and trace.** Stand up the out-of-process gate and the append-only decision trace before adding obligation checks.
- **Phase 2, tier derivation.** Implement effect-surface tier derivation, self-report recording, and cumulative-window aggregation.
- **Phase 3, input and vendor controls.** Add untrusted-input classification and vendor attribution.
- **Phase 4, independent verification.** Qualify verifiers and wire them to the high-tier gate.
- **Phase 5, escape rate and human approval.** Run initial backtests, declare per-tier tolerances, and implement the human-approval path with standing envelopes.

Institutions should be prepared to discuss their implementation roadmap and current state with their supervisory teams.

Supersession / Related Guidance

This guidance supplements and does not supersede SR 11-7 (*Guidance on Model Risk Management*, 2011) or SR 26-2 (its AI/ML successor). It adopts LAAS v1.1 (`standard/LAAS.md`), which supersedes LAAS v1.0, as the reference technical control set. Where this letter and `conformance/laas/data.json` differ on a threshold value, the data file is the machine source of truth. Related supervisory and technical references include SR 11-7, SR 26-2, the NIST AI Risk Management Framework (AI 100-1), and the LAAS design record (`docs/laas/proposal-v1.1.md`).

Distribution

Reserve Banks should distribute this letter to the supervised organizations in their districts and to appropriate supervisory and examination staff. Direct questions regarding this guidance to the Division of Supervision and Regulation.

[Signed]

Director, Division of Supervision and Regulation

Attachment: *LLM-Agent Assurance Standard (LAAS) v1.1*, the technical annex defining the consequence-tier framework, the twelve normative obligations, verifier independence and qualification criteria, residual escape-rate tolerances, enforcement-plane integrity requirements, and decision-trace evidence requirements referenced throughout this letter (`standard/LAAS.md`).