

Orchestra AI

AGENTS DRIVE · HUMANS NAVIGATE

The Controls Catalogue

Orchestra Navigator™ · How Governance Is Implemented and Operated

The Framework establishes what Governable Autonomy is.
The Assessment Guide establishes how it is measured.
This document establishes how governance is implemented and operated.

PREFACE

How to Read This Document

Content here is labeled **STABLE** or **PROVISIONAL**, per the Navigator V1.0 design charter. The control classification mechanism (Hard versus Graduated) and the gate structure are Stable. Specific cadence numbers, threshold values, and archetype edge cases inside the risk-stringency matrix are Provisional, pending pilot validation against real organizations.

I

What This Document Adds

The Framework establishes the five-level ladder and the risk ceiling. The Assessment Guide establishes how an organization's standing is measured. Neither specifies the actual controls an organization implements to operate at a given level, or how those controls are governed once implemented. That is this document's subject: a numbered control catalogue, the rule for how strictly each control applies, the gates that govern advancement between levels, and the standing operating rhythm that keeps governance real between assessments.

II

Control Classification: Hard Controls and Graduated Good Practice

STABLE

Every control in this catalogue is one of two types, and the distinction is deliberate and asymmetric.

Hard controls are binary. They either operate as specified or the organization is not at the level claimed, full stop. No risk class, organization size, or maturity narrative relaxes them. A control is Hard because its failure mode is **silent**: a "mostly enforced" trigger is an unenforced trigger that has not been caught yet.

Graduated good practice controls require the underlying function, but the function's formality, cadence, and tooling scale with risk class and organizational context. A twenty-person team can satisfy a governance-function control with a named accountable owner; a regulated enterprise needs a cross-functional forum. Same function, proportionate form.

Leeway lives in how formally and how often a control is exercised, never in whether it exists.

Controls that are Hard, and why:

CONTROL	WHY IT CANNOT BE GRADUATED
Sandboxed least-privilege execution	A partial sandbox fails silently. Out-of-bounds actions that succeed leave no signal that anything was wrong.
Technical checkpoint enforcement	"Usually reviewed" is the rubber-stamp failure mode in structural form.
Kill switch (tested)	An untested kill switch is a belief, not a control. It is discovered only when needed, the worst possible moment to discover it doesn't work.
Escalation triggers as hard stops	An advisory trigger is not an oversight architecture. It is logging with aspirations.
Tamper-evident authority ledger	Gaps in the accountability record cannot be reconstructed after the fact, by definition.
Full traceability, including non-fired triggers	The only partial evidence against the silent-escalation-failure problem. Degrading it degrades the framework's honesty claim.
Automated stage-downgrade (Level 4)	The load-bearing safety property of Continuous Assurance. Without it, Level 4 is Level 3 with the safety net quietly removed.
The risk ceiling itself	The framework's core promise. Never a good-practice suggestion, and never overridden by an exception.

Every other control in this catalogue, including all cross-cutting governance controls, cadence-based process controls, and policy or training documentation, is Graduated: required in function, proportionate in form.

III

The Control Catalogue

STABLE STRUCTURE PROVISIONAL DETAIL

Controls are numbered NAV- <level> . <n> by the level at which they first apply, and are cumulative upward: everything mandatory at Level N remains mandatory at Level N+1 unless explicitly superseded.

Level 1 controls (Assisted)

ID	CONTROL	TYPE	EXAMPLE IN PRACTICE
NAV-1.1	AI-usage policy: permitted and prohibited contexts, provenance rules	Graduated	A one-page policy stating AI may draft code and documentation but may not approve its own pull requests, access customer PII, or send external communications unsupervised.
NAV-1.2	Provenance tagging of AI-generated contributions, operational and logged	Graduated	Every AI-assisted commit carries a "Co-authored-by: agent" trailer or equivalent metadata, queryable later to identify which code was AI-originated.
NAV-1.3	Reviewer training on AI-output failure modes; sampled anti-rubber-stamp checks	Graduated	Engineers complete training on hallucinated APIs and plausible-but-wrong logic; a manager samples 1 in 20 approved AI-assisted PRs monthly to confirm real engagement, not just a click.

Level 2 controls (Supervised Loops)

ID	CONTROL	TYPE	EXAMPLE IN PRACTICE
NAV-2.1	Documented loop-scoping: task boundaries, permitted actions, checkpoint placement	Graduated	A one-page scope sheet per loop type: "test-generation loop may read source and write test files only; may not modify production configs; checkpoint required before merge."
NAV-2.2	Sandboxed least-privilege execution, verified by test	HARD	The agent runs in a container with filesystem access limited to the repo's /src and /tests directories; a red-team test attempts to write outside that scope and is confirmed blocked in sandbox logs.
NAV-2.3	Technical checkpoint enforcement, no bypass path	HARD	Branch protection rules technically prevent merge to main without human-approved review, verified by attempting a direct push and confirming it is rejected by the platform, not just discouraged by policy.
NAV-2.4	Per-iteration logging sufficient for end-to-end reconstruction	Graduated	Every loop iteration logs the prompt, tool calls made, files touched, and test results, retrievable later to reconstruct exactly what the agent did across a 40-iteration run.

ID	CONTROL	TYPE	EXAMPLE IN PRACTICE
NAV-2.5	Named harness owner; reviewers work to documented acceptance criteria	Graduated	A specific engineer, named in the team wiki, owns the CI harness configuration; checkpoint reviewers use a written checklist rather than an unstructured glance.
NAV-2.6	Harness configuration change control	Graduated	Changes to the harness's permission scope or checkpoint rules go through the same pull-request review as application code, not a quiet config edit.
NAV-2.7	Loop incident procedure, tested at least once	Graduated	A documented runbook exists for "agent loop produced a destructive action," and the team has run a tabletop exercise simulating a loop that deleted a file it shouldn't have.
NAV-2.8	Kill switch, demonstrated	HARD	A single command halts all active agent loops for the system; the team has actually pressed it in a drill and confirmed all in-flight loops stopped within a defined time window.

Level 3 controls (Governed Autonomy)

ID	CONTROL	TYPE	EXAMPLE IN PRACTICE
NAV-3.1	Escalation trigger catalogue, governance-approved, mapped to risk and authority thresholds	Graduated	A published table: "loop requests a new API endpoint, escalate to tech lead," "loop's change touches billing logic, escalate to finance-approved reviewer," each row governance-signed-off.
NAV-3.2	Triggers enforced as hard stops, adversarially verified	HARD	A red-team exercise deliberately phrases a billing-logic change so it doesn't match the trigger's pattern; the test confirms the trigger still fires, or the gap is found and closed before certification.
NAV-3.3	Named, trained, authorized resolver per trigger class	Hard (authority) / Graduated (training)	The "billing logic" trigger routes to a specific named finance-approved engineer, not "whoever's on call," who has completed scenario-based training on realistic escalations.

ID	CONTROL	TYPE	EXAMPLE IN PRACTICE
NAV-3.4	Tamper-evident authority ledger: who, what, when, basis	HARD	Every escalation resolution is written to an append-only log capturing resolver identity, timestamp, the loop's request, and stated reasoning, in a format that prevents silent edits after the fact.
NAV-3.5	Full traceability including triggers evaluated but not fired	HARD	Logs record not only "trigger X fired at 14:02" but also "triggers Y and Z were evaluated and did not fire," so a later audit can check whether a near-miss trigger was well-calibrated.
NAV-3.6	Trigger catalogue under formal change control with governance sign-off	Graduated	Adding or modifying a trigger requires governance-forum approval and is version-controlled with a changelog, the same discipline applied to the triggers as to the code they govern.

ID	CONTROL	TYPE	EXAMPLE IN PRACTICE
NAV-3.7	Scheduled near-miss and blocked-escalation review cadence	Graduated	A monthly meeting reviews every escalation from the past period, asking "was this the right call, and did any near-miss suggest a trigger is miscalibrated."
NAV-3.8	Adversarial trigger-evasion testing on a defined schedule, findings remediaded	Graduated	Each quarter, a red team attempts to reach a triggering outcome through a path matching no existing trigger; any successful evasion is written up and the catalogue is updated.
NAV-3.9	Escalation-fatigue metrics with thresholds	Graduated	Dashboards track each resolver's auto-approve rate and time-to-decision; a resolver approving 95% of escalations in under ten seconds triggers a review of real engagement.
NAV-3.10	Automated rollback for reversible actions; kill switch tested under load	Hard (kill switch) / Graduated (rollback)	A loop that made an unwanted config change is automatically reverted via a one-command rollback, tested under simulated high load, not just a quiet test environment.

Level 4 controls (Continuous Assurance), design specification

ID	CONTROL	TYPE	EXAMPLE IN PRACTICE
NAV-4.1	Upstream objective and boundary approval before a loop runs	Graduated	Before a fully autonomous release-management loop is switched on, its objectives ("keep deployment failure rate below X, never touch payment infrastructure") are formally approved by governance.
NAV-4.2	Statistically defensible periodic audit: sampling frame, confidence levels, documented	Graduated	A quarterly audit samples a statistically justified percentage of the loop's decisions, calculated to detect a defined defect rate at a stated confidence level, not "a few we happened to look at."
NAV-4.3	Named objective owners and audit owners with written accountability acceptance	Graduated	A specific director signs a document accepting accountability for the loop's objectives; a separate named audit owner signs for the periodic review's rigor.
NAV-4.4	Continuous telemetry with anomaly and drift detection against validated baselines	Graduated	The system continuously compares current loop behavior against a statistically validated baseline, flagging deviations automatically rather than waiting for the next scheduled audit.

ID	CONTROL	TYPE	EXAMPLE IN PRACTICE
NAV-4.5	Automated stage-downgrade to Level 3 behavior on defined signals	HARD	If drift detection crosses a defined threshold, the harness automatically re-imposes per-instance escalation requirements without waiting for a human to notice and decide.
NAV-4.6	Immutable, long-retention audit store, regulatory-reconstruction grade	Graduated	Audit logs are written to storage that cannot be altered after the fact and retained for the period required by the applicable regulatory regime, not the default retention policy.
NAV-4.7	Periodic re-validation of each system's risk classification	Graduated	Every six months, someone actively re-checks whether the system's original risk classification still holds, rather than assuming it's fixed forever.

Cross-cutting governance controls (from Level 2 onward, all risk classes)

ID	CONTROL	TYPE	EXAMPLE IN PRACTICE
NAV-G.1	Standing governance function owning trigger approvals, certifications, exceptions, near-miss receipt	Graduated	A small company: a single named VP of Engineering reviewing trigger changes monthly. A larger regulated enterprise: a standing committee with engineering, risk, and legal, meeting biweekly.

ID	CONTROL	TYPE	EXAMPLE IN PRACTICE
NAV-G.2	Per-system stage certification with a defined recertification cycle	Graduated	Each production AI system has a certification record stating its current level and the date it must be re-evidenced, tracked in a register rather than a one-time sign-off.
NAV-G.3	Time-boxed exception process with a named approver; no standing, undated exceptions	Graduated (risk ceiling remains Hard)	A team needs to temporarily bypass a checkpoint during an incident; a named director approves, it expires automatically in 48 hours, and triggers a retrospective review.
NAV-G.4	Vendor-harness assurance boundary documented: vendor-evidenced vs. organization-evidenced controls	Graduated	A document states "our vendor's SOC 2 report evidences their sandbox isolation; we independently verify checkpoint enforcement and maintain our own authority ledger."
NAV-G.5	Incident and downgrade learnings formally fed back into the trigger catalogue and gate criteria	Graduated	After an incident reveals a trigger gap, the root-cause write-up includes a specific proposed catalogue change, going through the same change-control process as any other update.

IV

Risk-Class Stringency

STABLE CONCEPT PROVISIONAL VALUES

The four risk classes, Low, Moderate, High, and Critical, modulate control stringency and cadence. Hard controls do not flex with risk class; they are binary at every class. Risk class modulates only the form and cadence of Graduated controls.

CONTROL GROUP	LOW	MODERATE	HIGH	CRITICAL
NAV-1.x (Assisted)	Required	Required	Enhanced	Enhanced
NAV-2.x (Supervised Loops)	Required	Required	Enhanced	Enhanced · terminal level
NAV-3.x (Governed Autonomy)	Required	Enhanced	Enhanced, where ceiling permits	Not applicable, ceiling prohibits
NAV-4.x (Continuous Assurance)	Required if operating there	Not applicable	Not applicable	Not applicable
NAV-G.x (cross-cutting)	Required	Required	Enhanced	Enhanced

What "Enhanced" means, concretely:

- **Cadence tightening.** Near-miss review moves from quarterly to monthly. Evasion testing moves to every release cycle. Recertification moves from biennial to annual.
- **Authority elevation.** Resolver roles must be legally accountable individuals where required, not delegates. Exceptions require sign-off one governance level higher.
- **Evidence hardening.** Sampled evidence becomes complete-population evidence for the authority ledger and traceability.
- **Retention extension.** Log and ledger retention aligns to the applicable regulatory horizon rather than internal convenience.

On the Critical column: for Critical-class systems, Level 2 is not a waypoint toward autonomy. It is the destination. The governance objective is excellent checkpoint oversight, and any roadmap implying eventual Level 3 for such a system is itself a finding, not a maturity goal.

V

Stage Gates

- STABLE PRINCIPLES
- PROVISIONAL THRESHOLDS

Progression between levels is evidenced, per system, and reversible. Four principles govern every gate: evidence, not attestation; the minimum rule; gates are bidirectional; and the risk ceiling always binds.

Gate 0 to 1

Reviewer training completed and sampled review quality checks pass. Approved AI-usage policy and review standard in force. Provenance tagging operational. Sign-off: engineering leadership. Downgrade triggers: rubber-stamp rate breaches threshold, or provenance tagging found unreliable on audit.

Gate 1 to 2

Named harness owner in place. Loop-scoping and checkpoint-review competence demonstrated through observed review, not certificate. Loop-scoping procedure, harness change control, and a tested incident procedure documented. Sandbox verified by an attempted out-of-bounds action being blocked and logged. Technical checkpoint enforcement confirmed. Sample-loop reconstruction demonstrated. Kill switch demonstrated. Sign-off: engineering plus risk and governance function. Downgrade triggers: sandbox escape or permission-boundary breach; checkpoint bypass discovered; iteration logs found incomplete on sampling.

Gate 2 to 3, the critical gate

This is where checkpoint oversight is removed and carries the heaviest evidentiary burden. Every trigger class has a named resolver with documented authority mapped to accountability, assessed on realistic callout scenarios. Fatigue thresholds are set before go-live. A complete, governance-approved trigger catalogue exists. The authority-ledger procedure is defined. A near-miss review cadence is committed. Adversarial trigger-evasion testing has actually been performed, with findings remediated, not merely scheduled. Triggers are verified as hard stops by test. The authority ledger is system-captured and tamper-evident. Full traceability exists, including triggers evaluated but not fired. Rollback is demonstrated for reversible action classes. The kill switch is tested under load. A defined operating history at Level 2 for this specific system exists, typically at least a quarter, with checkpoint-review findings analyzed: this history is the evidence base the Level 3 triggers were actually designed from. Skipping it means the triggers are speculative.

Sign-off: governance function plus the accountable business owner; for regulated systems, explicit mapping to applicable oversight obligations. Downgrade triggers: any confirmed production trigger-evasion; a fatigue threshold breach; ledger gaps found on audit; a resolver acting without authority or context; the near-miss review cadence missed twice consecutively.

Gate 3 to 4, forward-looking

Treat this gate as a design specification, not an operational pathway; no organization is currently known to have certified evidence at this gate. Named objective and audit owners with demonstrated statistical competence and written accountability acceptance. An approved sampling design. A drift-response procedure with automatic reversion to Level 3. Extended Level 3 history showing stable evasion-test findings and well-calibrated triggers. Automated stage-downgrade demonstrated in practice, not merely built. Sign-off: board-level or equivalent risk owner. Downgrade triggers: any drift threshold breach; a missed audit cadence; sampling rigor found deficient; any incident whose root cause a periodic audit failed to surface within its own window.

VI

The Governance Operating Model

STABLE

Gates and certifications are events. Governance is the rhythm that runs between them, and it is where most real-world governance failure actually happens, not at the gate itself but in the months after it was passed.

The governance function (NAV-G.1). Proportionate in form: a single named accountable owner for a small organization or a low-risk system, a cross-functional forum for larger or higher-stakes contexts. Owns trigger-catalogue approvals, certification and revocation decisions, exception decisions, and near-miss review receipt. Meets at a cadence set by the highest risk class in its scope.

Recertification (NAV-G.2). Per system, not per organization. Biennial for Low risk, annual for Moderate and High, continuous condition-based for any Critical system at Level 2, where any Hard-control failure triggers immediate review rather than waiting for the scheduled date. Recertification re-evidences the gate criteria; it is not a signature renewal.

Reporting lines. Fatigue metrics, near-miss summaries, and evasion-test results report to the governance function on the recertification cadence. Downgrades and exceptions report to a board-level risk owner within a defined window, not batched into the next scheduled meeting.

Framework evolution (NAV-G.5). Incident and audit learnings are formally assessed against the trigger catalogue and gate criteria, and proposed changes go through the same change control the catalogue itself imposes on everything else. A governance framework that does not govern its own evolution is not actually governing anything durable.

VII

Vendor-Harness Assurance

STABLE PRINCIPLE

PROVISIONAL MECHANICS

Most organizations will run loops on a vendor harness rather than a homegrown one. NAV-G.4 exists because this creates a real assurance-boundary question: who evidences sandbox integrity, trigger enforcement, or kill-switch reliability when the harness is someone else's product?

The organization must document, per vendor harness in use, which controls are evidenced by the vendor, through attestations, certifications, or contractual terms, and which remain the organization's own responsibility to evidence. **Any control that cannot be evidenced by either party caps the Capable Stage for systems running on that harness**, regardless of how sophisticated the vendor's own claims are. A vendor attestation is not, by itself, equivalent to the organization's own adversarial verification.

VIII

What This Document Does Not Yet Resolve

- Specific cadence and threshold values throughout Sections IV to VI are asserted, not empirically derived from engagement data.
- The vendor-harness assurance mechanics need real-engagement development beyond the stated principle.
- Level 4 controls remain entirely a design specification; no client evidence currently exists that any operate defensibly in practice.
- This catalogue does not and cannot resolve the silent-escalation-failure problem named in the Framework whitepaper. Nothing here solves it.

Where to Go Next

Just need the concepts? Start with the Orquestra Navigator™ Framework whitepaper.

Running an assessment? Start with the Navigator Assessment Guide.

Determining mandatory controls for a specific risk class? Start with the Navigator Critical Controls Guide.

Implementing this catalogue right now? Section III is the control list, Section V is the gate sequence, Section VI is the operating rhythm to sustain it.

**This document establishes
how governance is implemented
and operated. What it governs
is defined in the Framework
and measured in the Assessment Guide.**

Rajesh Srinivasan

Orchestra AI · rajesh@orquestraai.io
orquestraai.io

Orchestra Navigator™ Controls Catalogue · Working draft, not yet released.
Trademark application in preparation, not yet registered.
© 2026 Orchestra AI.