

Orchestra AI

AGENTS DRIVE · HUMANS NAVIGATE

The Critical Controls Guide

Orchestra Navigator™ · What Applies, and How Strictly, by Risk Class

The Controls Catalogue establishes every control and its type.
This document establishes exactly what applies, how strictly, at each risk class.

PREFACE

How to Read This Document

Labeled **STABLE** or **PROVISIONAL** throughout. The four risk classes and their determinants are Stable.

The specific cadence numbers, seniority requirements, and evidence thresholds assigned to each class are Provisional, pending pilot validation.

This guide is organized by risk class, not by control, because that is how the question actually arrives in practice: "I have a system at this risk class, what exactly do I need to do," not "here is control NAV-3.7, here are its four variants." The Controls Catalogue remains the source of truth for control definitions; this guide is the lookup table for what they mean at each class.

I

What This Guide Adds

The Controls Catalogue's risk-stringency matrix states that controls are "Required" or "Enhanced" by class, and defines Enhanced in general terms. It does not say what those look like concretely for a Low-risk internal tool versus a Critical safety system, and it does not walk through how a system gets classified into one of the four in the first place. Both are this document's subject.

II

Classifying a System Into a Risk Class

- STABLE DETERMINANTS
- PROVISIONAL THRESHOLDS

Risk class is assessed per system, never at the organizational level, and is determined by five factors, each a concrete question to ask at project charter time.

1. Reversibility. Can a wrong action be undone before harm occurs? **Full:** a draft nothing depends on until approved. **Effortful:** reversible but at real cost. **Limited:** possible in principle, rarely happens before consequences land. **Irreversible:** cannot be meaningfully undone.

2. Named accountability. Does the domain require a specific, identifiable accountable human per decision? **None,** **Aggregate** (a role owns the class of decisions), or **Per-decision** (a named individual required).

3. Explainability obligations. Must the decision path be reconstructable in human terms? **None,** **Internal** only, or **Legal**, where a regulator or court can demand it.

4. Regulatory classification. Assigned by domain and effect on people, not chosen by the organization. Reference points include the EU AI Act's risk tiers and applicable sectoral rules.

5. Blast radius. Single artifact, **Team** or **product scope**, or **Systemic**.

The system's risk class is set by its most restrictive factor, not an average across the five.

IF ANY FACTOR READS...	CLASS FLOOR
Irreversible, or per-decision accountability, or systemic blast radius	Critical
Limited reversibility, or legal explainability with sectoral high-stakes classification	High
Effortful reversibility, or legal explainability without high-stakes classification, or team/product blast radius	Moderate
Full reversibility, no special accountability, no legal explainability, single-artifact blast radius	Low

Classification is set at charter time by approvers from three functions: engineering leadership, risk or compliance, and the business owner, not by the engineering team alone.

III

LOW RISK CLASS

Standard Controls, Generous Ceiling

Typical profile: internal tooling, fully reversible, no regulatory exposure, single-artifact blast radius.
Example: an internal code-review assistant, a draft generator for internal documentation.

Maximum permitted level: 4, the only class where Continuous Assurance is available at all.

CONTROL GROUP	REQUIREMENT AT LOW
NAV-1.x	Standard form: usage policy, provenance tagging, basic reviewer training.
NAV-2.x	Standard form: sandbox, checkpoint enforcement, logging, kill switch, no enhancement.
NAV-3.x	Standard form if operating at Level 3: trigger catalogue, resolver authority, near-miss review quarterly.
NAV-4.x	Standard form if operating at Level 4: sampling-based audit, no elevated confidence requirement.
NAV-G.x	A single named accountable owner is sufficient; recertification biennial.

Sign-off authority: engineering leadership alone is sufficient at every gate.

What Low risk does not mean: it does not mean no controls. Every Hard control still applies in full; nothing about this class weakens a Hard control. Only the formality and frequency of Graduated controls changes.

IV

MODERATE RISK CLASS

Enhanced Escalation Governance

Typical profile: customer-facing or revenue-affecting systems, standard data-protection exposure, moderate blast radius. Example: a customer support agent handling account inquiries, a marketing content loop publishing under the brand's name.

Maximum permitted level: 3.

CONTROL GROUP	REQUIREMENT AT MODERATE
NAV-1.x	Standard form.
NAV-2.x	Standard form.
NAV-3.x	Enhanced: near-miss review monthly. Evasion testing at least twice yearly. Resolver roles are named individuals, not a shared queue.
NAV-4.x	Not applicable; ceiling prohibits.
NAV-G.x	Named owner or small cross-functional group; recertification annual.

Sign-off authority: engineering leadership plus a risk or compliance representative at Gate 2 to 3.

Common failure point: treating "customer-facing" as automatically requiring maximal scrutiny, when the actual governing factor is usually reversibility and blast radius, not customer-facing status alone. Do not over-classify by instinct; use the determinants.

V

HIGH RISK CLASS

Legally Accountable Resolvers, Complete Evidence

Typical profile: regulated workload, AI as decision-support in a sectoral high-stakes domain, limited reversibility. Example: a loan pre-screening assistant, a clinical documentation summarizer feeding into care decisions.

Maximum permitted level: 2 to 3, and only to Level 3 where the risk ceiling determinants genuinely support it; many High-class systems will find their honest ceiling is 2.

CONTROL GROUP	REQUIREMENT AT HIGH
NAV-1.x	Enhanced: reviewer-quality sampling moves to a defined statistical confidence level, not spot checks.
NAV-2.x	Enhanced: evidence retained at the regulatory retention horizon, not internal default.
NAV-3.x	Enhanced, where Level 3 is defensible: resolver must be the legally accountable individual. Near-miss review monthly. Evasion testing every release cycle. Ledger and traceability move to complete-population evidence.
NAV-4.x	Not applicable; ceiling prohibits.
NAV-G.x	Cross-functional governance forum required; recertification annual; exceptions sign off one level higher than Moderate.

Sign-off authority: the risk function itself, not engineering leadership alone, from Level 1 to 2 onward.

Common failure point: treating Level 3 as an achievable engineering milestone rather than a conditional one. Capability never overrides the ceiling; this is the class where that principle gets tested most often.

VI

CRITICAL RISK CLASS

Checkpoint Oversight Is the Destination

Typical profile: safety-of-life or irreversible decisions, AI genuinely in the decision loop for consequential outcomes. Example: an autonomous system making medical dosing recommendations acted on without further review, a system with authority to move funds without human confirmation.

Maximum permitted level: 1 to 2. Per-instance human decision is retained.

CONTROL GROUP	REQUIREMENT AT CRITICAL
NAV-1.x	Enhanced: complete-population review, not sampling, for any AI contribution.
NAV-2.x	Enhanced, and terminal: this is the destination, not a waypoint. Held to the highest evidentiary standard in the catalogue.
NAV-3.x	Not applicable. The risk ceiling prohibits Level 3; escalation-based oversight is not a substitute for per-instance human decision here.
NAV-4.x	Not applicable.
NAV-G.x	Full governance forum including legal and compliance; recertification continuous and condition-based.

Sign-off authority: board-level or equivalent risk owner for any exception; routine sign-off requires the risk function and a named executive, not engineering leadership alone.

A roadmap showing a Critical-class system progressing toward Level 3 is not an ambition. It is a governance finding.

VII

Reclassification

STABLE PRINCIPLE

A system's risk class is not fixed at charter time forever. A feature moves from internal to customer-facing. A new regulation lands. Scope drifts into a regulated domain without anyone deciding it should. Reclassification is a tracked event, not a silent drift, and requires the same three approvers who set the original classification, so the chain of evidence remains intact.

PROVISIONAL Reclassification review should be triggered by: a change in the system's action space, a change in its user base or exposure, a change in applicable regulation, or a periodic re-validation per NAV-4.7 where operating at Level 4. Any of these should pause further autonomy increases until reclassification is complete.

VIII

Worked Example

System: an internal engineering assistant that drafts pull requests for dependency updates.

Classification walkthrough. Reversibility: full, every change goes through normal code review before merge. Named accountability: none beyond normal engineering accountability. Explainability: internal only. Regulatory classification: unregulated. Blast radius: team or product scope at most.

Result: Low risk class. Maximum permitted level: 4.

Capable Stage assessment (illustrative): the organization has a solid harness but no escalation trigger catalogue yet and no named resolver role. Capable Stage: 2.

Diagnostic: Certified Stage = min(Actual Stage 2, Capable Stage 2, Risk Ceiling 4, Evidence-Supported Stage 2) = 2. Finding: **Aligned**. The organization is operating exactly where its capability and evidence say it should, well within a generous risk ceiling it has room to grow into.

A Low-risk classification does not mean the assessment is trivial. It means the ceiling is generous. This example correctly reflects a healthy, honest state, not a problem to fix.

IX

What This Guide Does Not Yet Resolve

- The classification floor table in Section II is a working hypothesis; real engagements will surface edge cases.
- Sign-off seniority requirements are stated functionally, not as specific titles, because org structures vary too much for one title to apply universally.
- The reclassification triggers in Section VII are asserted, not derived from observed drift patterns.
- This guide does not and cannot resolve the silent-escalation-failure problem. It applies with equal force at every risk class.

Where to Go Next

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

Running an assessment? Start with the Navigator Assessment Guide.

Implementing the control set itself? Start with the Navigator Controls Catalogue.

Classifying a specific system right now? Section II's determinants and classification table are what you need, followed by the matching section.

**This document establishes
exactly what applies,
and how strictly,
at each risk class.**

Rajesh Srinivasan

Orchestra AI · rajesh@orquestraai.io
orquestraai.io

Orchestra Navigator™ Critical Controls Guide · Working draft, not yet released.

Trademark application in preparation, not yet registered.

© 2026 Orchestra AI.