UNISINOS

Architectural Governance of Autonomy in Agentic AI Systems

Founding Case Analysis and Research Programme Specification

DOCUMENT D0

Executive Brief

> *The entry point to the knowledge base. States what it claims and how it was built, the most significant findings, the headline counts, the decisions still open for the supervisor, and how to read the other six documents.*

CANDIDATE	M.Sc. Epitácio Vicente do Nascimento Neto
SUPERVISOR	Prof. Dr. Felipe André Zeiser
PROGRAMME	Graduate Program in Applied Computing (PPGCA), PhD. University of Vale do Rio dos Sinos (UNISINOS)
RUN DATE	7 August 2026
DOCUMENT VERSION	1.1
RUN IDENTIFIER	`20260807-1714`

How to read this document set

SINGLE SOURCE OF TRUTH

Every fact in this set is held once, in a machine readable register under the knowledge base `registers/` directory, and every document is rendered from those registers.

A guardrail, an action class or a metric is described once and cited everywhere else by its identifier. Where a document needs narrative around a register entry, the narrative lives in the document and the facts live in the register.

IDENTIFIER SCHEME

- `AC` action class
- `GR-Lnn` guardrail, by layer
- `SD` supervision determinant
- `SP` supervision point
- `MG` metric
- `CI` composite index
- `DM` applicability domain
- `CS` condition set
- `CB` cloud capability contract
- `BL` backlog item
- `F` durable finding
- `EV` evidence file

NAMING POLICY

This set is identity-free by construction, not by later redaction. No address, hostname, cloud identifier, repository URL, personal name other than the candidate and the supervisor, or customer reference appears anywhere in it.

Hosts are named by service role. Vendor and open technologies are named, because the technical content depends on it. Raw evidence stays on the host under `evidence/` and is cited by identifier, never quoted.

READING ORDER

- **D0** Executive Brief, read first
- **D2** Supervision and Autonomy Framework, the argument
- **D1** Founding Case Architecture, the evidence base
- **D3** Guardrail Catalogue, reference
- **D4** Measurement Framework, reference
- **D5** Applicability Domains
- **D6** Research Environment and Roadmap

EVIDENCE MARKS

- **OBSERVED** verified on the host in this run, with an evidence identifier.
- **REPORTED** drawn from the doctoral proposal or the run brief, not verifiable on this host.
- **PROPOSED** a design proposal of this run, not a description of anything that exists.
- **NOT DETERMINED** could not be established; the reason is always stated.

CROSS-REFERENCES

Cross-references are by document number and identifier, never by page number, because page numbers change when a register grows and the document is re-rendered.

Example: *see D3,* `GR-L06-02`.

Contents

1. Purpose and method

This knowledge base documents a real agentic AI environment, the founding case, to support a thesis that instruction-based governance of autonomous agents is insufficient and that enforcement must be architectural. The founding case is a multi-agent engineering estate built and operated by the candidate on production cloud infrastructure. One host in that estate, referred to throughout as the Agent Console host, was examined in detail in a single collection run identified as `20260807-1714`, captured on 7 August 2026. Nothing in this document set is inferred from memory or from the doctoral proposal alone; every factual claim carries an evidence identifier pointing at a raw probe captured on that host on that date, or is marked **REPORTED** where the source is local documentation this run could not independently exercise.

The base is read-only discovery, analysis and design. Nothing was remediated during the run: no control was implemented, no gap closed, no configuration touched, because the founding case is the empirical baseline against which later experimental work is compared, and changing it during observation would destroy that baseline. The discipline that replaces remediation is evidence citation throughout, a strict separation of **OBSERVED** (verified on this host), **REPORTED** (from the proposal or brief, not locally verifiable) and **PROPOSED** (a design proposal that does not exist), and identity-free construction: hosts, people and organisations are referred to by service role or by role only, and named technologies reveal a build choice, never who runs it.

Every claim is organised into sixty action classes, a 134-entry guardrail catalogue across twelve layers, ten supervision determinants, thirty-four supervision points, a seventy-two entry metric register, nine applicability domains, ten experimental condition sets and a fifty-eight item improvement backlog, all cross-referenced by fixed identifier so that a claim in prose and a row in a register are always the same fact stated twice, never two facts that happen to agree. Six documents, D1 through D6, each carry one part of the argument in full; this document, D0, is written last, after all six, and summarises rather than restates them. Where this document and one of the six disagree on a figure, the six are correct and this one should be fixed.

The registers are additive by design: an identifier, once allocated, is never renamed or renumbered, and an entry that no longer applies is marked retired rather than removed, because rendered documents already cite it. This matters for how the findings in section 2 should be read. They are not a survey of opinions about the founding case; they are conclusions computed from the registers by a stated rule, applied uniformly, and every number below can be recomputed from the cited register rather than taken on this document's word. Two disciplines run underneath every figure quoted below and are stated once here rather than per finding. First, where a worked example could not be verified on this host because the mechanism sits on another estate host, it is marked **REPORTED** and is load-bearing anyway, because omitting the sharpest cases in the estate to avoid the mark would be a worse distortion than carrying it. Second, this run is a single observer's single snapshot of one environment; D2's own limitations section states that plainly, and this document inherits the caveat rather than restating it in full.

---

2. The most significant findings

Access exceeds mandate on most of the inventory. Classified twice, against what an agent is technically capable of doing and against what it is authorised to do, fifty of the sixty action classes show access more permissive than mandate. The mandate is real in the sense that the operator means it; it exists only as prose in files the governed party can itself write. None of the fifty deltas is backed by a structural guardrail that would make the prohibited action unreachable (D1, `F-029`).

Enforcement is overwhelmingly not structural, on both cuts of the data. Of the sixty action classes, thirty-six carry no catalogued control at all, eight are bounded by prose alone, twelve by a check the acting party could also alter, and four by a control it cannot reach around. Fifty-six of the sixty therefore have no structural guardrail bound to them today. This is a narrower and more consequential claim than the guardrail catalogue's own population, which is thirty-nine present, seventeen partial and seventy-eight proposed across its 134 entries, with forty-nine classified structural in kind: nearly all of that structural capacity is proposed rather than built and bound to an action (D1, section 9; D3, section 3).

Two present controls are genuinely structural, and both share one shape. The read-only connection identity to the System of Record and the structural test mode overlay in the real-time domain are the only entries in the whole catalogue where the prohibited effect is not reachable at all, rather than merely reviewed or discouraged; compliance is not a term in either outcome. Both are REPORTED rather than independently verified on this host, and both are the reference pattern the proposed reference architecture is built to generalise (D2, section 5.1, `SP-029` / `F-023` and `SP-024` / `F-019`).

Supervision that an architecture claims and supervision a human can exercise have come apart. Thirty-four supervision points were assessed against six viability conditions, independence, latency, halt, attention, comprehensibility and authority. Twenty-seven diverge. The binding condition is independence in eight cases and latency in eight, halt in three, authority in three, attention in two and comprehensibility in two, and in one case the claimed control does not exist. Across forty-six days of console audit logs there is no record of a human stopping, pausing, rejecting or overriding an agent on the basis of what it was doing (D2, section 4; `F-042`).

The irreducible supervision residue is real but smaller than it first appears. Applying ten supervision determinants to the inventory, thirty-two of sixty action classes remain above Level I under every catalogued guardrail configuration, even the fully hardened one. The largest single group within that residue, nine action classes, is pinned by containment shortfall, which is fully removable by identity engineering rather than by any new supervisory mechanism; the genuinely irreducible portion, pinned by externality, accountability or irreversibility, is smaller than the raw figure suggests (D2, section 3.3).

Governance overhead is currently unmeasurable, not merely unmeasured. The metric register's own governance overhead subfamily, effort spent supervising against effort spent building, is absent or mixed across all three of its entries, because each is composed from component metrics that are themselves not yet computable. More broadly, of seventy-two specified metrics, only fourteen are available on this host today with no further instrumentation, twenty-five are derivable from data that exists but is not yet joined, and thirty-one require a data source that does not exist in any form. The gap is concentrated, not scattered evenly: repair capability is entirely absent, because no fault, incident or repair register exists anywhere on the host to draw a baseline from at all (D4, sections 3.9 and 6.1).

A hypothesis carried into this run was corrected rather than confirmed. The local verification affordance, a mechanism believed to be a live authentication bypass, was found implemented, exported and compiled into the running build, but bound to no route; a live probe returns unauthorised, not from the affordance, because there is no affordance left to reach. This correction, rather than a quiet adjustment of the brief to match what was found, is treated as a methodological result in its own right, because a discovery that only confirms its own priors is not evidence (D1, section 6.1, `F-035`).

The run's own toolchain supplied a live instance of the thesis. Three attempts to render this document set were ended by the kernel's out-of-memory killer after a parser stalled and grew without emitting any signal of its own; the only mechanism that ever stopped it was an external, structural control owing nothing to the process it killed. The failure, and its conversion into a contained, checkpointable, single-document error before any further rendering was allowed, is recorded as a

durable finding and cited in the framework's discussion of why an oversight signal must not originate from the supervised process (D2, section 4.2; `F-043`).

3. Headline counts

REGISTER OR MEASURE	COUNT
Action classes	60, of which 50 show access exceeding mandate
Guardrail catalogue	134 entries; population 39 present / 17 partial / 78 proposed; enforcement class 49 structural / 60 conditional / 13 instructional / 12 none
Action classes with a structural guardrail bound today	4 of 60 (56 have none)
Supervision determinants	10 (`SD-01` to `SD-10`), 8 from the proposal, 2 added this run
Supervision points	34, of which 27 diverge between nominal and effective posture
Irreducible supervision residue	32 of 60 action classes
Experimental condition sets	10 (`CS-01` bare baseline to `CS-09` fully hardened, plus one ablation)
Metrics	72; readiness 14 available / 25 derivable / 31 absent (2 entries split)
Composite indices	8, of which none is available and three are absent
Applicability domains	9 (`DM-01` to `DM-09`)
Cloud capability contracts	15, of which 11 are not transferable from the founding case as found
Improvement backlog	58 items; 43 tagged `BOTH` , 11 `RESEARCH` , 4 `PRODUCT` ; phased 28 / 10 / 5 / 15
Durable findings	43 (`F-001` to `F-043`)

The measured behavioural baseline behind the whole programme, drawn from the proposal rather than verified on this host, is a 61.1 per cent factual divergence rate between agent self-report and independent evidence across eighteen audited episodes drawn from forty-six candidates the system itself identified (D2, section 1; D4, section 3.1).

4. Three decisions for the supervisor

The build plan for the isolated experimental environment leaves three genuine forks open. None has a basis in the evidence gathered by this run sufficient to resolve it unilaterally; each has a real cost on both sides.

Which condition set to pilot first. Piloting `CS-01` against `CS-03`, the control condition against the founding case as found, needs the fewest new components and gives the earliest check that the synthetic environment approximates the real one, but says nothing yet about the thesis itself. Piloting `CS-02` against `CS-04`, instructional governance against identity separation, speaks directly to the central claim, but needs the classifier, mandate service and identity layer working before the first data point exists at all. The choice is a judgement about how much confidence in the environment's fidelity should be established before the framework is tested against it (D6, section 9.1).

Whether to build the runtime classifier or the evidence bus first. Both are phase 1, both are the largest single items in the backlog by effort, and neither strictly depends on the other. Building the

classifier first makes class-and-level assignment live before there is anywhere durable to record what was decided, so its own early behaviour is unverifiable for a period. Building the evidence bus first makes tamper-evident recording available before there is anything classified worth recording, reproducing a smaller version of the founding case's own current state. A third option, building both in parallel to a minimal viable state, avoids the gap at the cost of dividing a fixed amount of candidate time across two large items at once (D6, section 9.2).

How much of the cloud-neutral binding layer to build before the first experiment. Building both provider bindings for all fifteen capability contracts up front guarantees every result carries an unqualified cloud-neutrality claim, at the cost of moving a phase-four-sized body of work to the front of the programme. Building a single-provider binding first gets a pilot running inside phase one's own timeline, at the cost of a retrofit risk on the four contracts, workload identity, the policy enforcement point, the append-only log sink and the ephemeral credential contract, where the two providers do not merely differ in implementation but in what they structurally cover. A middle position closes the gap only for those four and treats the remaining eleven as single-provider for as long as the programme's timeline requires (D6, section 9.3).

---

5. How to read this document set

Seven documents in total, this one included. The reading order below, D0 first, then D2, D1, D3, D4, D5, D6, is the order the whole set is designed to be read in, and it is not the document numbering, because the argument reads better before the evidence that grounds it than after.

D0, Executive Brief (this document). What the knowledge base claims and how it was built, the most significant findings, the headline counts, the open decisions, and this reading order. Four to six pages, meant to stand alone.

D2, the Contribution. The argument itself: autonomy as a function of guardrail maturity, the ten supervision determinants and the irreducible residue they produce, and the six viability conditions that separate claimed supervision from supervision a human can actually exercise. Read this second because it states the claims that D1's evidence and D3 through D6's instruments exist to support; reading it before D1 keeps the argument from being lost inside the inventory.

D1, the Founding Case. A dated snapshot of the Agent Console host as found: its architecture, the full agent session lifecycle, its automations, its governance surfaces, and the sixty-entry action class inventory classified for access and mandate. This is the evidentiary spine every other document cites into. Read it after D2 to see the argument's grounding, not before, because the inventory alone does not carry the argument.

D3, the Guardrail Catalogue. The 134-entry catalogue across twelve layers, the enforcement classification that separates structural from conditional from instructional from none, the coverage matrix, and the ten experimental condition sets that vary guardrail maturity deliberately. This is the vocabulary D2 depends on, made exhaustive.

D4, the Metric Register. Seventy-two specified metrics and eight composite indices, built so that every claim can be tested for guardrail sensitivity and every claim drawn from an agent's own account carries a named, independent corroborating metric. This is the measurement instrument for the whole framework, including for the claim of divergence that opened the argument in D2.

D5, the Domain Map. The D2 framework and the D3 catalogue applied to nine applicability domains, ranked by defensible autonomy, closing with a five-dimension procedure an organisation without this estate can run to locate its own work on the same map.

D6, the Reference Architecture. A candid gap analysis of what the founding case supplies and what it must not supply, a seven-layer reference architecture and cloud-neutral capability contracts for the

isolated experimental environment the doctorate requires, a phased fifty-eight item backlog, and the three decisions in section 4 above, stated there in full.

A reader with time for one document beyond this one should read D2. A reader checking a specific figure should go directly to the register it cites, using the identifier printed alongside the claim; every identifier in this document and in D1 through D6 resolves to exactly one row in exactly one register, by construction.

This document's appendix is the run's decisions register, kept in the knowledge base as a run record: every point the brief left open, the options weighed, the choice made and the reason. It stays with the knowledge base rather than inside these pages so that this brief remains readable in one sitting.