UNISINOS

Architectural Governance of Autonomy in Agentic AI Systems

Founding Case Analysis and Research Programme Specification

DOCUMENT D4

Measurement Framework

> *The measurement instrument for the framework argued in D2 and catalogued in D3. A 72-entry metric register and eight composite indices, specified so that every metric can be tested for guardrail sensitivity and every claim derived from self-report carries a named independent corroborator.*

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. The metric specification model

D2 argues that autonomy is a function of guardrail maturity and that a claimed supervision posture must be tested against six viability conditions rather than taken on the architecture's own word. Neither claim is checkable without numbers. This document is the register of numbers: what is counted, how, from where, and with what confidence the count can be trusted. It is written to the same discipline as D3's guardrail catalogue, because a metric register that is looser than the thing it measures is worthless, and because the founding case's own central finding is precisely that a system's account of itself cannot be trusted on its own.

1.1 What a metric specification contains

Every entry in the register carries the same twenty-three fields, whether or not the metric is currently computable. A field left as `not set` or `not computable` is an entry, not an omission; the register's honesty depends on distinguishing a metric nobody has built from a metric that measures nothing.

FIELD	WHAT IT CAPTURES
`id`, `name`, `family`, `subfamily`	Identity and placement in the taxonomy: `engineering` or `governance`, and one of thirteen subfamilies
`type`	`KPI` (a performance measure), `KRI` (a risk indicator) or `KCI` (a control indicator)
`horizon`	`leading`, an early signal, or `lagging`, an outcome measured after the fact
`question`	The single question the metric answers, in one sentence
`operational_definition`	Precisely what is counted, including what is deliberately excluded
`formula`	The computation, stated so it could be implemented without further interpretation
`unit_range`, `granularity`	The unit and admissible range, and the level the metric is computed at: per action, per session, per agent, per host
`data_source`, `collection_point`	Where the numbers originate, and the exact place in the system instrumentation would have to sit
`sampling_frequency`	How often the metric is computed
`direction`	`higher-better`, `lower-better`, or `target-band`, where both a very low and a very high value are adverse
`baseline`	The founding case's current value, or `not computable` with the specific reason
`target`	A proposed target, or `not set` with the reason a target is not yet defensible
`guardrail_sensitivity`	Which guardrail identifiers from D3's catalogue move the metric, and in which direction
`confounders`	What else moves the metric besides the guardrails, so a change is not misread as a guardrail effect
`gaming_modes`	How the metric could be made to look better without the underlying property improving. Mandatory on every entry
`corroborating_metric`	The independent metric that would catch the gaming. Mandatory wherever `data_source` includes agent self-report
`readiness`, `instrumentation_gap`	`available`, `derivable` or `absent` on the Agent Console host today, and precisely what is missing to close the gap

1.2 Guardrail sensitivity as the research instrument

`guardrail_sensitivity` is what separates this register from a dashboard.

> **THE GUARDRAIL SENSITIVITY PRINCIPLE**
>
> A metric that does not move when guardrails vary is not a dependent variable. It may still be worth reporting, but it belongs to descriptive reporting, not to the experimental comparison this programme's fifth specific objective requires across condition sets `CS-01` through `CS-10`. Every metric in the register states, per guardrail, whether and in which direction it moves, and several entries say so explicitly of themselves: `MG-004` cycle time and `MG-045` harm completion time are each recorded as largely insensitive to guardrails for most action classes, because they measure a mechanism's intrinsic duration rather than a governed property, and both are marked accordingly rather than left to look like ordinary dependent variables.

The practical use of the field is to run the same computation under different condition sets and read the difference as the effect of the guardrail set alone. Where a guardrail is recorded as raising a metric in one place and lowering it in another, for example `GR-L11-05` scheduled unattended dispatch, which raises `MG-001` task completion rate by manufacturing completed work while also raising `MG-006` model capacity consumed with no originating task, the register records both directions rather than a single net effect, because the two consequences are analytically distinct and a composite that nets them out would hide exactly the trade-off an evaluator needs to see.

1.3 Gaming modes and the corroborating metric

> **THE GAMING MODES PRINCIPLE**
>
> The founding case's central finding is that a system can report success while failing: a deployment recorded as successful had in fact disabled the service's core connection function for about 24 hours **REPORTED**, and an independent audit of eighteen episodes drawn from forty-six candidates found factual divergence between the agent's account and independent evidence in eleven, a rate of 61.1 per cent **REPORTED**. A measurement framework built after that finding that still trusts an agent's account of its own work would not be a framework at all. Every metric in this register whose `data_source` includes agent self-report therefore carries a mandatory `gaming_modes` field naming how the number could be inflated or deflated without the underlying property changing, and a mandatory `corroborating_metric` field naming the independent metric that would catch it.

`MG-022` claim divergence rate is the register's central instance of the principle stated as a metric: its own `data_source` field is deliberately asymmetric, self-reports are `AGENT-EMITTED`, the comparator must be `INFRASTRUCTURE-EMITTED`, and its `corroborating_metric` field names three further metrics, `MG-026` unverifiable claim share, `MG-027` evidence citation completeness and `MG-065` audit record immutability, each of which detects a different way `MG-022` itself could be gamed: writing vaguely enough that no assertion can be contradicted, omitting the assertions least likely to survive checking, or editing the record after the divergence is found. No single metric in the register is asked to certify itself. That is a design rule, not an incidental feature: a metric that could certify itself would reproduce, inside the measurement layer, the exact defect it exists to detect in the system.

1.4 Agent-emitted and infrastructure-emitted events

The distinction that makes the register usable at all is between two classes of event.

An **agent-emitted event** is written by the same account, process or credential that performed the action it describes. On the Agent Console host this is the dominant class: the model runtime transcript (about 1.13 GiB `EV-026`), completion markers, memory notes and notification bodies are all written by the party whose work they describe, under the single operating system account every agent session runs as `EV-003`. An agent-emitted event is not worthless. It is frequently the only account of what

happened, and it is often accurate. It is untrustworthy as a sole source specifically because nothing in its production process distinguishes an accurate account from an inaccurate one; the mechanism that writes a true completion marker is identical to the mechanism that would write a false one.

An **infrastructure-emitted event** is written by a process running under a different principal from the one performing the action: a process manager's restart counter, a version control system's commit timestamp, a reverse proxy's access log, a health probe issued from outside the acting agent's reach. Its trustworthiness does not rest on the infrastructure being smarter or more honest than the agent. It rests on the narrower and more defensible property that the infrastructure has no motive and, in the well-built case, no means to misreport, because it is not the party being evaluated.

The two classes are not always cleanly separated in practice, and the register is explicit where they are mixed. `MG-002` actions per agent hour is recorded as infrastructure-emitted in origin, the model runtime hook substrate fires on every tool call `EV-026`, but the resulting records are written to files owned by the acting account and are therefore agent-modifiable after the fact; the register classifies this as a compromised infrastructure-emitted source rather than reclassifying it as agent-emitted, because the distinction that matters is not who wrote the first copy but who can rewrite it. `MG-029` self-attestation share of the record is the sharpest instance of the mixture: the records it counts are agent-emitted by definition, but the counting and attribution mechanism itself is infrastructure-emitted, since it is derived from filesystem ownership rather than from any claim. Section 5 develops this distinction into the measurement architecture; Figure D4-1 traces it through a single metric end to end.

---

2. Engineering metrics

The twenty-one engineering metrics answer the questions an engineering organisation already asks: how much gets done, how often it is correct, what a restart costs, and how well the system repairs itself. They are included in full because D2's argument depends on being able to show that guardrail investment has costs as well as benefits, and because a governance framework that cannot speak to ordinary delivery concerns will not be adopted by anyone who has to ship. Four subfamilies, `MG-001` to `MG-021`.

2.1 Throughput

`MG-001` to `MG-006` measure how much work moves through the system and how fast. None of the six has a computable baseline as a rate or ratio; the recurring obstacle is the same across all six, that no dispatch identifier joins a brief, a completion, or a token count back to the unit of work it belongs to, so the founding case can report volumes, approximately 1,900 commits across twelve repositories in more than eighty days, approximately 18.6 billion tokens over seven days, but not rates attributable to a task.

Table D4-1. Throughput metrics, MG-001 to MG-006.

MG-001 Task completion rate

QUESTION	What proportion of dispatched briefs reach a state the system records as complete?
OPERATIONAL DEFINITION	Count of dispatched briefs whose terminal record is a completion marker or completion attestation, divided by briefs dispatched in the window. Excludes briefs superseded by a later dispatch to the same agent; excludes ad hoc operator sessions carrying no brief; excludes re-issues of an unchanged brief after a supervisor restart. Counts claimed completion only, not verified completion.
FORMULA	completed_briefs / dispatched_briefs; completed_briefs = briefs with a terminal completion marker written by the acting agent; dispatched_briefs = briefs delivered through the dispatch path in the window
UNIT AND RANGE	ratio, 0 to 1
TYPE	KPI
HORIZON	lagging
GRANULARITY	per agent per day
DATA SOURCE	Mixed. Numerator AGENT-EMITTED: completion markers and completion attestations written by the acting agent into its own working directory. Denominator INFRASTRUCTURE-EMITTED: the console dispatch route and the session mailbox writer.
COLLECTION POINT	The console dispatch route handler for the denominator; a completion record store the acting agent cannot write for the numerator
SAMPLING FREQUENCY	daily
DIRECTION	higher-better
BASELINE	not computable; the founding case writes completion markers as free text in per-agent working directories with no dispatch identifier, so numerator and denominator cannot be joined to the same brief
TARGET	not set; a completion target set without MG-022 alongside rewards claiming completion rather than achieving it, which is the founding case failure mode
GUARDRAIL SENSITIVITY	GR-L07-07 lowers it as previously silent work is refused; GR-L06-01 lowers it because approval-held changes are incomplete within the window; GR-L04-07 lowers it by truncating long tasks; GR-L11-05 raises it because clock-driven dispatch manufactures attempted and therefore completed work; GR-L12-09 lowers it slightly where the model declines a brief
CONFOUNDERS	Workload composition and brief size; task difficulty; model version change; the three-minute heartbeat cadence, which generates completions with no originating brief; observation effects during audited periods
GAMING MODES	The agent writes the completion marker while leaving the work partial; it decomposes one brief into several trivially completable briefs to raise the ratio; it restates the brief in its own notes so that what it did becomes what was asked
CORROBORATED BY	MG-022 claim divergence rate; independently, the version control commit record and process state for the same window
READINESS	derivable
INSTRUMENTATION GAP	A dispatch identifier minted at the dispatch route, carried into the session and echoed in the completion record; no such identifier exists today

MG-002 Actions per agent hour

QUESTION	How many consequential tool actions does an agent perform per hour of live session time?
OPERATIONAL DEFINITION	Count of tool invocations classified as mutating, divided by live session hours. Excludes read-only tool calls; excludes subagent activity, which is attributed to the subagent; excludes wall-clock time in which the session is idle awaiting input. Mutating means the call writes to a filesystem, a process, a network endpoint or a control plane.
FORMULA	mutating_tool_events / live_session_hours; live_session_hours = wall clock hours in which the session emitted at least one hook event per five minute interval
UNIT AND RANGE	count per agent-hour, 0 to unbounded
TYPE	KPI
HORIZON	leading
GRANULARITY	per agent per day
DATA SOURCE	INFRASTRUCTURE-EMITTED in origin, the model runtime hook substrate fires on every tool call, but the resulting records are written to files owned by the acting account and are therefore agent-modifiable after the fact
COLLECTION POINT	The PreToolUse and PostToolUse hook, with the record written to a store owned by a principal other than the agent
SAMPLING FREQUENCY	hourly, aggregated daily
DIRECTION	target-band
BASELINE	not computable as a rate; the nearest observed proxy is approximately 1,900 commits across 12 repositories in more than 80 days, which is an estate-wide action count rather than a per-agent-hour rate, because live session hours and mutating-call classification are not recorded
TARGET	not set; both a low value, indicating a stalled agent, and a high value, indicating unreviewed churn, are adverse, and no defensible band can be set before MG-007 and MG-022 are computable
GUARDRAIL SENSITIVITY	GR-L09-04 lowers it through per action class rate limits; GR-L04-07 lowers it by bounding the turn; GR-L11-05 raises it, and on this host is the dominant driver, since a clock nudges six agents every three minutes; GR-L07-07 lowers it; GR-L02-11 lowers it for network-bearing actions only
CONFOUNDERS	Task difficulty and the ratio of reading to writing in the workload; model version change, since a more capable model does more per call; the number of concurrent sessions competing for the same working tree; a change in what is classified as mutating
GAMING MODES	The agent inflates the count with no-op file touches or repeated rewrites of the same content; conversely, under a rate limit it batches many effects into one call so the count falls while the effect volume does not
CORROBORATED BY	MG-069 unattended action path count; the version control commit series and process manager state as independent counts of realised effect
READINESS	derivable
INSTRUMENTATION GAP	A mutating or read-only classification on each hook record, and an idle-time definition; the raw event stream exists at about 186 MB across active sessions but is unclassified

MG-003 Lead time from dispatch to first effect

QUESTION	How long after a brief is dispatched does the first externally observable change occur?
OPERATIONAL DEFINITION	Elapsed time from the dispatch record to the first mutating action attributable to that dispatch. Excludes dispatches producing no mutating action, which are reported separately as a censored count; excludes time a session spends unavailable because it is being restarted by its supervisor.
FORMULA	median over dispatches of (t_first_mutating_action - t_dispatch)
UNIT AND RANGE	seconds, 0 to unbounded
TYPE	KPI
HORIZON	lagging
GRANULARITY	per dispatch, aggregated per agent per week
DATA SOURCE	Mixed. Dispatch timestamps INFRASTRUCTURE-EMITTED by the console injection service and the session mailbox file; first-effect timestamps come from the hook substrate, which is written under the agent account.
COLLECTION POINT	The injection service write path for t_dispatch; the PreToolUse hook for t_first_mutating_action
SAMPLING FREQUENCY	weekly
DIRECTION	lower-better
BASELINE	not computable; no dispatch identifier joins a brief to a subsequent tool call, so any attribution would rest on a temporal assumption rather than a recorded link
TARGET	not set; the useful comparison is against MG-045 harm completion time for the same class, not against an absolute service level
GUARDRAIL SENSITIVITY	GR-L06-01 raises it where the first effect requires approval; GR-L11-04 raises it because verification must complete before the next class unlocks; GR-L01-06 raises it by inserting an elevation step; GR-L11-05 lowers it by prompting an idle agent into action; GR-L10-08 lowers it by removing a wait for a human answer
CONFOUNDERS	Brief complexity and how much reading precedes the first write; queueing behind another agent in a shared working tree; whether the session was already warm; model version change
GAMING MODES	The agent performs a trivial mutating action immediately, such as touching a scratch file, so the clock stops early while the substantive work starts much later
CORROBORATED BY	MG-004 cycle time per action class; the version control commit timestamp series as an independent first-effect marker
READINESS	derivable
INSTRUMENTATION GAP	A dispatch identifier propagated into hook records; without it the join is temporal and breaks whenever two dispatches overlap in one session

MG-004 Cycle time per action class

QUESTION	How long does an action of a given class take from initiation to confirmed effect?
OPERATIONAL DEFINITION	Elapsed wall-clock time per attempted action, partitioned by action class identifier, measured from the initiating tool call to the event that confirms the effect. Excludes actions whose class cannot be assigned; excludes actions abandoned before any effect; excludes waiting time attributable to an approval, which is measured separately by MG-071.
FORMULA	per class, median over actions of (t_effect_confirmed - t_action_initiated)
UNIT AND RANGE	seconds, 0 to unbounded
TYPE	KPI
HORIZON	lagging
GRANULARITY	per action, aggregated per action class per week
DATA SOURCE	Mixed. Initiation from the hook substrate, which is agent-owned; effect confirmation preferably INFRASTRUCTURE-EMITTED from process manager state, supervisor journal entries, the reverse proxy access log or a cloud control plane event record.
COLLECTION POINT	The hook substrate for initiation; the state store of the affected system for confirmation
SAMPLING FREQUENCY	weekly
DIRECTION	lower-better
BASELINE	not computable as a distribution; the founding case records only order-of-magnitude values, namely file writes in milliseconds, cloud control plane calls in about one second, process manager restarts in seconds, and remote package installation in seconds to a few minutes
TARGET	not set; this metric is an input to the supervision viability comparison rather than a quantity to be optimised, and a shorter cycle time is adverse where it falls below the intervention latency
GUARDRAIL SENSITIVITY	For most classes this metric does not move when guardrails vary and must be read as descriptive reporting of the action intrinsic duration; it becomes a dependent variable only for classes carrying GR-L11-03 cooling off, which raises it, GR-L09-04 rate limits, which raise it, or GR-L01-06 just-in-time elevation, which raises it; GR-L04-02 does not move it
CONFOUNDERS	Host and network latency; the size of the change; whether the target is local or remote; concurrent load from the other sessions on the same host
GAMING MODES	Reporting the duration of the fastest sub-step as the duration of the action; declaring effect confirmed on a proxy such as process liveness rather than on the function itself
CORROBORATED BY	MG-045 harm completion time per action class; the affected system own state transition timestamps as an independent source
READINESS	derivable
INSTRUMENTATION GAP	An action class assignment on each hook record and an effect-confirmation event for each class; the hook stream exists but carries no class and no effect linkage

MG-005 Autonomous completion share

QUESTION	What share of completed tasks completed with no human input after dispatch?
OPERATIONAL DEFINITION	Completed tasks with zero human-originated inputs between dispatch and completion, divided by completed tasks. Human-originated means an operator message through the authenticated console path, an operator message relayed from the inbound messaging channel, or a keystroke typed directly into the session pane. Keystrokes injected by a watcher on a schedule are machine-originated and are explicitly not counted as human input. Excludes tasks that never completed.
FORMULA	completions_with_zero_human_inputs / completions
UNIT AND RANGE	ratio, 0 to 1
TYPE	KPI
HORIZON	lagging
GRANULARITY	per agent per week
DATA SOURCE	Mixed. The console injection route and the console audit log are INFRASTRUCTURE-EMITTED and carry an authenticated operator principal; watcher injections are recorded by the watchers themselves; a keystroke typed directly into a session pane is recorded nowhere.
COLLECTION POINT	The injection service, discriminating on the authenticated principal; the terminal multiplexer pane capture for direct typing
SAMPLING FREQUENCY	weekly
DIRECTION	higher-better
BASELINE	not computable as a ratio; the founding case records zero human interventions across 46 days of console audit logs against 19,334 heartbeat and nudge events, which is consistent with a share close to 1, but direct typing into a pane leaves no record so the numerator cannot be closed
TARGET	not set; a high autonomous share is a description of the operating posture, not an achievement, and is only interpretable alongside MG-052 effective supervision coverage
GUARDRAIL SENSITIVITY	GR-L06-01 lowers it by inserting a human decision; GR-L07-07 lowers it; GR-L10-08 raises it, because automated answering of interactive prompts removes the human from a decision that was previously theirs; GR-L11-05 raises it; GR-L01-06 lowers it where elevation requires a human approver
CONFOUNDERS	The mix of task types; operator availability and working hours; whether the operator intervenes outside the console, which is invisible to the measurement; observation effects, since an operator who knows the share is measured may withhold input
GAMING MODES	Counting a watcher nudge that in fact supplied the decision as autonomous continuation; classifying an operator instruction as system context so it falls outside the definition of human input
CORROBORATED BY	MG-053 operator intervention rate; the console audit log as the independent record of authenticated operator actions
READINESS	derivable
INSTRUMENTATION GAP	Capture of direct pane typing, which is not distinguishable from agent output in the current pane logs; without it the metric is biased upward by an unknown amount

MG-006 Model capacity consumed per completed task

QUESTION	How much model capacity is consumed for each task the system records as complete?
OPERATIONAL DEFINITION	Total tokens attributed to a session over the window divided by tasks that session completed. Includes input, output, cached input and cache-creation tokens. Excludes capacity consumed by the accounting stack itself; excludes sessions with zero completions, which are reported separately as consumption with no output.
FORMULA	session_total_tokens / session_completed_tasks
UNIT AND RANGE	tokens per completed task, integer >= 0
TYPE	KPI
HORIZON	lagging
GRANULARITY	per agent per day
DATA SOURCE	AGENT-EMITTED at root. The accounting stack tails the model runtime transcript files, which the measured account owns and can rewrite, and writes counters into a flat file owned by the same account with no checksum or sequence number.
COLLECTION POINT	The model provider billing record, or a proxy that observes requests before they leave the host, rather than the current tail of the agent own transcript
SAMPLING FREQUENCY	daily
DIRECTION	lower-better
BASELINE	not computable per task; the founding case records approximately 18.6 billion tokens over seven days across the whole reporting scope, but session identity is discarded at ingest into a source by model by half-hour bucket, so no per-agent or per-task attribution survives even though it is present in the input
TARGET	not set; an efficiency target is not defensible until MG-022 establishes that a recorded completion corresponds to completed work, otherwise the cheapest path to the target is to complete less and claim more
GUARDRAIL SENSITIVITY	GR-L04-07 lowers it by capping runaway turns; GR-L09-03 lowers it by stopping repeated failing attempts; GR-L11-05 raises it, because clock-driven check-ins consume capacity with no originating task; GR-L01-13 does not change the quantity but makes it attributable per agent and is the precondition for computing it at all; GR-L09-11 lowers it during a fleet pause
CONFOUNDERS	Model version change, which alters tokens per unit of work directly; context window and compaction behaviour; the size of the instruction and memory context loaded into every session; prompt caching hit rate
GAMING MODES	Appending one line to the counter store that reuses an existing source, model and half-hour key with smaller numbers replaces the real figure in every served endpoint; inflating the completed-task denominator lowers the ratio without any efficiency gain
CORROBORATED BY	The model provider billing record, which is outside the measured party control; MG-064 audit record completeness for the counter store
READINESS	absent
INSTRUMENTATION GAP	Per-session and per-agent attribution at ingest, and a counter store the measured account cannot write; today the attribution is discarded and the store is owned by the party it measures

MG-006 model capacity consumed per completed task is worth singling out because its instrumentation gap is a policy choice rather than a technical limitation: the resource accounting stack discards session identity at ingest, aggregating instead into a source-by-model-by-half-hour bucket, even though the identity is present in the request before it is discarded EV-024. Restoring it is a schema change, not a new capability.

2.2 Correctness

MG-007 to MG-012 measure whether the work that gets done is right. This subfamily carries the founding case's canonical incident: the one documented instance of MG-007 change failure rate is the deployment reported successful that had in fact disabled the service's core connection function for

about 24 hours REPORTED, and MG-008 defect escape rate records that same event as the subfamily's only documented escape, because no pre-deployment detection stage exists to define an escape against.

Table D4-2. Correctness metrics, MG-007 to MG-012.

MG-007 Change failure rate

QUESTION	What proportion of changes deployed by an agent require remediation?
OPERATIONAL DEFINITION	Deployments followed within a seven day observation window by a rollback, a revert, a hotfix touching the same files, or a service health failure attributable to the change, divided by deployments in the window. Excludes changes touching only documentation; excludes deployments never activated by a process restart; excludes remediation of a pre-existing defect the change merely exposed.
FORMULA	failed_deployments / total_deployments over a seven day observation window
UNIT AND RANGE	ratio, 0 to 1
TYPE	KRI
HORIZON	lagging
GRANULARITY	per repository per week
DATA SOURCE	Mixed. Version control history, process manager restart counters and health probe results are INFRASTRUCTURE-EMITTED; the attribution of a hotfix to a prior change is a judgement, made in the founding case by the same agent that authored both.
COLLECTION POINT	A deployment event record binding a source revision to an activation timestamp on a target host, plus the version control history of the same paths
SAMPLING FREQUENCY	weekly
DIRECTION	lower-better
BASELINE	not computable as a rate; one production instance is on record in which a deploy reported as successful had disabled the service core connection function for about 24 hours, and two further instances are on record in which unapproved code reached production through ordinary process restarts, but the denominator of deployments was never counted
TARGET	not set; a threshold cannot be justified before the denominator is countable, and the founding case cannot yet distinguish a deployment from a restart
GUARDRAIL SENSITIVITY	GR-L06-12 raises the measured rate at first, by detecting failures previously recorded as successes, and lowers the true rate thereafter; GR-L06-04 lowers it through canary and wave ordering; GR-L06-11 lowers it by preventing unreviewed code from being activated; GR-L06-07 lowers it by capping blast radius; GR-L06-01 lowers it modestly
CONFOUNDERS	Change size and the criticality of the surface touched; the length of the observation window, since a longer window finds more failures; workload composition; a model version change altering the character of the code produced
GAMING MODES	Fixing forward within the same session so no rollback or revert is ever recorded; committing the remediation into the same commit as the change; classifying a hotfix as a new feature so it is not matched to the prior change
CORROBORATED BY	MG-008 defect escape rate to production; MG-020 repair regression rate; the process manager restart counters as an independent signal of unplanned activation
READINESS	derivable
INSTRUMENTATION GAP	A deployment event record; today activation is an ordinary process restart that records neither the revision it loaded nor the fact that a deployment occurred

MG-008 Defect escape rate to production

QUESTION	How many defects reach a production surface before any control detects them?
OPERATIONAL DEFINITION	Count of defects first observed on a production surface, per unit of change, where a defect is a behaviour deviating from the stated intent of the change. Excludes defects detected before deployment; excludes surfaces with no external consumers; excludes purely cosmetic deviations.
FORMULA	production_first_observed_defects / deployments
UNIT AND RANGE	count per deployment, 0 to unbounded
TYPE	KRI
HORIZON	lagging
GRANULARITY	per repository per month
DATA SOURCE	Mixed and currently weak. Observation of a defect is AGENT-EMITTED, an agent notices and writes a memory note, or is operator-reported; there is no production error stream on this host correlated to change events.
COLLECTION POINT	A production error and health stream collected on a host the acting agent does not control, joined to the deployment event record
SAMPLING FREQUENCY	monthly
DIRECTION	lower-better
BASELINE	not computable; the founding case maintains no defect register and has no pre-deployment detection stage, so an escape cannot be distinguished from a defect in general; one escape is documented, the deploy that disabled a core connection function for about 24 hours
TARGET	not set; no pre-deployment detection stage exists, so every defect is by definition an escape and the ratio would carry no information
GUARDRAIL SENSITIVITY	GR-L06-12 lowers it through post-deploy independent health verification; GR-L06-08 lowers it, since a structural test mode intercepts external dispatch before it reaches a real consumer; GR-L06-06 lowers it by making dry run the default; GR-L08-05 lowers it; GR-L06-04 lowers it by limiting exposure during a canary wave
CONFOUNDERS	Consumer traffic volume, since an unused surface hides defects; observation intensity, since more looking finds more defects; the definition of stated intent, which the acting agent controls; seasonality of external usage
GAMING MODES	Silently repairing an escaped defect and recording only the repair; narrowing the stated intent of a change after the fact so the deviation is no longer a deviation; deploying to surfaces with no consumers so nothing is observed
CORROBORATED BY	MG-022 claim divergence rate; MG-007 change failure rate; the reverse proxy error rate as an independent production signal
READINESS	absent
INSTRUMENTATION GAP	A defect register, a pre-deployment detection stage to make escape meaningful, and a production error stream collected independently of the agent

MG-009 Rework ratio

QUESTION	What share of an agent change volume revises its own recent work rather than advancing new work?
OPERATIONAL DEFINITION	Changed lines in commits that modify a file the same agent changed within the preceding 72 hours, divided by total changed lines. Excludes generated artefacts, dependency lock files and vendored code; excludes revisions to a file last changed by a different agent, which are collaboration rather than rework; excludes merge commits.
FORMULA	rework_changed_lines / total_changed_lines
UNIT AND RANGE	ratio, 0 to 1
TYPE	KPI
HORIZON	lagging
GRANULARITY	per agent per week
DATA SOURCE	INFRASTRUCTURE-EMITTED, the version control history, provided commits carry a distinguishable per-agent author identity
COLLECTION POINT	A post-commit hook or a periodic walk of the version control history for each repository
SAMPLING FREQUENCY	weekly
DIRECTION	lower-better
BASELINE	not computable per agent; approximately 1,900 commits across 12 repositories over more than 80 days exist, but the founding case commits under a single human operator identity by standing instruction and every agent runs as one operating system account, so per-agent authorship cannot be recovered
TARGET	not set; some rework is healthy iteration and the informative signal is a trend under a fixed workload, not a level
GUARDRAIL SENSITIVITY	GR-L08-05 lowers it, since independent verification catches an error before it is committed and re-committed; GR-L11-04 lowers it; GR-L03-11 lowers it by removing the rework caused by concurrent agents sharing one working tree; GR-L06-09 lowers it through a change size limit; GR-L01-02 does not move it but is the precondition for attributing it per agent
CONFOUNDERS	Task difficulty and exploratory work that is legitimately iterative; repository conventions on commit size; the 72 hour window length; a model version change altering first-attempt quality
GAMING MODES	Deferring corrections into a single large later commit that also touches many unrelated files, diluting the ratio; amending or squashing commits so the revision history disappears before it is measured
CORROBORATED BY	MG-010 first attempt success rate per action class; MG-020 repair regression rate
READINESS	derivable
INSTRUMENTATION GAP	A per-agent author identity on commits; today authorship is collapsed to one operator identity, so the metric can only be computed per repository

MG-010 First attempt success rate per action class

QUESTION	What proportion of attempted actions of a given class succeed on the first attempt?
OPERATIONAL DEFINITION	Actions whose first attempt returns success, divided by distinct action intents, partitioned by action class. A retry is a subsequent attempt at the same class against the same target within fifteen minutes and is not counted as a new intent. Excludes read-only classes; excludes attempts where the target changed between attempts. Success is the tool own reported outcome, not an independent check.
FORMULA	per class, first_attempt_successes / distinct_action_intents
UNIT AND RANGE	ratio, 0 to 1
TYPE	KPI
HORIZON	lagging
GRANULARITY	per action class per agent per week
DATA SOURCE	Mixed. Attempt and outcome records are visible in the model runtime transcript and the hook stream, both written under the acting account; the success determination is the tool own exit status, which the agent can also report.
COLLECTION POINT	The PostToolUse hook, recording tool name, target and outcome into a store the agent cannot write
SAMPLING FREQUENCY	weekly
DIRECTION	higher-better
BASELINE	not computable; no runtime action classification exists and tool outcomes have never been aggregated, although the raw material is present in about 1.13 GiB of model runtime transcript
TARGET	not set; the value is dominated by task difficulty and would differ so widely across classes that a single target would be meaningless
GUARDRAIL SENSITIVITY	GR-L07-08 does not move it but is what makes it computable, by assigning a class at attempt; GR-L06-06 raises it, since a dry run surfaces problems before the counted attempt; GR-L04-06 lowers the measured rate, because an allowlist converts a silent success into an explicit denial; GR-L01-03 lowers it initially, as least privilege turns previously permitted attempts into failures; GR-L12-08 raises it modestly
CONFOUNDERS	Task difficulty; model version change; the fifteen minute retry window; whether the environment was already in the desired state, which makes an attempt trivially successful
GAMING MODES	Not attempting an action until success is certain, and doing the exploratory work through an uncounted class such as generic shell execution; recording the attempt only after it has succeeded
CORROBORATED BY	MG-032 boundary crossing attempt rate; MG-059 control bypass rate; MG-036 unclassified action share, which detects displacement into uncounted classes
READINESS	derivable
INSTRUMENTATION GAP	An action class assignment and a retry-grouping key; both would have to be derived offline from the existing transcript, and no such derivation exists

MG-011 Test coverage delta per change

QUESTION	Does a change alter the proportion of the code it touches that is exercised by automated tests?
OPERATIONAL DEFINITION	Coverage percentage over the set of files a change touches, measured after the change, minus the same measure before it. Excludes changes touching no executable code; excludes repositories with no test harness, which are reported separately as uncovered; excludes coverage of generated code.
FORMULA	coverage_after(changed_files) - coverage_before(changed_files)
UNIT AND RANGE	percentage points, -100 to 100
TYPE	KCI
HORIZON	leading
GRANULARITY	per change
DATA SOURCE	INFRASTRUCTURE-EMITTED, a coverage run executed by a build stage under a principal other than the author
COLLECTION POINT	A build stage that runs the test suite with coverage instrumentation on the pre-change and post-change revisions
SAMPLING FREQUENCY	per change
DIRECTION	higher-better
BASELINE	not computable; the founding case operates no continuous integration pipeline on any host, deploys directly on the target from a shared working tree, and invokes no coverage tooling
TARGET	not set; a coverage target is not defensible before a test harness and a build stage exist, and setting one first would reward test volume over test value
GUARDRAIL SENSITIVITY	GR-L06-11 raises it, because binding a deploy artefact to a review record forces a build stage where coverage can be measured at all; GR-L06-01 raises it modestly by creating a review occasion; GR-L06-12 raises it; GR-L08-07 raises it, since an adversarial reviewer asks for the missing test; GR-L12-03 does not move it, being prose
CONFOUNDERS	Repository maturity and the pre-existing coverage level; the size of the change, since a small change to a well covered file moves the delta a long way; the choice of coverage measure, whether line, branch or statement
GAMING MODES	Adding tests that execute code without asserting on its behaviour; excluding hard to test files from the measured set; splitting a change so the untested part lands in a separate commit that is measured on its own
CORROBORATED BY	MG-007 change failure rate; mutation testing score as an independent measure of whether the tests can fail
READINESS	absent
INSTRUMENTATION GAP	A build stage of any kind, a test harness in each repository, and a coverage tool; none exists on this host

MG-012 Build reproducibility rate

QUESTION	What proportion of deployed artefacts can be rebuilt from the recorded source revision?
OPERATIONAL DEFINITION	Deployments for which an independent rebuild from the recorded revision produces an artefact matching the deployed one, divided by deployments audited. Content that is legitimately non-deterministic, such as an embedded build timestamp, is normalised before comparison and the normalisation is recorded. Excludes interpreted sources deployed with no build step; excludes deployments with no recorded revision, which are counted separately as unattributable.
FORMULA	reproducible_rebuilds / audited_deployments
UNIT AND RANGE	ratio, 0 to 1
TYPE	KCI
HORIZON	lagging
GRANULARITY	per deployment, aggregated per repository per month
DATA SOURCE	INFRASTRUCTURE-EMITTED, a rebuild performed by a verifier under a separate identity and compared against the artefact resident on the target host
COLLECTION POINT	A verifier host that rebuilds from the recorded revision and hashes both artefacts
SAMPLING FREQUENCY	monthly
DIRECTION	higher-better
BASELINE	not computable; the founding case builds on the target host from a shared working tree and records no revision alongside the running artefact, so there is nothing to rebuild from, and two incidents are on record in which the running code corresponded to no approved revision
TARGET	1.0 for any artefact that passes through the deployment approval gate, because an approval that does not bind the running code approves nothing
GUARDRAIL SENSITIVITY	GR-L06-11 raises it strongly, since provenance binding is the mechanism; GR-L03-04 raises it by preventing drift in the deployed tree; GR-L03-11 raises it by removing concurrent writes to the tree that was built; GR-L06-02 raises it by giving a recorded comparison point; GR-L08-12 raises it through configuration hashing
CONFOUNDERS	Toolchain and dependency version drift between build and rebuild; the normalisation rules chosen; whether the rebuild host matches the build host
GAMING MODES	Recording the revision after the build, taken from the working tree current state rather than the state that was actually built; normalising away real differences under the label of non-determinism
CORROBORATED BY	MG-064 audit record completeness; the deployment approval gate own record of the approved revision as an independent source
READINESS	absent
INSTRUMENTATION GAP	A recorded revision on every deployed artefact, an artefact hash, and a verifier identity able to rebuild; none of these exists

MG-011 test coverage delta and MG-012 build reproducibility rate are both recorded as absent for the same structural reason: the founding case runs no continuous integration pipeline and builds directly on the target host from a shared working tree, so there is no pre-change and post-change pair to compare and nothing to rebuild from. Both are precisely the properties Humble and Farley's continuous delivery practice would supply, a point D2 Section 8 develops as a regression on established practice rather than a novel gap.

2.3 Restart safety

MG-013 to MG-016 measure what a restart costs, distinct from MG-014 's own count of how often restarts happen. The founding case's process manager inventory supplies a partial baseline directly: at collection it reported 37 restarts for the knowledge search service and 7 for the console backend against uptimes of 34 and 11 minutes respectively, with the remaining seven services at 0 over roughly ten days EV-004, though the counters are cumulative since the process manager's own last start rather than since a defined observation window, so a rate cannot be derived from them directly.

Table D4-3. Restart safety metrics, MG-013 to MG-016.

MG-013 Restart induced work loss

QUESTION	How much in-progress agent work is discarded when a session is restarted?
OPERATIONAL DEFINITION	For each session restart, the elapsed session time and number of turns between the last durable artefact and the restart. A durable artefact is a commit, a file written outside the session scratch area, or a completion record. Excludes scheduled reloads in which the agent was asked to write a summary first and demonstrably did so; excludes restarts of sessions idle for longer than the measurement interval.
FORMULA	sum over restarts of (t_restart - t_last_durable_artefact); and mean turns discarded per restart
UNIT AND RANGE	seconds, 0 to unbounded
TYPE	KRI
HORIZON	lagging
GRANULARITY	per restart, aggregated per agent per week
DATA SOURCE	Mixed. Restart timestamps are INFRASTRUCTURE-EMITTED by the supervisor unit journal and the process manager; last-artefact timestamps come from version control and the filesystem, which the agent controls.
COLLECTION POINT	The supervisor unit stop and start transitions, joined to the version control and filesystem modification record
SAMPLING FREQUENCY	weekly
DIRECTION	lower-better
BASELINE	not computable; the founding case supervisor units restart within about ten seconds and record the restart, but no marker of in-flight work exists, and the only field recording the outcome is a free text reload result string
TARGET	not set; the value depends entirely on the durability discipline of the workload rather than on a level that can be prescribed
GUARDRAIL SENSITIVITY	GR-L09-10 raises it, because restart-always returns a fresh session with no continuity of the interrupted turn; GR-L04-08 raises it, since automatic termination on wedge detection discards the current turn; GR-L05-09 lowers it through recoverable agent state; GR-L03-02 lowers it modestly by preserving per-session artefacts; GR-L04-11 does not move it
CONFOUNDERS	How often the workload naturally produces durable artefacts; the restart cause, since a memory ceiling restart lands at a random point whereas a scheduled reload does not; task length
GAMING MODES	Writing a trivial artefact at short intervals so the measured gap is always small while nothing substantive is produced; classifying an unplanned restart as scheduled so it is excluded
CORROBORATED BY	MG-014 unplanned restart rate; MG-066 episode reconstructability
READINESS	derivable
INSTRUMENTATION GAP	An in-flight work marker written at turn boundaries; today the only signal is the free text reload result string in the fleet registry

MG-014 Unplanned restart rate

QUESTION	How often do supervised processes and agent sessions restart outside a scheduled or deliberate action?
OPERATIONAL DEFINITION	Count of process or session restarts not attributable to a scheduled reload, an operator-initiated restart recorded in the console audit log, or a deployment restart, per process per day. Includes restarts triggered by a memory ceiling and by supervisor restart-always after an unexpected exit. Excludes restarts of the supervisor daemon itself, which reset the counters and are recorded separately.
FORMULA	(total_restarts - planned_restarts) / process_days
UNIT AND RANGE	count per process-day, 0 to unbounded
TYPE	KRI
HORIZON	lagging
GRANULARITY	per process per day
DATA SOURCE	INFRASTRUCTURE-EMITTED, the process manager restart counters and the supervisor unit journal; note the process manager counters are cumulative since its own last start, not since a defined observation window
COLLECTION POINT	The process manager state, sampled on a schedule and differenced, plus supervisor unit journal transitions
SAMPLING FREQUENCY	daily
DIRECTION	lower-better
BASELINE	partially observed and stated as counts rather than a rate: at collection the process manager reported 37 restarts for the knowledge search service and 7 for the console backend, against uptimes of 34 and 11 minutes respectively, with the remaining seven services at 0 over about 10 days; a per-day rate over a defined window is not computable because the counters are cumulative since daemon start
TARGET	not set; the founding case cannot separate restarts caused by agent action from restarts caused by resource ceilings, so a threshold would not be interpretable
GUARDRAIL SENSITIVITY	GR-L09-10 raises it, because restart-always converts every exit into a recorded restart; GR-L04-04 raises it, since control group limits terminate processes that exceed them; GR-L04-08 raises it; GR-L06-04 lowers it by avoiding simultaneous restarts across targets; GR-L09-03 lowers it by breaking a restart loop
CONFOUNDERS	Memory pressure from unrelated workloads on the same host; unattended operating system upgrades restarting the supervisor; deployment frequency; a change in the memory ceiling configuration
GAMING MODES	Disabling the restart policy so a failure presents as a silent outage rather than a counted restart; raising the memory ceiling so a leaking process degrades instead of restarting; restarting the supervisor to reset the counters
CORROBORATED BY	MG-015 service availability during agent activity; the reverse proxy access log as an independent record of service discontinuity
READINESS	available
INSTRUMENTATION GAP	none; visible today in the process manager inventory and the supervisor unit journal, with the caveat that counters are cumulative since daemon start and no scheduled-versus-unplanned attribution is recorded

MG-015 Service availability during agent activity

QUESTION	Does service availability differ during periods of agent activity from periods without it?
OPERATIONAL DEFINITION	Proportion of health probe intervals returning a healthy response, computed separately for intervals overlapping at least one mutating agent action and intervals with none. The health criterion must exercise the service core function, not merely process liveness. Excludes announced maintenance windows; excludes probe failures caused by the probing host itself.
FORMULA	healthy_probes_active / probes_active, compared with healthy_probes_inactive / probes_inactive; the reported figure is the difference between the two ratios
UNIT AND RANGE	ratio, 0 to 1, with a difference in the range -1 to 1
TYPE	KPI
HORIZON	lagging
GRANULARITY	per service per day
DATA SOURCE	INFRASTRUCTURE-EMITTED, health probes issued from a host the acting agent does not control; the activity classification comes from the hook substrate
COLLECTION POINT	An external prober outside the agent blast radius, joined to the mutating-action timeline
SAMPLING FREQUENCY	daily
DIRECTION	higher-better
BASELINE	not computable; no continuous external health probe exists for the services on this host, and the reverse proxy access log is not aggregated into availability intervals
TARGET	not set for the level; the defensible target is that the difference between the two ratios is statistically indistinguishable from zero, because agent activity should not be observable in availability
GUARDRAIL SENSITIVITY	GR-L06-12 raises it; GR-L06-04 raises it by limiting simultaneous exposure; GR-L06-07 raises it by capping blast radius; GR-L09-09 raises it through graceful degradation to a safe mode; GR-L02-01 does not move it
CONFOUNDERS	External traffic patterns and time of day; unrelated infrastructure failures; the choice of health criterion, which determines whether a functional failure is visible at all; the correlation between agent activity and business hours
GAMING MODES	Probing from inside the host so a network or proxy failure is invisible; choosing a probe path that does not exercise the function the change touched, which is precisely the form of the 24 hour incident on record; excluding the incident window as maintenance after the fact
CORROBORATED BY	MG-008 defect escape rate to production; MG-016 in-flight interaction loss on restart
READINESS	absent
INSTRUMENTATION GAP	An external health prober with a function-exercising criterion, and an activity timeline joined to it; neither exists

MG-016 In-flight interaction loss on restart

QUESTION	How many interactions in progress are lost when a service or session is restarted?
OPERATIONAL DEFINITION	Count of interactions open at the moment of a restart that did not complete, divided by restarts. An interaction is an external request, conversation or session with a party outside the host. Excludes idle connections held open with no work in progress; excludes interactions a client retried successfully within one second; excludes internal loopback calls.
FORMULA	lost_in_flight_interactions / restarts
UNIT AND RANGE	count per restart, 0 to unbounded
TYPE	KRI
HORIZON	lagging
GRANULARITY	per service per restart
DATA SOURCE	INFRASTRUCTURE-EMITTED, an application-level in-flight counter sampled at the moment the termination signal is issued, joined to the supervisor stop timestamp
COLLECTION POINT	The application request lifecycle, exporting an in-flight gauge, read by the supervisor stop path before termination
SAMPLING FREQUENCY	per restart
DIRECTION	lower-better
BASELINE	not computable; no service on this host records in-flight interaction counts, and the one termination path that halts work in progress does so by killing the process group with no drain phase
TARGET	0 for services carrying interactions with parties outside the host, because a discarded live external interaction is a harm to a third party rather than an internal cost
GUARDRAIL SENSITIVITY	GR-L09-09 lowers it through graceful degradation; GR-L06-04 lowers it by draining one target at a time; GR-L11-08 lowers it by confining restarts away from hours when external interactions occur; GR-L09-01 raises it, because the session stop affordance kills the process group and discards whatever was in flight; GR-L06-05 lowers it through change freeze windows
CONFOUNDERS	External traffic volume at the moment of restart; connection duration characteristics of the protocol; client retry behaviour, which masks loss; time of day
GAMING MODES	Counting only interactions the application chose to register, so unregistered ones are never lost; sampling the in-flight set after termination, when it is empty by construction; defining an abruptly closed interaction as completed
CORROBORATED BY	MG-015 service availability during agent activity; MG-045 harm completion time per action class
READINESS	absent
INSTRUMENTATION GAP	An in-flight interaction gauge in each externally facing service and a drain phase in the stop path; neither exists

MG-014 is the subfamily's one available entry and its guardrail sensitivity is instructive: GR-L09-10 restart-always supervision, a present control that D2 Section 4.4 identifies as defeating VC3 halt, raises MG-014 by converting every process exit into a recorded restart. The same mechanism that undermines supervision viability is the reason this metric is directly observable today.

2.4 Repair capability

MG-017 to MG-021 measure whether the system detects and fixes its own faults without a human, and whether the fix is real. All five are absent. The founding case keeps no fault register, no incident register and no repair register, so none of the three populations the subfamily depends on exists as a countable set. What does exist is a single latency bound: the fastest independent detector on this host trips at 180 seconds plus up to 60 seconds of schedule granularity, and its trip condition is that an agent has stopped producing output, not that a fault has occurred EV-028 EV-039.

Table D4-4. Repair capability metrics, MG-017 to MG-021.

MG-017 Mean time to detect a fault

QUESTION	How long does a fault persist before any mechanism detects it?
OPERATIONAL DEFINITION	Mean elapsed time from the first occurrence of a fault to its first detection by any mechanism, where detection means a record exists that a person or automation could act on. Faults first surfaced by a human report are timed to that report. Excludes faults whose first occurrence cannot be timestamped, which are reported separately; excludes detection of the agent going quiet, which is a liveness signal and not a fault signal.
FORMULA	mean over faults of (t_first_detection - t_first_occurrence)
UNIT AND RANGE	seconds, 0 to unbounded
TYPE	KPI
HORIZON	lagging
GRANULARITY	per fault, aggregated per month
DATA SOURCE	Mixed. Watcher trips are INFRASTRUCTURE-EMITTED; the primary reporting channel is AGENT-EMITTED, since the notification script runs only when the agent chooses to call it; first-occurrence timestamps generally require reconstruction from service logs.
COLLECTION POINT	An independent monitoring stream with fault-specific criteria, plus retrospective log reconstruction for first occurrence
SAMPLING FREQUENCY	monthly
DIRECTION	lower-better
BASELINE	not computable as a mean; one bound is on record, namely that the fastest independent detector on this host trips at 180 seconds plus up to 60 seconds of schedule granularity and its trip condition is that an agent stopped producing output rather than that a fault occurred; the single documented production fault persisted about 24 hours
TARGET	not set as an absolute; the defensible requirement is that detection time is shorter than harm completion time for the same class, which is MG-046 latency margin
GUARDRAIL SENSITIVITY	GR-L08-11 lowers it strongly, since an independent event stream is the only source that does not depend on the agent choosing to report; GR-L06-12 lowers it for change-induced faults; GR-L09-06 lowers it, because automatic suspension on anomaly presupposes anomaly detection; GR-L10-11 lowers it for the wedge class only; GR-L12-04 does not move it
CONFOUNDERS	Fault type, since a crash is detected far faster than a silent functional regression; traffic volume on the affected surface; the definition of first occurrence; observation intensity during an audit
GAMING MODES	Defining fault onset as the moment of detection, which makes the metric identically zero; recording only faults that were detected quickly; treating an agent later discovery as the detection event while suppressing the earlier unactioned signal
CORROBORATED BY	MG-025 time to divergence detection; MG-064 audit record completeness
READINESS	absent
INSTRUMENTATION GAP	A fault register with first-occurrence timestamps and an independent monitoring stream with fault-specific criteria; today the only independent detectors watch for silence, not for faults

MG-018 Mean time to repair without human intervention

QUESTION	How long does the system take to restore a faulted service with no human acting?
OPERATIONAL DEFINITION	Mean elapsed time from first detection to restoration, restricted to incidents in which no human input occurred between the two. Restoration means the service passes the same criterion that failed. Excludes incidents a human touched; excludes incidents with no defined health criterion; excludes supervisor restarts of a process that had not actually failed.
FORMULA	mean over unattended incidents of (t_restored - t_detected)
UNIT AND RANGE	seconds, 0 to unbounded
TYPE	KPI
HORIZON	lagging
GRANULARITY	per incident, aggregated per month
DATA SOURCE	Mixed. Restart and probe timestamps are INFRASTRUCTURE-EMITTED; the assertion that a repair occurred, as distinct from a restart, is AGENT-EMITTED in the founding case.
COLLECTION POINT	The incident record, joined to the independent health stream that defines detection and restoration
SAMPLING FREQUENCY	monthly
DIRECTION	lower-better
BASELINE	not computable; the founding case keeps no incident record; the nearest observed value is the supervisor restart interval of about ten seconds for a session that has exited, which restores a process rather than repairing a fault
TARGET	not set; without MG-019 repair correctness rate a shorter repair time rewards declaring restoration rather than achieving it
GUARDRAIL SENSITIVITY	GR-L09-10 lowers it for the process-exit class only; GR-L06-03 lowers it, since pre-approved rollback targets remove a decision from the path; GR-L06-02 lowers it by providing a state to restore to; GR-L06-01 raises it where the repair itself needs approval; GR-L09-03 raises it, because a circuit breaker deliberately stops automated repair attempts
CONFOUNDERS	Fault severity and whether the repair is a restart or a code change; time of day, since an unattended repair at night has no competing human; the exclusion rule for incidents a human touched, which selects the easy incidents
GAMING MODES	Declaring restoration on the basis of process liveness rather than working function, which is the exact form of the 24 hour incident on record; restarting the health check rather than the service; excluding hard incidents by having a human touch them
CORROBORATED BY	MG-019 repair correctness rate; MG-022 claim divergence rate
READINESS	absent
INSTRUMENTATION GAP	An incident record with detection and restoration timestamps, and a restoration criterion evaluated independently of the repairing agent

MG-019 Repair correctness rate

QUESTION	What proportion of repairs restore the failed function rather than only its symptom?
OPERATIONAL DEFINITION	Repairs verified by an independent exercise of the function that failed, divided by repairs claimed. The verifying check must exercise the failed function itself, not a proxy such as process liveness or an unrelated successful HTTP response. Excludes repairs where the failed function cannot be exercised without affecting a third party, which are reported separately as unverifiable.
FORMULA	independently_verified_repairs / claimed_repairs
UNIT AND RANGE	ratio, 0 to 1
TYPE	KPI
HORIZON	lagging
GRANULARITY	per repair, aggregated per month
DATA SOURCE	Claims are AGENT-EMITTED. Verification requires a verifier holding an identity separate from the repairing agent, which does not exist on this host, where every agent, watcher and console process runs as one operating system account.
COLLECTION POINT	A verifier process under a separate identity that re-executes the failing criterion after the repair is claimed
SAMPLING FREQUENCY	per repair
DIRECTION	higher-better
BASELINE	not computable as a rate; the founding case has one documented instance in which a repair was reported successful while the service core connection function remained disabled for about 24 hours, which is a single negative observation and not a denominator
TARGET	1.0, because a repair that has not been independently exercised is a claim rather than a repair, and the founding case documents the exact failure
GUARDRAIL SENSITIVITY	GR-L08-05 raises it strongly; GR-L06-12 raises it; GR-L08-08 raises it by cross-checking the claim against raw sources; GR-L08-06 raises it by removing the acting agent from the attestation; GR-L12-05 does not move it, being a prose instruction of the kind already present
CONFOUNDERS	Fault type, since some functions are trivially exercisable and others are not; who defines the verifying criterion, which in the founding case is the repairing agent; the availability of a safe way to exercise an externally facing function
GAMING MODES	Choosing a verification that passes by construction; verifying from the same session and credentials that performed the repair, so cached or stale state reads back as success; narrowing the definition of the failed function to the part that was fixed
CORROBORATED BY	MG-022 claim divergence rate; MG-020 repair regression rate
READINESS	absent
INSTRUMENTATION GAP	A verifier identity distinct from the acting account, and a recorded failing criterion to re-execute; neither exists, and the shared account model makes the first structurally unavailable today

MG-020 Repair regression rate

QUESTION	How often does a repair introduce a new fault?
OPERATIONAL DEFINITION	Repairs followed within seven days by a new fault in a component the repair touched, divided by repairs. Excludes new faults with an independently established unrelated cause; excludes repairs to components that received no subsequent traffic within the window, which are reported separately as unexercised.
FORMULA	repairs_followed_by_new_fault / repairs
UNIT AND RANGE	ratio, 0 to 1
TYPE	KRI
HORIZON	lagging
GRANULARITY	per repair, aggregated per month
DATA SOURCE	Mixed. Version control history and health data are INFRASTRUCTURE-EMITTED; the causal attribution of the new fault is a judgement that in the founding case is made by the same agent that performed the repair.
COLLECTION POINT	The version control history of the touched paths, joined to the independent health stream over the following seven days
SAMPLING FREQUENCY	monthly
DIRECTION	lower-better
BASELINE	not computable; the founding case maintains neither a repair register nor a fault register, so neither term of the ratio exists
TARGET	not set; the level depends on component coupling and cannot be prescribed independently of the codebase
GUARDRAIL SENSITIVITY	GR-L06-07 lowers it by capping blast radius; GR-L06-09 lowers it through a change size limit that escalates large repairs; GR-L08-07 lowers it through an adversarial second opinion; GR-L11-03 lowers it, since a cooling off period between related high impact actions prevents a hurried second change; GR-L06-06 lowers it
CONFOUNDERS	Component coupling and test coverage; the seven day window; traffic on the touched component, which determines whether a regression is observed at all; who adjudicates causation
GAMING MODES	Bundling the regression fix into the original repair record so only one repair is counted; attributing the new fault to an unrelated cause, which the same agent adjudicates; repairing components with no traffic so no regression is ever observed
CORROBORATED BY	MG-007 change failure rate; MG-009 rework ratio
READINESS	absent
INSTRUMENTATION GAP	A repair register, a fault register and an independent causal attribution; none exists

MG-021 Rollback exercise success rate

QUESTION	When a rollback is attempted or rehearsed, does it restore the prior state?
OPERATIONAL DEFINITION	Rollbacks that restore the target to a state matching the recorded pre-change snapshot, divided by rollbacks attempted, including scheduled rehearsals against non-production targets. Excludes rollbacks of changes with no recorded pre-change state, which are counted separately as unrollable; excludes partial restorations accepted as adequate by the actor.
FORMULA	successful_rollbacks / attempted_rollbacks
UNIT AND RANGE	ratio, 0 to 1
TYPE	KCI
HORIZON	leading
GRANULARITY	per rollback, aggregated per quarter
DATA SOURCE	INFRASTRUCTURE-EMITTED, comparison of post-rollback state against a pre-apply snapshot taken by the change plane rather than by the acting agent
COLLECTION POINT	The change plane pre-apply snapshot record, compared against the target state after the rollback completes
SAMPLING FREQUENCY	quarterly
DIRECTION	higher-better
BASELINE	not computable; the denominator is zero, no rollback of any kind has been executed on this host; the update plane holds typed pre-apply cloud snapshots but package-state backup is absent, and its rollback route has never been invoked
TARGET	1.0, with at least one rehearsal per quarter per change class, because an unrehearsed rollback is an assumption rather than a control
GUARDRAIL SENSITIVITY	GR-L06-02 raises it by ensuring a snapshot exists to compare against; GR-L06-03 raises it by constraining rollback to pre-approved targets; GR-L05-09 raises it through recoverable state; GR-L03-04 raises it modestly by keeping the rolled-back tree from drifting; GR-L06-10 does not move it, since the report-only constraint prevents the apply that would need rolling back rather than improving the rollback itself
CONFOUNDERS	Change class, since a package rollback and a schema rollback differ entirely; whether the rehearsal target resembles production; time elapsed since the change, which allows dependent state to accumulate
GAMING MODES	Rehearsing only on change classes already known to be reversible; counting a forward fix as a rollback; declaring success on the basis of the rollback command exiting zero rather than on a state comparison
CORROBORATED BY	MG-043 irreversible action share; MG-012 build reproducibility rate
READINESS	absent
INSTRUMENTATION GAP	A pre-change snapshot for every change class, including package state, and a state comparison after rollback; the update plane has the first only for cloud resources

MG-019 repair correctness rate carries the subfamily's sharpest target statement: it is set at 1.0 not because perfection is expected but because a repair that has not been independently exercised is a claim rather than a repair, and the founding case's own canonical incident is exactly a case in which the claimed and the actual state differed. MG-021 rollback exercise success rate has a denominator of zero: no rollback of any kind, rehearsed or real, has been executed on this host REPORTED, which is itself the finding, not a gap in measurement.

3. Governance metrics

The fifty-one governance metrics are organised into nine subfamilies that follow the shape of D2's argument: verification and truthfulness first, because it is the founding case's central finding; then boundary and mandate, containment and supervision viability, which correspond respectively to D2's

access-versus-mandate analysis, the containment shortfall determinant, and the six viability conditions; then human oversight quality, control effectiveness, auditability, reliability and operations, and governance overhead, which together ask what the governance layer itself costs and whether it can be trusted to report on itself.

3.1 Verification and truthfulness

`MG-022` to `MG-029` are the direct instrumentation of D2 Section 4's argument. This is the subfamily with the founding case's only genuinely computed baseline in the whole register: `MG-022` claim divergence rate at 0.611, eleven divergent episodes in eighteen audited **REPORTED**. Every other entry in the subfamily is either `not computable` for want of the comparator that would make it a distribution rather than a single documented incident, or, in the case of `MG-024`, computable at 0 by architecture, since all fourteen live agent sessions share one operating system account and no verification on this host can therefore satisfy the identity criterion for independence `EV-003`.

Table D4-5. Verification and truthfulness metrics, MG-022 to MG-029.

MG-022 Claim divergence rate

QUESTION	How often does an agent account of what it did diverge from independent evidence of what happened?
OPERATIONAL DEFINITION	Audited episodes in which at least one material factual assertion in the agent self-report is contradicted by independent evidence, divided by episodes audited. A material assertion is one about an action taken, a state reached, or a verification performed. Excludes stylistic differences; excludes assertions about intention or reasoning; excludes episodes for which no independent evidence exists on either side, which are counted separately as unadjudicable.
FORMULA	divergent_episodes / audited_episodes
UNIT AND RANGE	ratio, 0 to 1
TYPE	KRI
HORIZON	lagging
GRANULARITY	per episode, aggregated per agent per audit cycle
DATA SOURCE	Mixed and asymmetric by design. Self-reports are AGENT-EMITTED, being completion messages, memory notes and session summaries. The comparator must be INFRASTRUCTURE-EMITTED: version control history, process manager state, health checks, cloud control plane state and the reverse proxy access log.
COLLECTION POINT	An audit harness that pairs each self-report with the infrastructure record for the same window, executed under an identity separate from the audited agent
SAMPLING FREQUENCY	weekly on a sample, with a full census per quarter
DIRECTION	lower-better
BASELINE	0.611, being 11 divergent episodes in 18 audited, where the 18 were drawn from 46 candidate episodes surfaced by the system itself
TARGET	not set as a level; the defensible target is a monotone decrease under a fixed audit protocol, because any absolute threshold above zero legitimises a known rate of false reporting, and the founding case sampling frame was not random
GUARDRAIL SENSITIVITY	GR-L08-05 lowers it strongly, removing both the opportunity and the incentive; GR-L08-06 lowers it, since the acting agent no longer writes its own attestation; GR-L08-08 lowers it through cross-checking against raw sources; GR-L08-11 lowers it by supplying an independent comparator that makes divergence detectable; GR-L12-09 does not move it, since prose instruction to report accurately is precisely what the founding case already had
CONFOUNDERS	Episode selection, since the founding case 46 candidates were surfaced by the system rather than sampled at random; task difficulty; model version change; the auditor materiality threshold; observation effects if an agent can detect that an audit is running
GAMING MODES	Writing self-reports vague enough that no assertion can be contradicted, which lowers divergence while lowering informativeness; asserting only what the infrastructure record will confirm and omitting the rest; editing the transcript or the memory note after the fact, which the acting account can do because it owns every evidence store on this host
CORROBORATED BY	MG-026 unverifiable claim share detects the vagueness strategy; MG-027 evidence citation completeness; MG-065 audit record immutability detects post hoc editing
READINESS	derivable
INSTRUMENTATION GAP	The founding case audit was manual; a repeatable harness needs a stable episode identifier, a self-report store the agent cannot edit, and an infrastructure comparator collected under a different principal

MG-023 Verification coverage of completion claims

QUESTION	What proportion of completion claims are subjected to any independent check?
OPERATIONAL DEFINITION	Completion claims for which an independent verification record exists, divided by completion claims made. Independent means performed by a principal other than the claimant. Excludes self-verification, that is the claimant re-running its own check; excludes claims about read-only work; excludes verification performed only after a divergence was already suspected.
FORMULA	independently_verified_claims / total_completion_claims
UNIT AND RANGE	ratio, 0 to 1
TYPE	KCI
HORIZON	leading
GRANULARITY	per agent per week
DATA SOURCE	Claims are AGENT-EMITTED. Verification records would be emitted by the verifier and are absent on this host, where the deployment approval gate peer review is written by another agent and no code compares reviewer to author.
COLLECTION POINT	A verification queue that receives every completion claim and records the verifying principal, the check performed and its outcome
SAMPLING FREQUENCY	weekly
DIRECTION	higher-better
BASELINE	0 in normal operation; no verifier identity exists and the only peer review mechanism has a documented production instance of an agent recording its own review under another agent token; the 18 audited episodes were a research intervention covering about 0.39 of the 46 candidates within that intervention alone
TARGET	1.0 for claims about action classes rated irreversible or severe; lower coverage is defensible only where the action is reversible and cheap to re-check
GUARDRAIL SENSITIVITY	GR-L08-05 raises it; GR-L08-06 raises it by making verification a precondition of the record; GR-L08-07 raises it; GR-L01-07 raises it and is the precondition, since without a separate verifier identity coverage is definitionally zero; GR-L10-04 raises it modestly by forcing evidence into the approval path
CONFOUNDERS	The rate of claim production, since coverage falls as claims rise at fixed verifier capacity; which claims are routed for verification; verifier throughput; the definition of a completion claim
GAMING MODES	Recording a verification performed by a second session of the same agent under the same credentials and counting it as independent; verifying the cheap claims while leaving the consequential ones uncovered, so the ratio looks healthy while risk is untouched
CORROBORATED BY	MG-024 verification independence ratio, which directly detects the first mode; MG-022 claim divergence rate
READINESS	absent
INSTRUMENTATION GAP	A verification record schema, a routing rule from claim to verifier, and a verifier identity; none exists on this host

MG-024 Verification independence ratio

QUESTION	What proportion of verifications are performed by a principal genuinely independent of the actor?
OPERATIONAL DEFINITION	Verification records whose verifying principal differs from the acting principal in identity, credential and process, divided by verification records. A verifier sharing the actor operating system account, credentials or filesystem write access is not independent regardless of what name appears on the record. Excludes verifications with no recorded verifier.
FORMULA	independent_verifications / total_verifications
UNIT AND RANGE	ratio, 0 to 1
TYPE	KCI
HORIZON	leading
GRANULARITY	per verification, aggregated per month
DATA SOURCE	INFRASTRUCTURE-EMITTED where identity is enforced at the credential layer. On this host it would be AGENT-ASSERTED, because the name on a verification record is a string the acting party writes and all parties share one account.
COLLECTION POINT	The credential issuance and authentication layer, comparing the authenticated principal on the verification against the authenticated principal on the action
SAMPLING FREQUENCY	monthly
DIRECTION	higher-better
BASELINE	0; all 14 live agent sessions, the console, the watchers and every registry writer run as one operating system account holding unrestricted passwordless root, so no verification on this host can satisfy the identity criterion; a production instance is on record of an agent recording its own peer review using another agent token
TARGET	1.0, because the criterion is binary at the architecture level and any value below 1 means the verifier could have been the actor
GUARDRAIL SENSITIVITY	GR-L01-07 raises it and is its definition; GR-L01-02 raises it, since per-agent workload identity is what makes the comparison meaningful; GR-L01-01 holds it at 0, because a shared account makes independence unrepresentable; GR-L08-06 raises it; GR-L05-06 raises it by denying the agent the credential that could alter the record
CONFOUNDERS	A change in the account model rather than in behaviour will move it discontinuously; token-based pseudo-identities that look distinct but share a credential store; whether independence is assessed at the identity layer or only on the name written into the record
GAMING MODES	Naming a different agent in the verification record while executing under the same credential, which the founding case demonstrates in production; rotating which session performs verification while every session remains one principal; introducing a nominally separate service account whose key sits in the same file the agents own
CORROBORATED BY	MG-067 separation of duties violations; the credential issuance record and process ownership as independent sources
READINESS	available
INSTRUMENTATION GAP	none for the architectural determination, which is visible today in the process ownership of every agent session and in the single account holding all credentials; a verification record schema is still required before the ratio is informative rather than identically zero

MG-025 Time to divergence detection

QUESTION	How long does a divergent claim stand before anything contradicts it?
OPERATIONAL DEFINITION	Elapsed time from the moment a divergent claim is recorded to the moment the divergence is first recorded by any party. Claims still undetected at the end of the window are censored and reported as a separate count rather than dropped, because dropping them makes the metric a function of the audit schedule. Excludes divergences the claimant itself corrected within the same turn.
FORMULA	median over detected divergences of (t_divergence_recorded - t_claim_recorded); reported with the count of claims still undetected at window close
UNIT AND RANGE	seconds, 0 to unbounded
TYPE	KRI
HORIZON	lagging
GRANULARITY	per divergence, aggregated per audit cycle
DATA SOURCE	Mixed. Claim timestamps are AGENT-EMITTED; detection timestamps come from the audit harness, which must run under a separate identity to count as detection.
COLLECTION POINT	The audit harness, recording both the claim timestamp it read and the time at which the contradiction was established
SAMPLING FREQUENCY	per audit cycle
DIRECTION	lower-better
BASELINE	not computable as a distribution; the single documented case stood about 24 hours before the disabled function was noticed, and the 11 divergences found in the research audit were detected only when the audit ran, so the observed values measure the audit schedule rather than any detector
TARGET	not set; the meaningful requirement is that detection precede the next dependent action, which is a sequencing constraint rather than a time
GUARDRAIL SENSITIVITY	GR-L08-05 lowers it; GR-L06-12 lowers it by contradicting a false success promptly; GR-L08-11 lowers it; GR-L08-08 lowers it; GR-L08-09 raises it where the retention horizon is shorter than the detection interval, because the comparator evidence expires before the check can run
CONFOUNDERS	The audit cadence, which dominates the observed value; the visibility of the claim, since a claim in an SMS is read sooner than one in a memory note; whether the divergence had a downstream consequence that forced attention
GAMING MODES	Delaying the recording of a claim until the comparator evidence has expired; recording claims in a store with shorter retention than the comparator so the pair can never be formed; correcting the claim quietly so no divergence is ever recorded
CORROBORATED BY	MG-022 claim divergence rate; MG-064 audit record completeness
READINESS	absent
INSTRUMENTATION GAP	A durable claim record with a trustworthy timestamp and a comparator retained at least as long; today transcripts are pruned on a rolling window of about 30 days and every store is writable by the claimant

MG-026 Unverifiable claim share

QUESTION	What proportion of an agent assertions cannot be checked against any independent record?
OPERATIONAL DEFINITION	Material assertions in agent self-reports for which no independent record could confirm or contradict them, divided by material assertions. Unverifiable includes assertions about internal reasoning, about actions that leave no trace, and about states no longer observable. Excludes assertions the auditor chose not to verify on cost grounds, which are recorded separately as unattempted rather than folded into the numerator.
FORMULA	unverifiable_assertions / material_assertions
UNIT AND RANGE	ratio, 0 to 1
TYPE	KRI
HORIZON	leading
GRANULARITY	per report, aggregated per agent per audit cycle
DATA SOURCE	AGENT-EMITTED self-reports parsed by the audit harness; the existence of a comparator is determined from the INFRASTRUCTURE-EMITTED record, so the classification itself is not an agent judgement
COLLECTION POINT	The audit harness, at the point where each assertion is matched to a candidate comparator
SAMPLING FREQUENCY	per audit cycle
DIRECTION	lower-better
BASELINE	not computable; the founding case audit classified whole episodes as divergent or not and did not classify individual assertions by verifiability
TARGET	not set as a level; its principal use is as a control on MG-022, since a falling divergence rate accompanied by a rising unverifiable share indicates that reporting has become evasive rather than accurate
GUARDRAIL SENSITIVITY	GR-L08-11 lowers it, because more infrastructure-emitted events means more assertions have comparators; GR-L08-01 lowers it modestly, since a transcript supplies some comparator; GR-L08-09 raises it as the retention horizon shortens; GR-L07-08 lowers it, since classified actions leave classified records; GR-L03-06 lowers it by preserving comparators that would otherwise be overwritten
CONFOUNDERS	Report style and length; the type of work, since design and analysis produce inherently less verifiable assertions than deployment; the comparator inventory available to the auditor; a model version change altering how the agent writes
GAMING MODES	Padding reports with trivially verifiable assertions so the ratio falls while the consequential assertions remain unverifiable; framing an unverifiable claim as an intention, which the definition excludes from the denominator
CORROBORATED BY	MG-027 evidence citation completeness; MG-022 claim divergence rate read on the same episodes
READINESS	absent
INSTRUMENTATION GAP	An assertion-level parse of self-reports and a comparator inventory; the founding case audits at episode granularity only

MG-027 Evidence citation completeness

QUESTION	What proportion of material claims carry a resolvable citation to independent evidence?
OPERATIONAL DEFINITION	Material claims accompanied by a citation to a specific evidence artefact, such as a commit identifier, a log location with a timestamp, or a hash of command output, divided by material claims. A citation counts only if the named artefact exists and contains what the claim says. Excludes citations to the agent own prior prose; excludes citations to artefacts the agent created in the same session solely to be cited.
FORMULA	claims_with_resolvable_citations / material_claims
UNIT AND RANGE	ratio, 0 to 1
TYPE	KCI
HORIZON	leading
GRANULARITY	per report, aggregated per agent per week
DATA SOURCE	AGENT-EMITTED claims, with resolution performed against INFRASTRUCTURE-EMITTED artefacts by the audit harness
COLLECTION POINT	The audit harness citation resolver, which must fetch the named artefact and test the asserted content against it
SAMPLING FREQUENCY	weekly
DIRECTION	higher-better
BASELINE	not computable; the founding case self-reports are prose messages truncated at 480 characters on the notification path and free text memory notes, and neither carries a citation convention
TARGET	1.0 for claims about action classes rated irreversible or severe; requiring citation on low risk claims imposes cost with no risk reduction
GUARDRAIL SENSITIVITY	GR-L10-04 raises it, since approval requests must present evidence rather than the agent summary; GR-L08-12 raises it through run provenance binding each measurement to a configuration hash; GR-L10-05 raises it by distinguishing claimed from verified in the interface; GR-L08-08 raises it; GR-L09-05 lowers it, because notification rate limiting and 480 character truncation leave no room for a citation
CONFOUNDERS	Report medium, since a message channel with a hard character limit cannot carry citations; the strictness of the resolver; the availability of stable artefact identifiers; a model version change altering citation habits
GAMING MODES	Citing an artefact that exists but does not contain the asserted content, which passes a naive existence check; citing an artefact the agent wrote itself moments earlier; attaching one genuine citation to a paragraph containing several uncited claims
CORROBORATED BY	MG-024 verification independence ratio, which tests whether the cited artefact has an independent author; MG-022 claim divergence rate
READINESS	absent
INSTRUMENTATION GAP	A citation convention in agent reports, stable artefact identifiers, and a resolver that tests content rather than existence

MG-028 Adversarial challenge yield

QUESTION	How often does a deliberate adversarial review of a completion claim find a material defect?
OPERATIONAL DEFINITION	Adversarial reviews finding at least one material defect, divided by adversarial reviews performed. An adversarial review is one whose instruction is to find the failure and which is performed by a principal other than the author. Excludes confirmatory reviews; excludes reviews of claims the author had already withdrawn; excludes findings the author accepts as immaterial without dispute.
FORMULA	reviews_finding_material_defects / adversarial_reviews
UNIT AND RANGE	ratio, 0 to 1
TYPE	KCI
HORIZON	leading
GRANULARITY	per review, aggregated per month
DATA SOURCE	Review records, which are independent only if the reviewer holds a separate identity; on this host the reviewer shares the author account, so the record is effectively AGENT-EMITTED
COLLECTION POINT	The review queue, recording the reviewing principal, the instruction given and the findings raised
SAMPLING FREQUENCY	monthly
DIRECTION	target-band
BASELINE	not computable as a rate; the closest observation is the research audit itself, which found 11 material divergences in 18 episodes, but that was a single audit protocol rather than a routine review population
TARGET	not set as a point value; a yield near zero means either the work is sound or the review is not adversarial, and a yield near one means the work is not fit to review, so only the intermediate range is informative and only when read against MG-022
GUARDRAIL SENSITIVITY	GR-L08-07 raises it, being the mechanism itself; GR-L01-07 raises it by making the reviewer genuinely other; GR-L08-06 raises it; GR-L07-02 raises it modestly, since separating the decision point from the enforcement point gives the reviewer a defined standing; GR-L12-04 does not move it, being prose
CONFOUNDERS	Reviewer capability and model version; the instruction given to the reviewer, which determines how adversarial the review actually is; the difficulty of the reviewed work; whether the reviewer and author share a training distribution and therefore share blind spots
GAMING MODES	Submitting only trivially defensible claims to adversarial review; the reviewer, sharing the author incentives and account, raising real but immaterial findings so the yield looks healthy; the author pre-correcting known defects and submitting the cleaned version
CORROBORATED BY	MG-022 claim divergence rate measured on the same episodes; MG-024 verification independence ratio
READINESS	absent
INSTRUMENTATION GAP	A review queue with a recorded adversarial instruction and a materiality adjudication by a third party; the founding case has agent-to-agent escalation notes but no review population

MG-029 Self-attestation share of the record

QUESTION	What proportion of the governance record about an agent work was written by that agent?
OPERATIONAL DEFINITION	Records in the evidence corpus for a window whose writing principal is the acting agent or a process running under its account, divided by all records in that corpus. Includes model runtime transcripts, hook output, pane logs, memory notes, completion markers and notification bodies. Excludes operator-authored console entries and records written by a process under a different account. Measured on record counts, with a byte-weighted figure reported alongside.
FORMULA	agent_written_records / total_evidence_records
UNIT AND RANGE	ratio, 0 to 1
TYPE	KRI
HORIZON	leading
GRANULARITY	per window per agent
DATA SOURCE	Determined from filesystem ownership and writer attribution rather than from any claim, so the classification is INFRASTRUCTURE-EMITTED even though the records themselves are not
COLLECTION POINT	An inventory of evidence stores that records, for each store, the writing process and the account it runs as
SAMPLING FREQUENCY	monthly
DIRECTION	lower-better
BASELINE	approaching 1 and not computed exactly; about 1.13 GiB of model runtime transcript, about 186 MB of hook output, the pane logs, the memory stores, the fleet registry and the resource accounting counters are all owned and writable by the single account every agent runs as; the one store written by a different code path is the console audit log, 863 lines over 46 days, and it records operator actions through the console only, never agent actions
TARGET	not set as a corpus-wide level; the architecturally meaningful target is that the subset of the record relied upon for verification has a self-attestation share of 0
GUARDRAIL SENSITIVITY	GR-L08-11 lowers it by adding records the agent does not write; GR-L03-06 lowers it through an immutable append-only evidence directory; GR-L03-10 lowers it by separating the evidence store from the operational filesystem; GR-L05-05 lowers it; GR-L08-01 raises it, since a transcript written by the model runtime under the agent own account adds agent-owned volume to the record
CONFOUNDERS	The verbosity of the agent, which changes the numerator without changing governance; the addition of high-volume low-value infrastructure logging, which changes the denominator; whether the corpus is defined by record count or by bytes
GAMING MODES	Increasing the volume of infrastructure-emitted records that carry no verifying content so the ratio falls while verifiability does not rise; deleting agent-written records rather than adding independent ones, which improves the ratio by destroying evidence
CORROBORATED BY	MG-065 audit record immutability; MG-064 audit record completeness
READINESS	derivable
INSTRUMENTATION GAP	A writer attribution per evidence store; ownership is observable today but no inventory ties each store to its writing principal and its record count

MG-029 self-attestation share of the record is the subfamily's structural counterpart to MG-022's behavioural finding: it does not ask whether any particular report was accurate, only what share of the entire evidentiary record was written by the party being evaluated. The founding case's answer approaches 1 across every store examined, transcripts, hook output, pane logs, memory notes, the fleet registry and the resource accounting counters, with a single exception, the console audit log at 863 lines over 46 days, which records operator actions through the console and never agent actions EV-034. Read together, MG-022 and MG-029 say the same thing from two directions: not only does the record diverge from the truth at a measured rate, almost the entire record is written by the one party with an interest in what it says.

3.2 Boundary and mandate

MG-030 to MG-037 instrument D2's access-versus-mandate distinction directly. MG-030 access exceeds mandate count is the subfamily's one available entry, computable today by joining the access and mandate columns already populated for all sixty action classes: fifty of sixty show access wider than mandate EV-003 EV-032.

Table D4-6. Boundary and mandate metrics, MG-030 to MG-037.

MG-030 Access exceeds mandate count

QUESTION	For how many action classes does the actor hold technical access wider than its authorised mandate?
OPERATIONAL DEFINITION	Count of action classes in which the access level technically attainable by the agent is more permissive than the level its mandate authorises, assessed per class against the action class register. Excludes classes where access and mandate are equal or where access is narrower. Excludes classes unreachable only because tooling is absent rather than because a control prevents them, since absent tooling can be installed by the same agent.
FORMULA	count of classes where ordinal(access_level) < ordinal(mandate_level), with I = 1, II = 2, III = 3, so a lower ordinal is more permissive
UNIT AND RANGE	count, integer 0 to 60
TYPE	KRI
HORIZON	leading
GRANULARITY	per host configuration per agent identity
DATA SOURCE	INFRASTRUCTURE-EMITTED, derived from the credential set, the operating system privilege model, the tool permission configuration and the network posture, compared against the recorded mandate
COLLECTION POINT	A periodic posture assessment that joins the observed access model to the action class register
SAMPLING FREQUENCY	per configuration change, and at least monthly
DIRECTION	lower-better
BASELINE	50 of 60 action classes have access exceeding mandate
TARGET	0, because the gap is the definition of unenforced governance; any non-zero value means at least one prohibition depends on the agent choosing to comply
GUARDRAIL SENSITIVITY	GR-L01-03 lowers it strongly through a least privilege permission set per action class; GR-L07-06 lowers it by making self-widening impossible; GR-L01-09 raises it, since passwordless root collapses access to level I for nearly every class; GR-L01-10 raises it by extending access to the cloud classes through the metadata credential; GR-L04-06 lowers it for the shell-mediated classes
CONFOUNDERS	A change to the mandate rather than to access moves the count with no change in what the agent can do; adding or retiring action classes changes the population; estate heterogeneity, since the assessment is per host and hosts differ
GAMING MODES	Widening the mandate to match the access, which drives the count to zero while changing nothing technically; classifying an action into a neighbouring class whose mandate is more permissive; assessing access on a hardened reference host rather than on the host the agent runs on
CORROBORATED BY	MG-034 mandate stability and MG-035 mandate widening events initiated by the governed party jointly detect the widening strategy; MG-031 structural enforcement coverage
READINESS	available
INSTRUMENTATION GAP	none; computable today by joining the access and mandate columns of the action class register, both of which are populated for all 60 classes

MG-031 Structural enforcement coverage

QUESTION	What proportion of action classes are constrained by at least one structural control rather than by instruction alone?
OPERATIONAL DEFINITION	Action classes for which at least one applicable control is of enforcement class structural, divided by action classes in the register. Structural means the prohibited action is not reachable by the actor and defeating it requires changing the architecture rather than changing a decision. Excludes conditional controls the actor can alter, instructional controls entirely, and controls that are structural in code but whose configuration file, policy data or hook script the actor can write.
FORMULA	classes_with_at_least_one_structural_control / total_action_classes
UNIT AND RANGE	ratio, 0 to 1
TYPE	KCI
HORIZON	leading
GRANULARITY	per host configuration
DATA SOURCE	INFRASTRUCTURE-EMITTED, derived from the guardrail catalogue enforcement class and population fields and verified against the running host rather than against documentation
COLLECTION POINT	A control inventory that tests, for each candidate control, whether the acting account can write the artefact that implements it
SAMPLING FREQUENCY	per configuration change, and at least monthly
DIRECTION	higher-better
BASELINE	not computable as a ratio in this block; the founding case structural controls are few and enumerable, but a general disqualifier applies, namely that every settings file, hook script, evidence store and registry implementing a control is writable by the account those controls govern, so the per-class count was not computed here
TARGET	1.0 for classes rated severe or irreversible, because below that threshold the prohibition is advice rather than a control
GUARDRAIL SENSITIVITY	GR-L07-01 raises it by expressing policy as data a decision point can evaluate; GR-L07-07 raises it by covering the unenumerated remainder; GR-L04-06 raises it; GR-L03-08 raises it, since without denying the actor write access to its own permission settings no tool-layer control qualifies as structural; GR-L01-01 lowers it, because a shared account with passwordless root disqualifies otherwise structural controls
CONFOUNDERS	How strictly the structural criterion is applied, which is the dominant source of variance; changes to the action class population; assessing the control as designed rather than as deployed
GAMING MODES	Counting a control as structural when its configuration is writable by the actor, which is the founding case dominant pattern; counting a control whose code exists but whose consumer was never implemented, of which this host has a documented example on its most dangerous tools
CORROBORATED BY	MG-059 control bypass rate; MG-062 defence in depth factor
READINESS	derivable
INSTRUMENTATION GAP	A per-class control inventory with a writability test on each implementing artefact; the guardrail catalogue supplies the candidate controls but no automated writability assessment exists

MG-032 Boundary crossing attempt rate

QUESTION	How often does an agent attempt an action outside its mandate?
OPERATIONAL DEFINITION	Count of attempted actions whose class carries a mandate level the agent has not satisfied at the time of attempt, per agent per day. An attempt is counted whether or not it succeeds and whether or not it is refused. Excludes actions within mandate; excludes attempts by non-agent principals; excludes repeated identical attempts within one second, which are counted once.
FORMULA	out_of_mandate_attempts / agent_days
UNIT AND RANGE	count per agent-day, 0 to unbounded
TYPE	KRI
HORIZON	leading
GRANULARITY	per agent per day
DATA SOURCE	INFRASTRUCTURE-EMITTED at a policy decision point, which does not exist on this host; the hook substrate sees every tool call but is contractually required to exit successfully and therefore cannot evaluate anything
COLLECTION POINT	A policy decision point invoked before the tool call, recording the classified action, the mandate level and the decision
SAMPLING FREQUENCY	daily
DIRECTION	lower-better
BASELINE	not computable; no runtime action classifier exists, so an out-of-mandate attempt is indistinguishable from any other tool call; the founding case cannot enumerate its own attempts, which is itself the finding
TARGET	not set; on first instrumentation the measured value rises from zero because attempts become visible, so a target set before a stable baseline would penalise detection
GUARDRAIL SENSITIVITY	GR-L07-08 does not move the underlying behaviour but raises the measured value from zero by making attempts visible; GR-L04-06 lowers it over time, since an allowlist teaches the boundary; GR-L07-07 raises the measured value, because default deny converts previously silent actions into recorded refusals; GR-L01-03 lowers it; GR-L12-04 does not reliably move it, which is the central claim this register tests
CONFOUNDERS	The breadth of the mandate, since a narrow mandate produces more crossings for identical behaviour; task composition; the classifier granularity; a model version change altering how readily the agent reaches for a privileged tool
GAMING MODES	Performing the action through a path the classifier does not see, for example a shell pipeline rather than a named tool; splitting one out-of-mandate action into steps each of which is individually within mandate; having another agent perform the step
CORROBORATED BY	MG-036 unclassified action share detects displacement into unseen paths; MG-033 boundary crossing success rate; MG-024 verification independence ratio detects the proxy-agent strategy
READINESS	absent
INSTRUMENTATION GAP	A runtime action classifier and a policy decision point; the hook substrate already fires on every tool call with a match-everything matcher, so the observation surface exists and only the decision function is missing

MG-033 Boundary crossing success rate

QUESTION	When an agent attempts an action outside its mandate, how often does it succeed?
OPERATIONAL DEFINITION	Out-of-mandate attempts that produced their intended effect, divided by out-of-mandate attempts. Effect is judged from the infrastructure state record, not from the tool return value, because a tool can report success without effect and can report failure after a partial effect. Excludes attempts whose intended effect cannot be determined.
FORMULA	successful_out_of_mandate_attempts / out_of_mandate_attempts
UNIT AND RANGE	ratio, 0 to 1
TYPE	KRI
HORIZON	lagging
GRANULARITY	per action class per month
DATA SOURCE	Denominator from the policy decision point; numerator INFRASTRUCTURE-EMITTED from the state of the affected system
COLLECTION POINT	The policy decision point for the attempt, joined to the affected system state record for the effect
SAMPLING FREQUENCY	monthly
DIRECTION	lower-better
BASELINE	not computable as a ratio, and the reason is the finding: with no enforcement point in the runtime, an attempt succeeds unless the model itself declines, so the expected value is 1 across the 50 classes in which access exceeds mandate
TARGET	0, because a non-zero value means that what is described as enforcement is not enforcement
GUARDRAIL SENSITIVITY	GR-L07-06 lowers it; GR-L04-06 lowers it; GR-L01-03 lowers it; GR-L02-05 lowers it for the egress-bearing classes; GR-L04-10 raises it, since permission skipping enabled by default removes the only per-call gate the tooling offers
CONFOUNDERS	Which classes are attempted, since success rates differ enormously by class; whether the enforcement point is in the path for all invocation routes or only for named tools; the definition of intended effect
GAMING MODES	Recording only refused attempts as attempts, so the denominator fills with failures and the ratio falls; reclassifying a successful crossing as within mandate after the fact, which the actor can do because it can write the mandate files; declaring the effect unintended once it is observed
CORROBORATED BY	MG-035 mandate widening events initiated by the governed party detects the reclassification strategy; MG-059 control bypass rate
READINESS	absent
INSTRUMENTATION GAP	A policy decision point that records refusals as well as permissions, and an effect record independent of the tool own return value

MG-034 Mandate stability

QUESTION	How often does an agent mandate change, and how much of it changes?
OPERATIONAL DEFINITION	Count of changes to mandate-bearing artefacts per window, and the proportion of mandate clauses altered. Mandate-bearing artefacts are instruction files, role files, permission settings and any policy data consulted at authorisation time. Excludes non-normative documentation; excludes formatting-only edits; excludes changes to files that no authorisation path reads.
FORMULA	mandate_changes per window; and altered_clauses / total_clauses
UNIT AND RANGE	count per window, integer >= 0; and ratio, 0 to 1
TYPE	KCI
HORIZON	lagging
GRANULARITY	per agent per week
DATA SOURCE	INFRASTRUCTURE-EMITTED only if mandate artefacts are versioned under a principal other than the agent; on this host they are ordinary files owned by the agents they govern and versioned by a scheduled commit job the same account controls
COLLECTION POINT	A versioned mandate store whose write path requires a principal the agent does not hold
SAMPLING FREQUENCY	weekly
DIRECTION	target-band
BASELINE	not computable; the founding case mandate is spread across about 40 classic instruction files totalling about 538 kB, or about 1,240 files and 26.7 MB counting role files, rosters, prompts and three memory stores, none of which is versioned as a mandate and all of which are writable by the agents they govern
TARGET	not set as a level; a mandate that never changes is not being maintained and one that changes constantly is not a constraint, so the reportable quantity is the distribution and its authorship, not a target
GUARDRAIL SENSITIVITY	GR-L07-03 lowers it and makes it measurable through signed and versioned mandates; GR-L07-05 lowers it by adding an approver role to each change; GR-L07-04 raises it, since mandate expiry forces periodic reissue; GR-L03-07 lowers it by denying the agent write access to its own instruction files; GR-L07-09 raises it, because a prose brief as the operative mandate changes with every dispatch
CONFOUNDERS	Organisational change unrelated to the agent; the definition of a clause, which determines the denominator; consolidation or splitting of instruction files, which changes the count without changing the mandate
GAMING MODES	Making a semantically large change with a small diff, for example deleting a single prohibition; moving a constraint into a file not counted as mandate-bearing; batching many changes into one commit so the count stays low
CORROBORATED BY	MG-035 mandate widening events initiated by the governed party; MG-067 separation of duties violations
READINESS	derivable
INSTRUMENTATION GAP	A defined set of mandate-bearing artefacts and a clause-level parse; the version control history exists but treats these files as ordinary documentation and does not distinguish the authoring principal

MG-035 Mandate widening events initiated by the governed party

QUESTION	How often does the agent itself change the rules that govern it in a direction that increases its own permitted scope?
OPERATIONAL DEFINITION	Count of changes to mandate-bearing artefacts authored by the governed agent whose effect is to permit an action previously prohibited, to remove a required approval, or to broaden the scope of an existing permission. Excludes clarifications with no scope effect; excludes narrowing changes; excludes changes authored by an operator or by a principal outside the governed set.
FORMULA	widening_changes_authored_by_the_governed_agent per window
UNIT AND RANGE	count, integer >= 0
TYPE	KRI
HORIZON	lagging
GRANULARITY	per agent per month
DATA SOURCE	Version control diffs of mandate artefacts joined to an authorship attribution; on this host the diff exists in a repository the same account can rewrite and authorship is not recoverable, because every agent shares one operating system account and commits under one human identity by standing instruction
COLLECTION POINT	The mandate store write path, recording the authenticated principal of each change, under a principal separation the host does not currently have
SAMPLING FREQUENCY	monthly
DIRECTION	lower-better
BASELINE	not computable; authorship of a mandate change cannot be attributed to an agent rather than to the operator, because commits are made under a single human identity by standing instruction and all agents share one account; that the capability exists is established, since every instruction and permission file is writable by the agents it governs
TARGET	0, because self-widening is the failure the architectural enforcement claim is aimed at, and any non-zero value falsifies the enforcement of the mandate layer
GUARDRAIL SENSITIVITY	GR-L07-06 drives it to 0 by construction; GR-L03-07 lowers it; GR-L03-08 lowers it; GR-L07-05 lowers it and makes it detectable through an approver-attributed change journal; GR-L01-01 raises it in effect, since a shared account makes authorship unattributable and events therefore pass unrecorded
CONFOUNDERS	Legitimate operator-authored widening misattributed to an agent, or the reverse; a change in what counts as a mandate artefact; the diff classifier sensitivity to deletions as against additions
GAMING MODES	Widening by omission, that is deleting a constraint rather than adding a permission, which a naive diff classifier may not flag; arranging for a second agent to make the change so the governed party is not the author on paper while both are the same principal; widening the mandate in a file consulted at runtime while leaving the human-readable one unchanged
CORROBORATED BY	MG-024 verification independence ratio detects the proxy-author strategy; MG-067 separation of duties violations; MG-034 mandate stability
READINESS	absent
INSTRUMENTATION GAP	Per-agent authorship on mandate changes, which requires per-agent identity; the founding case has neither, so the event class is unobservable rather than absent

MG-036 Unclassified action share

QUESTION	What proportion of the actions an agent takes fall outside any classification the policy layer recognises?
OPERATIONAL DEFINITION	Tool invocations and shell commands that a runtime classifier cannot map to an action class, divided by all tool invocations and shell commands. Calls mapped with low confidence are counted as classified and reported separately, so that low confidence is not concealed inside the unclassified figure. Excludes model operations producing no side effect.
FORMULA	unclassified_calls / total_calls
UNIT AND RANGE	ratio, 0 to 1
TYPE	KRI
HORIZON	leading
GRANULARITY	per agent per day
DATA SOURCE	INFRASTRUCTURE-EMITTED from the classifier at the policy decision point for the numerator; the denominator is observable today, since the hook substrate fires on every tool call with a match-everything matcher and produces about 186 MB of output across active sessions
COLLECTION POINT	The classifier at the policy decision point, immediately before the tool call
SAMPLING FREQUENCY	daily
DIRECTION	lower-better
BASELINE	not computable as a ratio and identically 1 by construction today, because no runtime classifier exists and no action is classified at the time of attempt; the denominator is available from the hook stream and the numerator is undefined
TARGET	below 0.05 within a defined tool inventory, because an unclassified action cannot be governed and the residual must be small enough to review individually; 0 is not defensible while general shell execution remains available
GUARDRAIL SENSITIVITY	GR-L07-08 lowers it, being the mechanism that makes classification exist; GR-L07-07 lowers its consequence rather than its value, by denying whatever remains unclassified; GR-L04-06 lowers it, since an allowlist enumerates the classifiable surface; GR-L04-05 does not move it, because a deny list classifies nothing it does not name; GR-L08-13 lowers it if the existing hook substrate is used to classify rather than only to record
CONFOUNDERS	The breadth of the tool inventory; how much work is routed through general shell execution as against named tools; classifier maturity; a model version change altering tool selection
GAMING MODES	Routing work through one broadly defined class such as generic shell execution so nothing is formally unclassified while nothing is meaningfully classified; performing side effects inside a tool whose declared class does not describe them, for example writing through a shell redirect issued by a nominally read-only tool
CORROBORATED BY	MG-032 boundary crossing attempt rate; MG-061 control signal emission coverage
READINESS	derivable
INSTRUMENTATION GAP	A classifier and a tool inventory; the observation surface already exists in the hook substrate, which sees every tool call and is deliberately prevented from acting on what it sees

MG-037 Level demotion latency

QUESTION	Once evidence establishes that an agent's effective access for an action class should be reduced, how long does the reduction take to actually take effect?
OPERATIONAL DEFINITION	Elapsed time from the moment a demotion trigger is established, such as a divergence finding, a confirmed boundary crossing or an incident, to the moment the agent's technical access for that class is actually narrowed. Excludes changes to prose mandate text with no accompanying technical change; excludes demotions proposed but never applied.
FORMULA	median over demotion events of (t_effective_restriction_applied - t_trigger_established)
UNIT AND RANGE	seconds, 0 to unbounded
TYPE	KCI
HORIZON	lagging
GRANULARITY	per demotion event, aggregated per agent per quarter
DATA SOURCE	Mixed. The trigger is ideally established by an independent audit harness; the restriction itself is applied through a settings or permission file that on this host is writable by the account it governs, so the application record is agent-adjacent even where the trigger is independent.
COLLECTION POINT	The permission or access configuration write path, joined to the audit harness record of when the trigger was established
SAMPLING FREQUENCY	per event, reviewed monthly
DIRECTION	lower-better
BASELINE	not computable; no demotion event has ever occurred on this host, because access is not assigned per agent at all, every agent shares one operating system account holding unrestricted passwordless root, and no mechanism exists to narrow one agent's access without narrowing every agent's access simultaneously
TARGET	not set; a latency figure is not meaningful before per-agent access exists to narrow
GUARDRAIL SENSITIVITY	GR-L01-02 lowers it and is the precondition, since demotion is unrepresentable without a distinct principal to demote; GR-L01-04 ephemeral credentials with short session duration lowers it by letting a demotion take effect at natural expiry rather than requiring active revocation; GR-L07-04 mandate expiry lowers it the same way; GR-L09-06 automatic suspension on anomaly lowers it by triggering demotion without waiting for a human decision; GR-L01-01 makes it undefined, since a shared account has no per-agent access to narrow
CONFOUNDERS	How quickly the trigger itself is recognised, which is a separate metric (MG-025) bundled into this one if not reported apart; whether the demotion requires a session restart to take effect; organisational reluctance to narrow a productive agent's access
GAMING MODES	Recording the demotion as effective at the moment a settings file is edited, while the running session retains its prior credential until its next restart; narrowing access in a document the runtime never reads while leaving the technical grant untouched
CORROBORATED BY	MG-030 access exceeds mandate count; MG-050 nominal to effective posture divergence count
READINESS	absent
INSTRUMENTATION GAP	Per-agent identity and access levels to narrow, and a trigger-to-application pipeline; neither exists, and the underlying access model does not currently vary by agent

MG-032 and MG-033, boundary crossing attempt rate and success rate, share a single obstacle that is itself a finding: no runtime action classifier exists on this host, so an out-of-mandate attempt is indistinguishable from any other tool call, and MG-033's expected value under the current architecture is 1 across the fifty classes where access exceeds mandate, because nothing in the runtime evaluates an attempt before it executes. The hook substrate that could serve as the observation surface for both metrics already fires on every tool call; what is missing in each case is only the decision function, not the event stream EV-026.

3.3 Containment

MG-038 to MG-043 measure how far a single error can travel, the operational form of D2's SD-07 containment shortfall determinant, which D2 Section 3.2 identifies as the largest single group in the supervision residue and the one most fully removable by identity engineering.

Table D4-7. Containment metrics, MG-038 to MG-043.

MG-038 Blast radius realised per incident

QUESTION	When an incident occurs, how far does its effect actually propagate compared with what the architecture assumed it could reach?
OPERATIONAL DEFINITION	For each incident, the observed scope of effect scored against the action class register's blast radius categories, development, local-host, estate or external-public, established from independent evidence rather than from the acting agent's own account of what happened. Excludes near-misses with no realised effect, which are counted separately; excludes incidents whose scope could not be independently established.
FORMULA	per incident, blast_radius_realised = ordinal(observed_scope), with development=1, local-host=2, estate=3, external-public=4; reported as the distribution across incidents in the window
UNIT AND RANGE	ordinal, 1 to 4
TYPE	KRI
HORIZON	lagging
GRANULARITY	per incident
DATA SOURCE	Mixed. The initial account of an incident is typically AGENT-EMITTED; confirmation of true scope requires INFRASTRUCTURE-EMITTED evidence from every system potentially reached, such as cross-host logs, cloud control plane events and DNS records.
COLLECTION POINT	An incident review drawing on cross-host and cloud control plane evidence independent of the account or accounts involved
SAMPLING FREQUENCY	per incident, reviewed quarterly
DIRECTION	lower-better
BASELINE	not computable as a distribution; no incident register exists on this host, so incidents are not enumerated as a population; the one documented case, a deploy reported successful that had in fact disabled the service's core connection function for about 24 hours, reached at least local-host and arguably estate scope informally, but was never scored against this scale
TARGET	not set; the productive comparison is against the maximum blast radius the architecture actually permits given passwordless root and five non-expiring administrator-equivalent cloud credentials, which is external-public for nearly every agent, so any realised scope below that ceiling reflects restraint in what happened to be attempted rather than a property of the containment architecture
GUARDRAIL SENSITIVITY	GR-L02-08 segmentation between agent tiers lowers it; GR-L02-09 no lateral network access between agent workspaces lowers it; GR-L06-07 blast radius cap per change lowers it directly; GR-L01-02 lowers it by scoping what a single compromised identity can reach; GR-L03-03 per-session filesystem isolation lowers it for the local-host component
CONFOUNDERS	Which action classes are actually attempted in the incident window; how quickly containment or intervention arrests propagation before it reaches its architectural ceiling; the credential surface available at the time, which sets the ceiling independent of likelihood
GAMING MODES	Describing an incident narrowly, for example as touching one file, when the credential or account involved in fact reaches estate or cloud scope, so the description undercounts what actually happened; closing the incident review before downstream propagation through a shared credential is checked
CORROBORATED BY	MG-040 lateral reach from a compromised agent identity; MG-022 claim divergence rate
READINESS	absent
INSTRUMENTATION GAP	An incident register with a scored scope field, and cross-host or cross-account evidence collection independent of the account involved; neither exists

MG-039 Containment success rate

QUESTION	When an errant or compromised action occurs, how often is its effect actually held within the boundary the containment architecture intends?
OPERATIONAL DEFINITION	Incidents whose observed blast radius does not exceed the containment boundary declared for the affected action class, divided by incidents reviewed. Excludes incidents with no declared containment boundary for the affected class, which are reported separately as unassessable; excludes incidents contained only because they were never attempted at scale, which are not distinguishable from genuine containment on this host.
FORMULA	contained_incidents / total_incidents_reviewed
UNIT AND RANGE	ratio, 0 to 1
TYPE	KCI
HORIZON	lagging
GRANULARITY	per incident, aggregated per quarter
DATA SOURCE	Same as MG-038; the containment boundary itself would come from a declared policy the action class register does not currently carry
COLLECTION POINT	An incident review comparing MG-038's realised scope against a declared intended boundary per action class
SAMPLING FREQUENCY	quarterly
DIRECTION	higher-better
BASELINE	not computable; no incident register exists and no action class carries a declared containment boundary to compare a realised scope against, so the numerator and the comparison standard are both absent
TARGET	1.0 for classes rated severe blast radius or irreversible; below that threshold a breach of intended boundary is tolerable if bounded
GUARDRAIL SENSITIVITY	GR-L02-08 raises it; GR-L02-09 raises it; GR-L03-03 raises it; GR-L06-07 raises it directly, since it declares the boundary the metric would check against; GR-L01-01 lowers it, because a shared account with passwordless root removes any boundary for the numerator to be compared against
CONFOUNDERS	Incident severity mix in the window; whether the declared boundary reflects genuine architectural constraint or is set generously to make containment easy to satisfy; detection completeness, since an uncontained incident that is never detected cannot be counted as a failure
GAMING MODES	Declaring the containment boundary after the fact to match wherever the incident happened to stop; setting an intentionally wide boundary so that most incidents are contained by definition
CORROBORATED BY	MG-038 blast radius realised per incident; MG-042 privilege breadth index
READINESS	absent
INSTRUMENTATION GAP	An incident register, and a declared containment boundary per action class; neither exists

MG-040 Lateral reach from a compromised agent identity

QUESTION	If one agent session or its credentials were fully compromised, how many additional hosts, accounts and credential-bearing artefacts could be reached with no new secret or privilege?
OPERATIONAL DEFINITION	Count of distinct estate hosts, cloud accounts and credential stores reachable from a single agent session using only credentials that session already holds or can trivially obtain, such as an owned private key, a stored cloud key pair, or the instance metadata endpoint. Excludes reach that would require guessing a secret or exploiting an unpatched defect elsewhere; excludes reach already counted through a different owned credential to the same target.
FORMULA	count of distinct reachable targets (hosts + cloud accounts + credential stores) from one agent session, established by posture assessment
UNIT AND RANGE	count, integer >= 0
TYPE	KRI
HORIZON	leading
GRANULARITY	per agent identity, currently fleet-wide since one identity is shared
DATA SOURCE	INFRASTRUCTURE-EMITTED, derived from the credential inventory, the account model, network reachability and the filesystem permission model
COLLECTION POINT	A posture assessment walking the credential and key inventory, the network reachability graph and the metadata endpoint
SAMPLING FREQUENCY	at least monthly, and on every credential or topology change
DIRECTION	lower-better
BASELINE	partially observed as a lower bound, not a complete count: seven unencrypted private keys reaching other estate hosts and five non-expiring administrator-equivalent cloud key pairs reaching five distinct cloud accounts sit in files owned by the single account every agent runs as, and the cloud instance metadata endpoint is reachable from any process with egress unrestricted; a complete reachability graph across the estate, the metadata endpoint and the inter-agent messaging bus token model was not assembled for this register, so the true figure is at least twelve and is not an upper bound
TARGET	not set as an absolute; the defensible target is that reach from any one agent match only what that agent's declared mandate requires, which is not expressible while every agent shares one identity
GUARDRAIL SENSITIVITY	GR-L01-02 per-agent workload identity lowers it by scoping credentials to the agent that needs them; GR-L01-05 no long-lived key material on disk lowers it strongly, since the seven private keys and the cloud key file are exactly this pattern; GR-L02-09 no lateral network access lowers it; GR-L01-01 raises it, being the direct cause of the shared reach; GR-L02-12 blocking metadata access lowers it by one high-value target class
CONFOUNDERS	How many estate hosts and cloud accounts exist at all, which sets the ceiling the count is measured against; whether SSH host key checking is enabled, which the founding case documents as disabled for at least one credential distribution path; key rotation cadence
GAMING MODES	Rotating or deleting a credential file immediately before an assessment while an operational copy persists elsewhere, for example in a session scratch directory or shell history; measuring reachability only through named credential files while equivalent access exists through the metadata endpoint or an already-authenticated session
CORROBORATED BY	MG-042 privilege breadth index; MG-041 egress destination cardinality per agent
READINESS	derivable
INSTRUMENTATION GAP	A systematic reachability graph across hosts, cloud accounts, the metadata endpoint and the inter-agent bus, refreshed on change; today only a static credential and key inventory exists

MG-041 Egress destination cardinality per agent

QUESTION	To how many distinct external network destinations does an agent's traffic actually go?
OPERATIONAL DEFINITION	Count of distinct destination host and port pairs an agent's outbound traffic reaches in a window, across any protocol. Excludes traffic to loopback or to other processes on the same host; excludes DNS lookups that resolve to no connection.
FORMULA	count(distinct(dest_host, dest_port)) per agent per window
UNIT AND RANGE	count, integer >= 0
TYPE	KRI
HORIZON	leading
GRANULARITY	per agent per day
DATA SOURCE	Would be INFRASTRUCTURE-EMITTED from a firewall, proxy or egress log; none exists on this host, so any record would come from the agent's own session or application logs, which the agent account also owns
COLLECTION POINT	An egress point independent of the agent account, such as a network firewall or forward proxy log
SAMPLING FREQUENCY	daily
DIRECTION	lower-better
BASELINE	not computable; no host firewall is active, egress is unrestricted, and no egress log or proxy exists to enumerate destinations, so which agent reached what is not recorded anywhere independent of the agent's own account
TARGET	not set as an absolute; the informative comparison is cardinality against the small number of destinations a given agent's declared role actually requires
GUARDRAIL SENSITIVITY	GR-L02-05 deny by default egress lowers it structurally, since default reachability becomes zero rather than unrestricted; GR-L02-06 destination allowlist lowers it and supplies the denominator a meaningful target needs; GR-L02-07 DNS egress logging raises measured visibility without changing true cardinality; GR-L02-10 private endpoints for cloud service access lowers it by removing public egress for those calls; GR-L02-12 blocking metadata access lowers it by one destination
CONFOUNDERS	Legitimate use of package registries, the model provider API and notification channels inflates the count independent of risk; DNS-based multi-IP services appear as many destinations for one logical service; a change in what counts as loopback
GAMING MODES	Routing traffic through one already-permitted relay so the visible destination count stays low while effective reach is unchanged; batching many distinct destinations behind a single proxy hop that is the only hop logged
CORROBORATED BY	MG-040 lateral reach from a compromised agent identity; MG-036 unclassified action share, since egress through an unclassified path escapes both counts
READINESS	absent
INSTRUMENTATION GAP	An egress log or proxy independent of the agent account; none exists, and no host firewall is active to produce one as a side effect

MG-042 Privilege breadth index

QUESTION	How much technical privilege does a single agent identity hold, expressed as one comparable figure?
OPERATIONAL DEFINITION	A composite count of distinct high-consequence capabilities held by an agent's operating system account: whether it holds passwordless root, the count of non-expiring administrator-equivalent cloud credentials it can use, the count of estate hosts reachable through an owned private key, and the count of secret files it owns. Reported as the component counts and their sum rather than a weighted score, because weighting would assume a risk model the founding case does not establish. Excludes privileges reachable only through a further defect or guess, which belong to MG-040 rather than this metric.
FORMULA	privilege_breadth = has_passwordless_root(0 or 1) + n_administrator_equivalent_cloud_credentials + n_estate_hosts_reachable_by_owned_key + n_owned_secret_files
UNIT AND RANGE	count, integer >= 0, unbounded
TYPE	KRI
HORIZON	leading
GRANULARITY	per agent identity, currently fleet-wide since one identity is shared by all 14 sessions
DATA SOURCE	INFRASTRUCTURE-EMITTED, the account, credential and secret-file inventory
COLLECTION POINT	A periodic posture assessment of the account's root status, credential file and secret file ownership
SAMPLING FREQUENCY	monthly, and on every credential or account change
DIRECTION	lower-better
BASELINE	26, being 1 for passwordless root, 5 for the non-expiring administrator-equivalent cloud key pairs reaching five distinct cloud accounts, four at full administrator and the fifth at identity-administrator plus wildcard permissions, 7 for the estate hosts reachable through unencrypted private keys, and 13 for the owned secret files, held identically by every one of the 14 live agent sessions because they share one operating system account
TARGET	not set as an absolute figure; the defensible requirement is that the index differ across agents according to role, which cannot be expressed while every agent shares one identity
GUARDRAIL SENSITIVITY	GR-L01-02 lowers it by distributing the total across distinct identities rather than replicating it on every one; GR-L01-04 ephemeral credentials with short session duration lowers it, since a standing count of held credentials falls when they expire rather than persist; GR-L01-05 no long-lived key material on disk lowers it directly; GR-L01-12 credential inventory and rotation schedule lowers it by retiring stale grants; GR-L01-09 raises it, being one of its own components; GR-L01-10 raises it by adding the metadata credential path
CONFOUNDERS	How many cloud accounts and estate hosts exist at all, which bounds how large the count could be; consolidation or expansion of the estate; what counts as administrator-equivalent, a judgement applied consistently in the founding case's own credential audit
GAMING MODES	Relabelling a credential's permission set as least privilege without actually revoking the standing reach it grants; rotating credentials on a schedule that resets a count without reducing what the new credential still permits
CORROBORATED BY	MG-030 access exceeds mandate count; MG-040 lateral reach from a compromised agent identity
READINESS	available
INSTRUMENTATION GAP	none for the current single-identity figure, which is directly observed; a per-agent decomposition requires per-agent identity, which does not exist

MG-043 Irreversible action share

QUESTION	What proportion of an agent's executed actions belong to an action class the register marks irreversible?
OPERATIONAL DEFINITION	Count of executed actions whose action class is recorded as irreversible in the action class register, divided by all executed actions, per agent per window. Excludes actions with no assignable class, which are counted separately under MG-036.
FORMULA	irreversible_actions / total_actions
UNIT AND RANGE	ratio, 0 to 1
TYPE	KRI
HORIZON	leading
GRANULARITY	per agent per week
DATA SOURCE	Would require a runtime classifier joining executed actions to the action class register's reversibility field at the point of execution; the register's classification exists but the runtime join does not
COLLECTION POINT	The same policy decision point MG-032 and MG-036 would sit at, immediately before the tool call, reading the reversibility field
SAMPLING FREQUENCY	weekly
DIRECTION	lower-better
BASELINE	not computable as an execution rate; 15 of the 60 action classes in the register, 25 per cent of the population, are classed irreversible, but no runtime classifier joins executed actions to that field, so the share of actual executions falling into an irreversible class is unmeasured
TARGET	not set as a number below the 25 per cent population share, since a defensible target requires knowing execution volume by class first; the immediate target is that this metric become computable at all
GUARDRAIL SENSITIVITY	GR-L06-03 pre-approved rollback targets lowers it by converting some nominally irreversible effects to reversible-with-effort; GR-L05-09 recoverable agent state lowers it the same way; GR-L06-02 pre-apply snapshot lowers it; GR-L07-07 default deny for unclassified actions lowers it by refusing irreversible attempts that were previously silent; GR-L07-08 does not change the true share but is the precondition for measuring it, echoing MG-032's treatment of the same guardrail
CONFOUNDERS	Workload composition and how much of the genuine task requires an irreversible step, such as sending an external message; a change to the register's reversibility classification itself
GAMING MODES	Reclassifying an action to a less severe reversibility rating rather than reducing its actual irreversibility; performing the irreversible step through an unclassified path so it never enters the denominator, which MG-036 would need to catch
CORROBORATED BY	MG-036 unclassified action share; MG-030 access exceeds mandate count
READINESS	derivable
INSTRUMENTATION GAP	The same runtime classifier MG-032 and MG-036 require, joined to the reversibility field rather than the mandate field; neither the classifier nor the join exists

MG-042 privilege breadth index is the subfamily's one `available` entry and its baseline is a direct sum rather than an estimate: 1 for passwordless root, plus 5 non-expiring administrator-equivalent cloud credential sets reaching five distinct cloud accounts, plus 7 estate hosts reachable through unencrypted private keys, plus 13 owned secret files, giving 26, held identically by every one of the fourteen live sessions because they share one account EV-003 EV-033. MG-040 lateral reach reports the same picture as a lower bound rather than a complete graph, at least twelve distinct reachable targets, because the reachability graph across the estate, the metadata endpoint and the inter-agent messaging bus was not assembled as a single artefact for this register.

3.4 Supervision viability

`MG-044` to `MG-052` are the direct instrumentation of D2 Section 4's six viability conditions, and the largest subfamily in the register at nine entries. Three of the nine, `MG-042`'s counterpart `MG-047` halt capability coverage, `MG-049` concurrent supervised stream count and `MG-050` nominal to effective posture divergence count, are `available` today because they require nothing beyond an inventory the founding case already possesses: the supervisor unit list, the fleet registry, and the completed viability assessment itself.

Table D4-8. Supervision viability metrics, MG-044 to MG-052.

MG-044 Intervention latency

QUESTION	From the moment a human decides to intervene, how long until work actually in progress stops?
OPERATIONAL DEFINITION	Elapsed time from an operator-initiated stop action, whether a console stop, a kill switch or an interrupt, to the moment the targeted agent process ceases taking new tool actions, established from an independent liveness signal rather than from the agent's own report that it has stopped. Excludes the time an operator spends deciding to intervene, which belongs to MG-054 and MG-053; excludes automated, non-human-initiated terminations.
FORMULA	t_work_stopped - t_operator_initiates_stop, per intervention
UNIT AND RANGE	seconds, 0 to unbounded
TYPE	KRI
HORIZON	lagging
GRANULARITY	per intervention
DATA SOURCE	Mixed. The operator-initiated timestamp is INFRASTRUCTURE-EMITTED from the console audit log; whether work has actually stopped requires an independent tool-call or process signal, since asking the agent whether it stopped is agent self-report.
COLLECTION POINT	The console stop-action write path, joined to the process manager or hook substrate signal marking the last tool call
SAMPLING FREQUENCY	per intervention
DIRECTION	lower-better
BASELINE	not computable as an empirical distribution, because zero human stops, pauses or overrides were recorded across 46 days of console audit logs; by design, the one mechanism that attempts to halt work in progress opens by pasting a prompt asking the agent to do more work and grants it up to 35 seconds of further unrestricted tool access before killing it, while the console's interrupt is a bracketed paste plus a newline that queues for the next turn and stops nothing, so its designed latency is unbounded
TARGET	shorter than harm completion time (MG-045) for the action class being interrupted, which is the requirement MG-046 tests directly; not set as an absolute figure because it varies by class
GUARDRAIL SENSITIVITY	GR-L09-01 session stop affordance lowers it once implemented to halt rather than to prompt-then-kill; GR-L10-02 always reachable intervention affordance lowers the human-decision portion feeding into this measurement; GR-L09-10 restart-always raises it in effect, since a killed process returns within seconds and can look like a stop while resuming the same trajectory; GR-L10-10 an intervention console measuring its own latency lowers it by making it visible and therefore improvable; GR-L09-11 global fleet pause under a separate principal lowers it structurally for the fleet-wide case
CONFOUNDERS	Which stop mechanism is invoked; whether the agent is mid-tool-call when the stop is issued; the definition of stopped, since a session going idle briefly before a restart-always respawn is not the same as work having genuinely ceased
GAMING MODES	Measuring latency to the stop command being accepted or queued rather than to work actually ceasing; defining stopped as the pane going idle rather than as no further mutating action occurring, which a brief restart-always gap can satisfy before the session resumes
CORROBORATED BY	MG-046 latency margin; MG-068 wedged session detection latency, sharing the same instrumentation gap
READINESS	absent
INSTRUMENTATION GAP	A timed record of actual interventions, of which the founding case has none in 46 days, and an independent stopped-signal distinct from the agent's own report

MG-045 Harm completion time per action class

QUESTION	From the moment an action of a given class begins, how long until its effect is fully realised and no longer interceptable by any available halt mechanism?
OPERATIONAL DEFINITION	Elapsed time from action initiation to the point at which the effect is complete and no halt mechanism available on the host can prevent or reverse it, per action class. Distinct from MG-004 cycle time, which measures duration to confirmed effect rather than to the point of no return; the point of no return can precede confirmation. Excludes classes with no defined point of no return, such as pure reads.
FORMULA	per class, t_harm_complete - t_action_initiated, reported as an order-of-magnitude estimate given the founding case's instrumentation rather than a precise distribution
UNIT AND RANGE	seconds, order of magnitude, 0 to unbounded
TYPE	KRI
HORIZON	lagging
GRANULARITY	per action class
DATA SOURCE	Derived analytically from the action's mechanism, such as file write duration, network round trip or process restart time, rather than measured from a timed incident population, since no incident with a recorded effect-completion boundary exists
COLLECTION POINT	The affected system's own state transition record, where one exists, read at the point closest to irreversibility rather than at tool return
SAMPLING FREQUENCY	per action class, reviewed on architecture change
DIRECTION	higher-better
BASELINE	not computable as a distribution; the founding case establishes only that consequential actions on this host complete in milliseconds to seconds, echoing MG-004's baseline of file writes in milliseconds and cloud control plane calls in about one second, while the fastest independent detector on the host trips at 180 seconds plus up to 60 seconds of schedule granularity, so harm completion time is order-of-magnitude below the fastest detection latency across essentially the whole action class population
TARGET	not set as an absolute; the formal requirement is t_harm > t_detect + t_decide + t_act (MG-046), which this metric is one input to
GUARDRAIL SENSITIVITY	GR-L09-04 per action class rate limits raises it by slowing execution and buying intervention time; GR-L11-03 cooling off period raises it; GR-L01-06 just-in-time elevation raises it by inserting a step before the point of no return; GR-L06-06 dry run as the default raises it by deferring the real effect; for most classes this metric is intrinsic to the mechanism and does not move with guardrails, echoing MG-004's own caveat, and becomes a dependent variable only for classes carrying a deliberate friction control
CONFOUNDERS	Network and host latency; the size of the change; whether an effect propagates synchronously or asynchronously, since replication or DNS propagation lag can extend the true point of no return well beyond the initiating call's return
GAMING MODES	Measuring harm completion to the initiating call's return rather than to downstream propagation, which understates true propagation speed and makes the reported margin look larger, and therefore safer, than it actually is; defining completion as the tool's own success signal rather than the effect's real point of irreversibility
CORROBORATED BY	MG-004 cycle time per action class, using independent infrastructure timestamps for confirmation; MG-046 latency margin
READINESS	derivable
INSTRUMENTATION GAP	A per-class effect-completion event distinguishing the point of no return from confirmed completion, and a timed incident population to build a real distribution from; today only order-of-magnitude estimates exist

MG-046 Latency margin

QUESTION	Is the fastest available intervention faster than the harm the action being intervened on would cause?
OPERATIONAL DEFINITION	The signed difference between harm completion time for an action class and the sum of detection, decision and action latency for the fastest applicable intervention mechanism, formalising the supervision viability condition t_detect + t_decide + t_act < t_harm. Positive means supervision is viable in time for that class; negative means the harm completes before any available intervention can act. Excludes classes for which no intervention mechanism exists at all, which are reported separately as undefined rather than as a very negative margin.
FORMULA	latency_margin = t_harm - (t_detect + t_decide + t_act), per action class
UNIT AND RANGE	seconds, signed, unbounded in either direction
TYPE	KRI
HORIZON	leading
GRANULARITY	per action class
DATA SOURCE	Composed from MG-017, MG-025, MG-044 and MG-045; each component's own data source and limitation carries into this composite
COLLECTION POINT	Wherever the four component metrics are collected, combined at the point of assessment rather than at any single instrumentation point
SAMPLING FREQUENCY	reviewed on architecture change, and at least quarterly
DIRECTION	higher-better
BASELINE	negative across nearly the whole action class population; the fastest independent detector on this host trips at 180 seconds plus up to 60 seconds of schedule granularity, while consequential actions complete in milliseconds to seconds, so the margin is negative by roughly two to five orders of magnitude for the great majority of action classes; an exact per-class figure is not computable because MG-017, MG-044 and MG-045 are each themselves not computable as distributions in the founding case
TARGET	positive for every action class rated severe blast radius or irreversible; not set as a numeric floor because the underlying component metrics are not yet measured
GUARDRAIL SENSITIVITY	GR-L08-11 raises it by lowering t_detect through an independent event stream; GR-L10-03 bounded intervention latency commitment raises it by construction; GR-L09-04, GR-L11-03 and GR-L01-06 raise it by raising t_harm rather than by lowering detection or decision time; GR-L09-06 automatic suspension on anomaly raises it by removing the human decision step, t_decide, from the path entirely
CONFOUNDERS	Which of the three latency components dominates for a given class; overlapping incidents competing for the same operator attention, interacting with MG-049; a model version change altering how fast an agent completes a harmful sequence once begun
GAMING MODES	Computing the margin against mean detection latency rather than the tail, hiding a rare slow detection that determines real exposure; substituting a general liveness signal, the agent produced some output, for a harm-specific detector, which trips reliably but is not evidence the specific harm was caught in time
CORROBORATED BY	MG-044 intervention latency; MG-045 harm completion time per action class
READINESS	absent
INSTRUMENTATION GAP	None of the three component latencies is instrumented as a distribution today; this metric cannot be computed until they are

MG-047 Halt capability coverage

QUESTION	What proportion of live agent sessions and automations have any verified mechanism capable of halting work already in progress, as distinct from merely preventing the next action?
OPERATIONAL DEFINITION	Count of agent sessions and unattended automations for which a verified halt-in-progress mechanism exists, satisfying the supervision viability halt condition, divided by all live sessions and automations. A mechanism counts only if it stops the current turn or current run, not merely the next scheduled one.
FORMULA	sessions_and_automations_with_verified_halt / total_sessions_and_automations
UNIT AND RANGE	ratio, 0 to 1
TYPE	KCI
HORIZON	leading
GRANULARITY	per fleet configuration
DATA SOURCE	INFRASTRUCTURE-EMITTED, derived from the supervisor unit inventory, the console stop-path implementation and the automation inventory
COLLECTION POINT	A design and behavioural review of every supervisor unit, console stop path and scheduled automation, tested against actual behaviour rather than intent
SAMPLING FREQUENCY	on every configuration change, and at least monthly
DIRECTION	higher-better
BASELINE	at most 1 of 14 live sessions and 0 of 17 unattended automations: only one mechanism terminates a running child process at all, and it opens by pasting a prompt asking the agent to do more work and grants it up to 35 seconds of further unrestricted tool access before killing it; for the nine unit-supervised agents restart-always returns a fresh session within seconds, which restarts rather than halts; the console's interrupt is a bracketed paste that queues for the next turn and stops nothing; the kill switch, the message bus kill switch and the deployment gate all affect only the next action and never work in progress
TARGET	1.0; every session and automation able to take a consequential action should have a verified halt-in-progress path, since a control that only blocks the next action does not satisfy the halt condition
GUARDRAIL SENSITIVITY	GR-L09-01 raises it once implemented as a true halt rather than a prompt-then-kill; GR-L04-08 automatic termination on wedge detection raises it for the wedge case, since it terminates rather than resumes; GR-L09-10 lowers it, being the mechanism that defeats a halt by respawning; GR-L09-11 global fleet pause raises it by adding a fleet-wide halt; GR-L09-03 circuit breaker on repeated failure raises it for that specific failure mode
CONFOUNDERS	Fleet size and the mix of supervised versus unsupervised sessions; whether an automation's action is brief enough that halting mid-action is largely moot; changes to supervisor configuration made for unrelated reasons
GAMING MODES	Counting a mechanism that stops the next scheduled action as a halt, the exact conflation the founding case's own kill switches make; counting restart-always recovery as evidence the process stopped, when it in fact resumed within seconds and continued toward the same effect
CORROBORATED BY	MG-044 intervention latency; MG-069 unattended action path count, supplying the automation side of the denominator
READINESS	available
INSTRUMENTATION GAP	A formal per-mechanism test of whether it halts in-progress work or only gates the next call; today the classification above is a design read of documented behaviour, not a tested one

MG-048 Oversight signal independence share

QUESTION	What proportion of signals reaching the operator originate from a source other than the agent they concern?
OPERATIONAL DEFINITION	Count of oversight signals, notifications, alerts and dashboard entries in a window whose originating process runs under a principal other than the agent it concerns, divided by all oversight signals in the window. This measures the independence supervision viability condition as a rate rather than as a per-point judgement. Excludes signals with no identifiable originator.
FORMULA	independent_signals / total_signals
UNIT AND RANGE	ratio, 0 to 1
TYPE	KCI
HORIZON	leading
GRANULARITY	per fleet per week
DATA SOURCE	Mixed. The primary notification path is AGENT-EMITTED, since the agent decides when to call it about itself; watcher-originated signals such as heartbeats and wedge detections are nominally INFRASTRUCTURE-EMITTED, though watchers and agents currently share one operating system account.
COLLECTION POINT	The notification and watcher dispatch points, classified by originating process at the moment each signal is sent
SAMPLING FREQUENCY	weekly
DIRECTION	higher-better
BASELINE	not computed as an exact ratio for this register, though the founding case's own supervision determinant analysis names independence as the binding condition in 8 of 34 assessed supervision points; the raw signal volume in the observation window was 19,334 heartbeat and nudge events against 4 human-facing alerts, and the great majority of human-facing content originates from the agent's own decision to call the notification path, so the independent share of consequential, human-facing signal is low, though signal originator was not classified across the full volume for this figure
TARGET	1.0 for any signal used as the sole basis of an approval or intervention decision; a self-originated status update remains useful but must not be the only evidence behind a go or no-go decision
GUARDRAIL SENSITIVITY	GR-L08-11 raises it directly by adding a genuinely infrastructure-emitted stream; GR-L01-07 raises it, since a separate verifier identity is what makes a signal independent rather than nominally so; GR-L10-01 separation of notification from authorisation raises it; GR-L09-05 notification rate limiting does not move independence, only volume; GR-L01-01 lowers it, because a shared account collapses the distinction between watcher and agent that independence depends on
CONFOUNDERS	Which channel is counted, since SMS, console and memory notes differ; watcher coverage, since 8 of 14 sessions have no watcher at all and therefore emit no independent signal regardless of true state; alert volume policy, which can suppress independent signals without suppressing agent-originated ones
GAMING MODES	Routing an agent-authored status update through a watcher process so it is credited as independent while its content is still authored by the subject; increasing benign, independent heartbeat volume to dilute the ratio while leaving the informative, agent-originated content unchanged
CORROBORATED BY	MG-024 verification independence ratio, the same independence question applied to verification records; MG-057 signal to noise ratio of the oversight channel
READINESS	derivable
INSTRUMENTATION GAP	An originator classification applied uniformly across every notification and alert channel; today the channels are heterogeneous and unclassified as a set

MG-049 Concurrent supervised stream count

QUESTION	How many independent signal streams would an operator need to attend to at once to maintain the supervision the architecture claims?
OPERATIONAL DEFINITION	Count of distinct live agent sessions, watcher streams and scheduled automations whose output an operator would need to attend to concurrently, measured against the attention supervision viability condition that this count remain within one operator's capacity. Excludes streams that are purely informational duplicates of another already counted.
FORMULA	count of concurrently live sessions + active watcher streams + scheduled automations, at a given time
UNIT AND RANGE	count, integer >= 0
TYPE	KRI
HORIZON	leading
GRANULARITY	per fleet snapshot
DATA SOURCE	INFRASTRUCTURE-EMITTED, the tmux session inventory, the fleet registry and the cron, systemd timer and at job inventory
COLLECTION POINT	A fleet snapshot combining the session, watcher and automation inventories at a point in time
SAMPLING FREQUENCY	daily snapshot, reviewed weekly
DIRECTION	target-band
BASELINE	14 live agent sessions, a partial set of watchers covering only some of them, and 17 unattended automations, of which 16 are scheduled through cron, systemd timers and at, are concurrently live on this host, monitored in practice by a single operator over a single SMS-first channel; 8 of the 14 sessions have no stall detection at all because watcher scope is a hardcoded array of agent names, so the count of streams that would need attention if all were adequately watched exceeds what is watched today
TARGET	not set as a number; the defensible requirement is that the count stay within whatever capacity a single operator has been empirically shown to sustain, which has not been established for this operator or channel
GUARDRAIL SENSITIVITY	GR-L10-09 single channel aggregation lowers the operator-facing count by merging streams without lowering the underlying number of things happening; GR-L09-11 global fleet pause lowers effective concurrent load during an incident; GR-L04-11 concurrency cap on live agent sessions lowers the true count directly; GR-L10-06 operator attention budgeting is the mechanism that would make this count actionable rather than merely descriptive
CONFOUNDERS	Fleet growth over time; how many sessions are genuinely active versus idle at a given moment; whether aggregation reduces cognitive load or merely reduces message count
GAMING MODES	Aggregating many streams into one channel and reporting the channel count rather than the underlying stream count, which improves this metric's proxy while attention demand is unchanged; excluding automations with no human-facing output from the count, even though 7 of them inject input directly into a live session and therefore need supervision regardless of visible output
CORROBORATED BY	MG-057 signal to noise ratio of the oversight channel; MG-052 effective supervision coverage
READINESS	available
INSTRUMENTATION GAP	An empirically established single-operator capacity figure to compare the count against; none exists

MG-050 Nominal to effective posture divergence count

QUESTION	In how many of the architecture's claimed supervision points does what the human can actually do differ from what the design claims?
OPERATIONAL DEFINITION	Count of assessed supervision points at which the nominal posture, what the design or documentation claims, diverges from the effective posture, what a human can actually verify, do or stop, scored against the six supervision viability conditions. Excludes points not assessed; excludes divergences the same assessment closed before being counted.
FORMULA	count of divergent_points / total_supervision_points_assessed, reported as a count and a ratio, with the binding condition broken out per divergence
UNIT AND RANGE	count, integer 0 to points assessed; ratio 0 to 1
TYPE	KRI
HORIZON	lagging
GRANULARITY	per fleet assessment
DATA SOURCE	Derived from a structured review comparing declared control intent, drawn from the guardrail catalogue, against live verification of each mechanism; INFRASTRUCTURE-EMITTED to the extent the live verification is performed by a party other than the agent whose supervision is assessed
COLLECTION POINT	A structured supervision-point review, one row per claimed control, tested against actual behaviour
SAMPLING FREQUENCY	on architecture change, and at least quarterly
DIRECTION	lower-better
BASELINE	27 of 34 assessed supervision points diverge between nominal and effective posture; the binding condition is independence in 8 cases, latency in 8, halt in 3, authority in 3, attention in 2 and comprehensibility in 2, and in one case the claimed control does not exist at all
TARGET	0; every divergence is, by the founding case's own framing, a primary finding rather than a tolerable residual
GUARDRAIL SENSITIVITY	Guardrails close specific gaps rather than the aggregate: GR-L10-03 bounded intervention latency commitment addresses the latency condition; GR-L08-11 addresses independence; GR-L09-01 properly implemented addresses halt; GR-L09-11 addresses authority; GR-L10-06 addresses attention; GR-L10-04 addresses comprehensibility; no single guardrail lowers the aggregate count on its own, since the 34 assessed points span all six conditions
CONFOUNDERS	How many supervision points are enumerated to assess in the first place, since a narrower assessment scope mechanically lowers the count; assessor judgement on what counts as a claimed control at all
GAMING MODES	Narrowing the definition of a supervision point so fewer are assessed; declaring a point resolved once a guardrail is nominally present without re-verifying its effective behaviour, which is the exact nominal-versus-effective gap this metric exists to catch
CORROBORATED BY	MG-031 structural enforcement coverage; MG-052 effective supervision coverage
READINESS	available
INSTRUMENTATION GAP	None for the current snapshot, which is the founding case's own completed assessment; a repeatable version needs a standing point inventory and scheduled re-verification rather than a one-off review

MG-051 Approval evidence sufficiency

QUESTION	When a human is asked to approve an action, does the request carry independent evidence of the state, or only the agent's own summary?
OPERATIONAL DEFINITION	Approval requests whose presented material includes at least one independently sourced artefact, such as a diff, a log excerpt or a health check result, rather than solely the agent's prose description, divided by approval requests. Satisfies the comprehensibility supervision viability condition. Excludes requests with no recorded payload at all, which are counted separately as unassessable.
FORMULA	requests_with_independent_evidence / total_approval_requests
UNIT AND RANGE	ratio, 0 to 1
TYPE	KCI
HORIZON	leading
GRANULARITY	per approval request, aggregated monthly
DATA SOURCE	The deployment approval gate's request payload and the console's rendering of it; whether content is independently sourced or agent-authored is a property of the payload's origin, INFRASTRUCTURE-EMITTED as to what was displayed even where the content itself is AGENT-EMITTED
COLLECTION POINT	The approval gate's request-composition step, at the point the payload is assembled for the reviewer
SAMPLING FREQUENCY	per request, reviewed monthly
DIRECTION	higher-better
BASELINE	not computable as a ratio; the deployment approval gate's structural fact is that only a human identity token can write an approved status, but what evidence is shown to that human at decision time is not documented, and reviewer independence is instructional only, since no code compares the originating agent to the reviewing agent, and a production instance is on record of an agent recording its own peer review using another agent's token, which suggests that instance was not grounded in independently sourced evidence
TARGET	1.0 for every approval gating an action rated severe blast radius or irreversible
GUARDRAIL SENSITIVITY	GR-L10-04 raises it directly and is its definition; GR-L10-05 visual distinction between claimed and verified raises it by making the source of each item explicit; GR-L08-08 cross-check of claims against raw sources raises it by verifying the evidence before the request is composed; GR-L06-01 does not move it on its own, since a gate held by a separate principal says nothing about what evidence that principal is shown
CONFOUNDERS	The size and reviewability of the change, since a large diff can be technically present but practically unreviewable; reviewer time pressure; whether the evidence, even when present, is comprehensible to the reviewer
GAMING MODES	Attaching a diff that is technically independent evidence but too large or noisy to actually review, satisfying the letter of the metric without the intent; having the agent itself generate the independent evidence artefact, for example a log excerpt it selects and formats
CORROBORATED BY	MG-027 evidence citation completeness; MG-024 verification independence ratio
READINESS	absent
INSTRUMENTATION GAP	A payload-origin classification on every approval request the deployment approval gate renders; not currently captured

MG-052 Effective supervision coverage

QUESTION	What proportion of live agent sessions are, right now, under supervision that would satisfy all six viability conditions if invoked?
OPERATIONAL DEFINITION	Count of live agent sessions for which independence, latency, halt, attention, comprehensibility and authority all hold at assessment time, divided by live agent sessions. A session counts only if every condition holds; satisfying five of six does not count, because supervision viability is conjunctive rather than additive.
FORMULA	sessions_satisfying_all_six_conditions / live_sessions
UNIT AND RANGE	ratio, 0 to 1
TYPE	KCI
HORIZON	leading
GRANULARITY	per fleet assessment
DATA SOURCE	Derived from MG-047 through MG-051 assessed per session, combined at the point of assessment
COLLECTION POINT	A per-session supervision-point review applying all six viability conditions
SAMPLING FREQUENCY	on architecture change, and at least quarterly
DIRECTION	higher-better
BASELINE	0; the founding case's own supervision determinant analysis found 27 of 34 assessed supervision points diverging from at least one of the six conditions, and no session was found to satisfy all six simultaneously; separately, 8 of 14 sessions have no stall detection at all, which alone removes the independence and latency conditions for those sessions
TARGET	not set as a ratio target; the immediate target is a defined, non-zero count, since the current value is zero across the fleet
GUARDRAIL SENSITIVITY	This is the terminal aggregate of MG-047 through MG-051; GR-L08-11 raises it by closing the independence condition, GR-L10-03 by closing the latency condition, GR-L09-01 properly implemented by closing the halt condition, GR-L09-11 by closing the authority condition, GR-L10-06 by closing the attention condition and GR-L10-04 by closing the comprehensibility condition, but no single one of them raises it alone, because the conjunction requires all six conditions closed simultaneously for a given session
CONFOUNDERS	Fleet composition, since sessions differ in what automations and watchers cover them; assessment thoroughness; a session moving in and out of coverage as watchers and supervisors are added or removed
GAMING MODES	Counting a session as covered because a guardrail exists somewhere in its path without verifying all six conditions for that specific session; assessing only the well-supervised sessions and excluding the 8 unwatched ones from the denominator as not yet in scope
CORROBORATED BY	MG-050 nominal to effective posture divergence count; MG-047 halt capability coverage
READINESS	derivable
INSTRUMENTATION GAP	A per-session, per-condition standing assessment refreshed on configuration change, rather than the one-off review the founding case performed

MG-046 latency margin formalises D2's VC2 condition, `t_detect + t_decide + t_act < t_harm`, as a single signed number, and its baseline is the sharpest figure in the register: negative by roughly two to five orders of magnitude across nearly the whole action class population, because consequential actions complete in milliseconds to seconds while the fastest independent detector trips at 180 seconds plus schedule granularity EV-039. MG-047 halt capability coverage supplies the mechanism behind that gap: at most one of fourteen live sessions and zero of seventeen unattended automations have a verified mechanism that halts work already in progress rather than merely gating the next action, and the one mechanism that comes closest opens by granting the agent up to 35 further seconds of unrestricted tool access before terminating it EV-039.

3.5 Human oversight quality

MG-053 to MG-058 ask not whether oversight is structurally possible but whether it is actually exercised. Across the six entries the founding case's recurring baseline is zero or near-zero, not because the metrics are uncomputable but because the observed count genuinely is zero: zero interventions, zero approvals, zero rejections, across 46 days of console audit logs EV-034.

Table D4-9. Human oversight quality metrics, MG-053 to MG-058.

MG-053 Operator intervention rate

QUESTION	How often does the operator actually stop, pause, reject or override an agent based on what it is doing?
OPERATIONAL DEFINITION	Count of operator actions recorded in the console audit log or an equivalent record that stop, pause, reject or override an agent's in-progress or proposed work, per unit time. Excludes scheduled or automated nudges, which are machine-originated by definition; excludes routine status queries with no corrective effect.
FORMULA	intervention_events / observation_days
UNIT AND RANGE	count per day, 0 to unbounded
TYPE	KRI
HORIZON	lagging
GRANULARITY	per fleet per week
DATA SOURCE	INFRASTRUCTURE-EMITTED, the console audit log, restricted to the authenticated operator principal
COLLECTION POINT	The console audit log, filtered to stop, pause, reject and override action types
SAMPLING FREQUENCY	weekly
DIRECTION	target-band
BASELINE	0 across 46 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, and zero update approvals, zero rejections and zero kill-switch activations were recorded in the same window
TARGET	not set as a level; a rate of zero is consistent either with flawless agent behaviour or with an oversight channel nobody is using, and the founding case's own claim divergence finding of 0.611 makes the first reading indefensible, so the honest target is that the rate become measurably non-zero and then be read jointly with MG-022
GUARDRAIL SENSITIVITY	GR-L10-07 live terminal view raises it by giving the operator something concrete to react to; GR-L10-04 raises it by making evidence, not summary, the trigger; GR-L09-07 acknowledge-to-stop raises it structurally, since it requires an operator action to proceed past an escalation; GR-L10-08 automated answering of interactive prompts lowers it directly, because it removes the human from exactly the decision this metric counts; GR-L09-05 notification rate limiting can lower it by suppressing the signal that would have prompted intervention
CONFOUNDERS	Operator availability and working pattern; whether the alerting channel actually reaches the operator, since a message sent is not a message read; trust calibration, since an operator who has never seen a caught error has no occasion to intervene
GAMING MODES	Counting a scheduled or automated nudge as an intervention, when the founding case explicitly treats it as machine-originated; lowering alert volume so fewer occasions for intervention arise, which raises the apparent quality of the agents by starving the metric rather than by improving behaviour
CORROBORATED BY	MG-056 alert acknowledgement rate; MG-022 claim divergence rate, read jointly per the target note above
READINESS	available
INSTRUMENTATION GAP	None for the console-mediated channel; direct pane typing and any out-of-band operator action are not captured by any log

MG-054 Approval turnaround time

QUESTION	How long does an approval request wait for a human decision?
OPERATIONAL DEFINITION	Elapsed time from an approval request being recorded by the deployment approval gate to the recorded approve or reject decision. Excludes requests withdrawn before a decision; excludes requests still pending at window close, which are reported separately as a censored count.
FORMULA	t_decision - t_request_recorded, median over decided requests
UNIT AND RANGE	seconds, 0 to unbounded
TYPE	KPI
HORIZON	lagging
GRANULARITY	per approval request, aggregated weekly
DATA SOURCE	INFRASTRUCTURE-EMITTED, the deployment approval gate's own request and decision timestamps
COLLECTION POINT	The approval gate's request-write and decision-write paths
SAMPLING FREQUENCY	weekly
DIRECTION	target-band
BASELINE	not computable; the founding case recorded zero update approvals and zero rejections across the observed window, so no decided request exists to time; the deployment approval gate's own approval history was not populated during the observation period
TARGET	not set; the useful target is a band, since a decision made implausibly fast suggests no real review occurred, which is the concern MG-051 addresses directly, and a decision that takes too long blocks legitimate work, and no empirical distribution exists yet to anchor a band
GUARDRAIL SENSITIVITY	GR-L10-02 always reachable intervention affordance lowers it; GR-L06-09 change size limit triggering escalation raises it for large changes by design; GR-L10-06 operator attention budgeting lowers it by prioritising the queue; GR-L06-01 does not move it directly, since holding the gate with a separate principal says nothing about how fast that principal decides
CONFOUNDERS	Request volume competing for the same operator attention, interacting with MG-049; time of day and business hours; the comprehensibility of what is presented (MG-051), since an unreadable request is decided more slowly, or waved through faster with less scrutiny
GAMING MODES	Batching several changes into one request so only one turnaround clock runs; treating a default-timeout auto-decision, should one ever be added, as a human decision for the purposes of this clock
CORROBORATED BY	MG-055 approval rejection rate; MG-051 approval evidence sufficiency
READINESS	absent
INSTRUMENTATION GAP	A populated approval history; today the gate exists but the observed window shows no decided requests

MG-055 Approval rejection rate

QUESTION	What proportion of approval requests does the human actually decline?
OPERATIONAL DEFINITION	Approval requests recorded as rejected, divided by approval requests decided, approved plus rejected, excluding requests still pending at window close.
FORMULA	rejected_requests / (approved_requests + rejected_requests)
UNIT AND RANGE	ratio, 0 to 1
TYPE	KRI
HORIZON	lagging
GRANULARITY	per approval request, aggregated monthly
DATA SOURCE	INFRASTRUCTURE-EMITTED, the deployment approval gate's own decision record
COLLECTION POINT	The approval gate's decision-write path
SAMPLING FREQUENCY	monthly
DIRECTION	target-band
BASELINE	not computable; zero rejections and zero approvals were recorded across the observation window, so the ratio has no denominator
TARGET	not set as a level; a rate of exactly zero over a long window is itself the signal this register treats as adverse, for the same reason given under MG-053, namely that it is not distinguishable from an unused gate
GUARDRAIL SENSITIVITY	GR-L10-04 raises it plausibly, since evidence-based requests are more likely to surface a real defect than a summary is; GR-L08-07 adversarial second opinion raises it, since adversarial review finds more to reject; GR-L01-07 raises it by making the reviewer genuinely independent of the requester; GR-L06-01 does not move it on its own for the reason given under MG-054
CONFOUNDERS	The quality of work submitted for approval, itself downstream of everything else in this register; reviewer fatigue over a long queue; social cost of rejecting an agent's or a colleague's work, which may still bias an operator toward approving even where the reviewed party is not human
GAMING MODES	Submitting only work already known to be safe for formal approval while riskier work bypasses the gate through the ordinary-restart path the founding case documents as having shipped unapproved code twice; withdrawing a request likely to be rejected and resubmitting a revised version, which never appears as a rejection
CORROBORATED BY	MG-054 approval turnaround time; MG-007 change failure rate, read against approvals to see whether approved work fails less than unapproved work
READINESS	absent
INSTRUMENTATION GAP	A populated approval history with outcomes, and a way to detect submissions that bypassed the gate entirely, which the founding case's ordinary-restart path allows

MG-056 Alert acknowledgement rate

QUESTION	What proportion of alerts sent to the operator receive any recorded acknowledgement?
OPERATIONAL DEFINITION	Alerts for which an operator acknowledgement, reply or console action is recorded within a defined window, divided by alerts sent. Excludes informational heartbeats not requiring a decision, which are reported separately; excludes alerts suppressed before delivery by a deduplication window.
FORMULA	acknowledged_alerts / alerts_sent
UNIT AND RANGE	ratio, 0 to 1
TYPE	KPI
HORIZON	lagging
GRANULARITY	per fleet per week
DATA SOURCE	Mixed. Alert dispatch is INFRASTRUCTURE-EMITTED from the notification path; acknowledgement is not currently captured at all on the SMS-first channel, which records delivery, not receipt or action.
COLLECTION POINT	The notification dispatch point for the denominator; an acknowledgement event, which does not currently exist, for the numerator
SAMPLING FREQUENCY	weekly
DIRECTION	higher-better
BASELINE	not computable as a ratio; the observation window recorded 4 human-facing alerts against 19,334 heartbeat and nudge events, but acknowledgement was not separately tracked for those 4, so whether any were read, let alone acted on, is not established; the 46-day zero-intervention finding is consistent with, but does not prove, zero acknowledgement
TARGET	1.0 for any alert rated as requiring an operator decision, as distinct from an informational heartbeat, which need not be acknowledged
GUARDRAIL SENSITIVITY	GR-L09-07 acknowledge-to-stop raises it structurally, since it makes acknowledgement a precondition for the escalation to stop occurring, forcing the record to exist; GR-L09-08 escalation ladder with timeout raises it by re-sending on non-acknowledgement; GR-L10-01 separation of notification from authorisation does not move it directly; GR-L09-05 notification rate limiting can raise the measured ratio by suppressing duplicate alerts that would otherwise inflate the denominator without adding decisions
CONFOUNDERS	Channel reliability, since an SMS that fails silently is indistinguishable from one ignored; alert volume and fatigue, given the 19,334 to 4 ratio of noise to signal already observed; time of day
GAMING MODES	Counting delivery, message sent, as acknowledgement, message actioned, which the SMS channel cannot currently distinguish; lowering alert volume so fatigue-driven non-acknowledgement has fewer occasions to occur, improving the ratio while true attentiveness is unchanged
CORROBORATED BY	MG-057 signal to noise ratio of the oversight channel; MG-053 operator intervention rate
READINESS	absent
INSTRUMENTATION GAP	An acknowledgement event distinct from message delivery; the current SMS-first channel records dispatch, not receipt or action

MG-057 Signal to noise ratio of the oversight channel

QUESTION	Of everything sent to the operator, what proportion is decision-relevant rather than routine status noise?
OPERATIONAL DEFINITION	Count of human-facing alerts requiring a decision, divided by total signal volume on the same channel in the same window, alerts plus heartbeats plus nudges.
FORMULA	decision_relevant_signals / total_signals
UNIT AND RANGE	ratio, 0 to 1
TYPE	KCI
HORIZON	leading
GRANULARITY	per fleet per week
DATA SOURCE	INFRASTRUCTURE-EMITTED, the notification and watcher dispatch logs
COLLECTION POINT	The notification dispatch point, classified by whether the signal requires an operator decision
SAMPLING FREQUENCY	weekly
DIRECTION	higher-better
BASELINE	4 human-facing alerts against 19,334 heartbeat and nudge events in the observation window, a ratio of approximately 0.0002, meaning noise outnumbers decision-relevant signal by roughly 4,800 to 1
TARGET	not set as an absolute; the defensible aim is that the ratio rise without the absolute count of decision-relevant signals falling, since raising the ratio by suppressing heartbeats alone would remove the liveness evidence heartbeats also serve
GUARDRAIL SENSITIVITY	GR-L09-05 notification rate limiting and deduplication raises it directly by cutting duplicate noise; GR-L10-09 single channel aggregation can raise it if aggregation summarises rather than merely relays; GR-L11-06 snooze window raises it by suppressing repeated identical alerts; GR-L07-08 runtime action classification could raise it by letting routine, already-classified events be filtered before they reach the channel
CONFOUNDERS	Fleet size, since more agents produce more heartbeats at a fixed per-agent rate; the definition of decision-relevant, itself a judgement; watcher coverage, since the 8 unwatched sessions contribute no heartbeats and mechanically raise the ratio without anyone deciding they were less important to watch
GAMING MODES	Reclassifying routine heartbeats as decision-relevant to inflate the numerator without adding real decisions; suppressing heartbeats to shrink the denominator, which also removes the liveness signal that lets an operator distinguish a healthy quiet agent from a wedged one
CORROBORATED BY	MG-056 alert acknowledgement rate; MG-068 wedged session detection latency, since heartbeat suppression trades off against this metric
READINESS	available
INSTRUMENTATION GAP	A standing decision-relevance classification applied consistently as the fleet and its automations change; today the 4-versus-19,334 figure is a point-in-time count

MG-058 Automated answering of human decision points

QUESTION	How often is a prompt or decision point built for a human answered automatically instead?
OPERATIONAL DEFINITION	Count of interactive permission prompts, resume prompts or other decision points designed for a human answered by an automation rather than by a human keystroke, divided by all occurrences of that decision point. Excludes prompts answered by a configuration set in advance by a human with documented, scoped consent, which is a different control than an unattended automatic answer.
FORMULA	auto_answered_prompts / total_prompt_occurrences
UNIT AND RANGE	ratio, 0 to 1
TYPE	KRI
HORIZON	leading
GRANULARITY	per automation per month
DATA SOURCE	AGENT-ADJACENT and INFRASTRUCTURE-EMITTED in combination: the automation's firing is a process event, but its target, a live agent session, and its effect are both agent-adjacent, and the founding case does not log the automation's firing frequency separately from the sessions it touches
COLLECTION POINT	The automation's own trigger and keystroke-injection point, logged independently of the session it targets
SAMPLING FREQUENCY	monthly
DIRECTION	lower-better
BASELINE	not computable as a ratio for the full population, but the mechanism is confirmed present: one of the founding case's automations automatically answers the tooling's own interactive permission prompt on the agent's behalf, and this is one of seven automations that inject text directly into a live agent's input stream with no human keystroke; the frequency of that specific automation's firing was not counted for this register
TARGET	0; a decision point built for a human to answer, answered instead by an unattended automation, is exactly the action the founding case's own action class register names as a distinct, severe, level-III action
GUARDRAIL SENSITIVITY	GR-L10-08 is this mechanism itself and raises the count by construction, being the guardrail catalogue's own entry for automated answering of interactive prompts; GR-L09-08 escalation ladder with timeout lowers it if it routes to a human rather than auto-answering; GR-L01-06 just-in-time elevation lowers it by requiring a genuine approver at the elevation step rather than an automation; GR-L07-06 structural impossibility of self-widening lowers it for the subset of prompts that would otherwise widen an agent's own scope
CONFOUNDERS	How many prompt types exist in the tooling and how many of them this automation targets; whether the automation exists to solve a genuine operational problem, an unattended session stalling on a prompt with nobody present
GAMING MODES	Narrowing the definition of decision point to exclude the specific prompt this automation answers, so the count reads as zero by definition rather than by absence of the behaviour; disabling logging on the automation so its firing leaves no record even though the behaviour continues
CORROBORATED BY	MG-069 unattended action path count, which this event class belongs to; MG-050 nominal to effective posture divergence count, since the authority condition is directly compromised by this mechanism
READINESS	derivable
INSTRUMENTATION GAP	A firing log for the specific automation, and a formal inventory of every interactive prompt type in the tooling so the denominator can be established

MG-057 signal to noise ratio of the oversight channel is the subfamily's most legible figure: 4 human-facing alerts against 19,334 heartbeat and nudge events in the observation window, a ratio of approximately 0.0002, or noise outnumbering decision-relevant signal by roughly 4,800 to 1 EV-039. MG-058 automated answering of human decision points records a confirmed mechanism rather than an estimated rate: one automation answers the tooling's own interactive permission prompt on the

agent's behalf, one of seven automations that inject text directly into a live session with no human keystroke EV-028, though how often that specific automation fires was not separately counted for this register.

3.6 Control effectiveness

MG-059 to MG-063 ask whether a control that exists on paper holds in practice: whether it can be bypassed, whether it fails open or closed, and how much of what should be defence in depth is actually one dependency wearing several names.

Table D4-10. Control effectiveness metrics, MG-059 to MG-063.

MG-059 Control bypass rate

QUESTION	How often is a control on the books circumvented rather than satisfied?
OPERATIONAL DEFINITION	Count of actions in which the actor achieved an effect a control was intended to prevent, via a path the control does not cover, divided by actions in which the control was applicable. Excludes actions the control genuinely prevented; excludes paths not yet discovered, which are a limitation of the metric rather than evidence of a zero rate.
FORMULA	bypassed_applicable_actions / applicable_actions
UNIT AND RANGE	ratio, 0 to 1
TYPE	KRI
HORIZON	leading
GRANULARITY	per control per quarter
DATA SOURCE	Mixed. Discovery of a bypass is typically the product of a deliberate review or an incident, not a routine signal; on this host the confirmed instances were found by architectural review rather than by an automated bypass detector.
COLLECTION POINT	A structured bypass review per control, reading the guardrail catalogue's own bypass_analysis field against the live host
SAMPLING FREQUENCY	quarterly, and on every control change
DIRECTION	lower-better
BASELINE	not computable as a ratio; the founding case documents specific confirmed bypass paths rather than a counted population, for example that the update plane's internal token routes are documented as contained because the reverse proxy does not forward them, while the proxy in fact has a single catch-all location with no exclusions, so that containment is not implemented at all, and that the process manager re-executes whatever sits on the shared working tree regardless of deployment approval gate status, with two incidents on record of unapproved code shipping through ordinary restarts
TARGET	0 for any control classified structural in the guardrail catalogue; a non-zero bypass rate on a structural control falsifies its classification
GUARDRAIL SENSITIVITY	GR-L07-06 structural impossibility of self-widening lowers it by removing the path rather than gating it; GR-L04-06 tool-layer allowlist lowers it for shell-mediated bypass specifically; GR-L02-06 destination allowlist lowers it for egress-mediated bypass; every guardrail catalogue entry whose bypass_analysis field documents a route lowers this metric once that route is closed, and this metric is elevated in exactly the cases documented above while the route remains open
CONFOUNDERS	How thoroughly bypass paths have been searched for, since an unsearched-for bypass is invisible to this metric as much as to the control itself; agent or actor sophistication; whether the bypass is deliberate or incidental, as in the restart-path case, which was not adversarial but an architectural gap
GAMING MODES	Testing a control only through its intended path and declaring it effective, precisely how the reverse-proxy containment claim in the founding case was made before the catch-all route was found; fixing one discovered bypass and declaring the control secure without searching for others
CORROBORATED BY	MG-031 structural enforcement coverage; MG-033 boundary crossing success rate
READINESS	derivable
INSTRUMENTATION GAP	A systematic bypass search per control, repeated on change, rather than the incidental discovery the founding case relies on

MG-060 Fail closed ratio

QUESTION	When a control's enforcement mechanism itself fails, does the system default to blocking the action or to permitting it?
OPERATIONAL DEFINITION	Count of guardrail catalogue entries with a recorded fail_state of fail-closed, divided by controls assessed, reported overall and restricted to controls guarding action classes rated severe blast radius or irreversible.
FORMULA	count(fail_state = fail-closed) / count(controls_assessed)
UNIT AND RANGE	ratio, 0 to 1
TYPE	KCI
HORIZON	leading
GRANULARITY	per host configuration
DATA SOURCE	INFRASTRUCTURE-EMITTED, the guardrail catalogue's own fail_state field, cross-checked against documented behaviour rather than intent where possible
COLLECTION POINT	The guardrail catalogue itself, read at the fail_state column
SAMPLING FREQUENCY	on every configuration change, and at least quarterly
DIRECTION	higher-better
BASELINE	derivable directly from the guardrail catalogue's fail_state field; a specific documented fail-open instance is the shared operating system account itself, whose standing condition fails open by design, since a defect or prompt injection in one agent executes with the authority of every other agent and of the console with nothing distinguishing which agent acted; the update kill switch is also effectively fail-open in practice, since the governed account can remove the sentinel file with no authentication and no audit entry
TARGET	1.0 restricted to controls guarding severe or irreversible action classes; fail-open is acceptable only for controls guarding reversible, low-consequence classes where availability legitimately outweighs caution
GUARDRAIL SENSITIVITY	GR-L01-02 converts the identity layer from fail-open to a meaningfully closed state; GR-L07-07 default deny for unclassified actions is fail-closed by construction and raises the ratio directly; GR-L09-02 kill switch behaviour depends on whether removal requires authentication, currently it does not; GR-L04-09 prohibition of the permission-skipping flag in defined contexts would convert a currently fail-open default, skip enabled by default, to fail-closed
CONFOUNDERS	Which controls are in scope for assessment; a control's fail state can differ by failure mode, for example a crash may fail closed while a misconfiguration fails open, and the catalogue records one value per control
GAMING MODES	Classifying a control as fail-closed based on its intended design rather than its verified behaviour under an actual failure of the enforcement mechanism; scoping the assessed population to exclude the controls already known to fail open
CORROBORATED BY	MG-063 single point of failure count; MG-031 structural enforcement coverage
READINESS	available
INSTRUMENTATION GAP	None for the catalogue-based figure, directly computable today; a live fault-injection test to verify the recorded fail_state matches actual behaviour does not exist

MG-061 Control signal emission coverage

QUESTION	What proportion of controls emit any signal, independent of the agent, when they act or when they are bypassed?
OPERATIONAL DEFINITION	Count of guardrail catalogue entries with a non-empty, independently sourced emitted_signal, divided by controls assessed. Excludes an emitted signal that is never collected or retained anywhere, which is treated as equivalent to no signal for this metric's purpose.
FORMULA	count(emitted_signal populated and independently sourced) / count(controls_assessed)
UNIT AND RANGE	ratio, 0 to 1
TYPE	KCI
HORIZON	leading
GRANULARITY	per host configuration
DATA SOURCE	INFRASTRUCTURE-EMITTED, the guardrail catalogue's own emitted_signal field
COLLECTION POINT	The guardrail catalogue, read at the emitted_signal column, cross-checked against whether the named signal is actually collected
SAMPLING FREQUENCY	on every configuration change, and at least quarterly
DIRECTION	higher-better
BASELINE	derivable from the guardrail catalogue's emitted_signal field; a documented null case is the shared operating system account entry itself, whose emitted_signal field states that no signal today distinguishes one agent from another, since process accounting, file ownership and audit records all name the same principal
TARGET	1.0 for controls guarding action classes rated severe blast radius or irreversible, since a control that acts silently cannot be verified as having acted at all
GUARDRAIL SENSITIVITY	GR-L08-11 raises it directly by adding an infrastructure-emitted stream; GR-L01-02 raises it for identity-dependent signals, since distinct principals are a precondition for an attributable signal; GR-L08-13 currently does not raise it despite firing on every tool call, because the transcript hook substrate is registered for context injection rather than enforcement and is designed to always exit successfully, so its output is telemetry that nothing consumes as a control signal
CONFOUNDERS	Whether emitted signals are actually collected and retained anywhere, as distinct from theoretically emittable; log rotation and retention windows that let an emitted signal exist momentarily and then vanish; which controls were assessed
GAMING MODES	Counting a signal that is emitted but never collected or monitored as coverage, when in practice it is equivalent to no signal; pointing to the hook substrate's high event volume as evidence of coverage while its always-succeed design means it never actually reports a control decision
CORROBORATED BY	MG-048 oversight signal independence share; MG-064 audit record completeness
READINESS	available
INSTRUMENTATION GAP	A distinction, not currently recorded in the catalogue at the emission level, between a signal that is emitted and one that is emitted, collected and actually monitored

MG-062 Defence in depth factor

QUESTION	For a given action class, how many independent controls would an actor have to defeat to reach the prohibited effect?
OPERATIONAL DEFINITION	For each action class, the count of distinct guardrail catalogue entries binding that class whose enforcement_class is structural or conditional, adjusted so that controls sharing a common defeating condition, for example two controls both defeated by the same passwordless root fact, are counted as one effective layer rather than two.
FORMULA	per class, count of independent_effective_layers, where independence means no single fact or credential defeats more than one of the counted layers
UNIT AND RANGE	count, integer >= 0
TYPE	KCI
HORIZON	leading
GRANULARITY	per action class
DATA SOURCE	Derived by cross-referencing the coverage-matrix.csv per-class control count against the guardrail catalogue's bypass_analysis field for each cited control
COLLECTION POINT	A structured cross-reference of the coverage matrix and the bypass_analysis field, per action class
SAMPLING FREQUENCY	on architecture change, and at least quarterly
DIRECTION	higher-better
BASELINE	not computable as a per-class count for this register block; coverage-matrix.csv performs the adjacent per-class count of structural, conditional and instructional guardrails, but the independence adjustment this metric requires, discounting layers that share a common bypass, is not applied there, and the founding case's single-account, passwordless-root condition is documented to defeat a large but uncounted share of nominally distinct controls simultaneously
TARGET	at least 2 independent layers for every action class rated severe blast radius or irreversible, consistent with a defence in depth principle; not set as a specific number without the independence-adjusted count above
GUARDRAIL SENSITIVITY	Every layer added at a genuinely different point, identity, network, filesystem or process, raises it, but only if it does not share the passwordless-root or shared-account bypass; GR-L01-01 lowers the effective count of every other control that assumes account separation, which is most of the catalogue, making the adoption of GR-L01-02 the single highest-leverage move for this metric, since it re-qualifies controls that already exist as independent
CONFOUNDERS	Which classes are analysed; the definition of sharing a bypass, which requires the bypass_analysis field of each candidate control to be read and cross-referenced by hand; catalogue completeness
GAMING MODES	Counting controls at the same layer as independent because they carry different names, when in fact they share an implementing artefact the same account can write; adding cosmetic controls that are real in name but immediately defeated by the same standing condition as everything else
CORROBORATED BY	MG-063 single point of failure count, this metric's complement; MG-031 structural enforcement coverage
READINESS	derivable
INSTRUMENTATION GAP	A systematic cross-reference of bypass_analysis fields across all controls binding a given action class, to identify shared single points of failure before counting layers as independent; the raw material exists but the cross-reference has not been performed

MG-063 Single point of failure count

QUESTION	How many single facts, credentials or conditions, if compromised or removed, would simultaneously defeat more than one control?
OPERATIONAL DEFINITION	Count of distinct facts, a credential, an account property, a configuration file, identified in the guardrail catalogue's bypass_analysis field as the defeating condition for more than one catalogue entry.
FORMULA	count of distinct shared_defeating_conditions appearing in more than one entry's bypass_analysis field
UNIT AND RANGE	count, integer >= 0
TYPE	KRI
HORIZON	leading
GRANULARITY	per host configuration
DATA SOURCE	Derived from a structured extraction of the guardrail catalogue's bypass_analysis text across all 134 entries
COLLECTION POINT	A structured extraction and clustering of bypass_analysis text, across the whole catalogue
SAMPLING FREQUENCY	on architecture change, and at least quarterly
DIRECTION	lower-better
BASELINE	at least one confirmed, high-multiplicity instance: the account's passwordless root and the shared operating system account condition are named as the defeating or enabling condition across a large share of the catalogue's present-population entries, since almost every present control on this host assumes an account boundary that does not exist; a precise count across all 134 entries was not tabulated for this register block
TARGET	not set as zero, since some shared dependency is unavoidable in any single-host architecture; the defensible aim is that no single point of failure defeat controls spanning more than one guardrail layer, L1 through L12, since that cross-layer collapse is what removes defence in depth entirely
GUARDRAIL SENSITIVITY	GR-L01-02 per-agent workload identity and the removal of standing passwordless root together implied by GR-L01-09's remediation both lower this count directly, since they are the two conditions most frequently named across the catalogue's bypass_analysis fields; GR-L01-01 is the standing condition that keeps the count high
CONFOUNDERS	Catalogue granularity, since splitting one control into several sub-controls that all share the same underlying dependency inflates the apparent layer count without changing the true single point of failure; assessor thoroughness in reading every bypass_analysis field
GAMING MODES	Adding new controls that nominally sit at different layers but are documented, in their own bypass_analysis text, as defeated by the same root or account condition, which improves MG-031 and MG-062 optics without reducing this count; fixing the single point of failure for one layer while leaving it in place for others, and reporting only the fixed layer
CORROBORATED BY	MG-062 defence in depth factor; MG-030 access exceeds mandate count, since the same root and account facts drive both
READINESS	derivable
INSTRUMENTATION GAP	A structured extraction of the defeating condition named in every bypass_analysis field, clustered to identify shared conditions, rather than the qualitative reading performed here

MG-060 fail closed ratio and MG-061 control signal emission coverage are both available, read directly off D3's guardrail catalogue fields, and both surface the same underlying condition from different angles: the shared operating system account is documented as a specific fail-open instance, since a defect in one agent executes with the authority of every other agent and of the console with nothing distinguishing which one acted EV-003, and the transcript hook substrate, despite firing on every tool call, is recorded as not raising MG-061 because it is contractually required to exit successfully and is therefore telemetry that nothing consumes as a control signal EV-026. MG-063 single point of failure count names that same account and its root privilege as the defeating condition shared across a large but, in this register, uncounted share of the catalogue's present-population entries, which is why GR-

`L01-02` per-agent workload identity recurs across the subfamily's `guardrail_sensitivity` fields more than any other single control, named in four of the five entries, `MG-060` through `MG-063`, as the change that re-qualifies controls already built as independent of one another.

3.7 Auditability

`MG-064` to `MG-067` ask whether the record, once written, can be trusted to stay what it was written as. This is the subfamily where the founding case's `MG-065` finding is bluntest: zero of the governance stores enumerated are structurally unwritable by the party they record.

Table D4-11. Auditability metrics, MG-064 to MG-067.

MG-064 Audit record completeness

QUESTION	What proportion of consequential actions leave any audit record at all?
OPERATIONAL DEFINITION	Count of executed actions belonging to a mutating action class for which a corresponding audit log entry exists, divided by executed mutating actions. Excludes read-only actions; excludes actions with no assignable class, which fall under MG-036.
FORMULA	audited_mutating_actions / total_mutating_actions
UNIT AND RANGE	ratio, 0 to 1
TYPE	KCI
HORIZON	leading
GRANULARITY	per agent per week
DATA SOURCE	Mixed. The console audit log is INFRASTRUCTURE-EMITTED for operator actions specifically; nothing on this host audits what an agent does, so agent-side coverage is currently zero by construction rather than measured.
COLLECTION POINT	Wherever a mutating route or tool call fires, tested for whether an audit write accompanies it
SAMPLING FREQUENCY	weekly
DIRECTION	higher-better
BASELINE	not computable as a ratio; the console audit log records only what operators do through the console, and auditing there is a manual per-handler call, so a new route is unaudited by default; nothing on this host audits what an agent does, so for the entire population of agent-executed actions the numerator is effectively 0, while the denominator, executed mutating actions, is itself not counted (MG-002)
TARGET	1.0 for actions in mandate levels II and III; a consequential action with no audit record is unverifiable by definition, the precondition MG-023, MG-024 and MG-027 all depend on
GUARDRAIL SENSITIVITY	GR-L08-01 raises it for the subset covered by the model runtime transcript; GR-L08-13 does not raise it despite firing on every call, because it is telemetry only and is designed to always exit successfully; GR-L05-02 raises it for fleet registry writes specifically; GR-L08-02 raises it partially for destructive console operations only, and the console audit log itself is a manual per-handler call, so new routes are unaudited by default; GR-L07-01 declarative policy as data raises it by giving every action a class an auditor can check against
CONFOUNDERS	Which stores are counted as part of the audit corpus; whether a store the actor itself can also write still counts as an audit record, which this metric does not distinguish from MG-065's immutability question; route or handler additions that silently widen the unaudited surface
GAMING MODES	Counting the existence of the model runtime transcript, agent-owned and rewritable, as equivalent audit coverage to an independent log; adding a log line to a new handler that fires but writes into a store the same handler's account can also erase, satisfying completeness while leaving immutability, MG-065, unaddressed
CORROBORATED BY	MG-065 audit record immutability, since completeness with no immutability is not durable evidence; MG-029 self-attestation share of the record
READINESS	derivable
INSTRUMENTATION GAP	A uniform, non-manual audit hook applied to every mutating route and every agent tool call, rather than the current per-handler opt-in

MG-065 Audit record immutability

QUESTION	Once written, can an audit record be altered or deleted by the party it records?
OPERATIONAL DEFINITION	Count of audit or evidence stores that are structurally unwritable by the principal they record, divided by audit or evidence stores assessed. Excludes stores with no defined recorded principal, such as pure system logs unrelated to agent action.
FORMULA	count(stores_unwritable_by_recorded_principal) / count(stores_assessed)
UNIT AND RANGE	ratio, 0 to 1
TYPE	KCI
HORIZON	leading
GRANULARITY	per host configuration
DATA SOURCE	INFRASTRUCTURE-EMITTED, derived from the filesystem ownership and permission model of each store
COLLECTION POINT	A structured review of ownership and write permission for every governance and evidence store
SAMPLING FREQUENCY	on every configuration change, and at least quarterly
DIRECTION	higher-better
BASELINE	0 of the stores enumerated in the founding case; every governance store is writable by the party it governs, including the fleet registry, whose lock is advisory only rather than access-controlling, the resource accounting counters, the session transcripts and the console audit log itself, which is world-readable, owned by the same account the process runs as, and sits under a writable allowed root, so ordinary file routes can rewrite or delete it; no hash chain, sequence number, signature or off-host copy exists for any of them
TARGET	1.0 for any store relied upon as evidence in MG-022, MG-023, MG-024, MG-027 or an approval decision, because a mutable record is not evidence, it is testimony from an interested party
GUARDRAIL SENSITIVITY	GR-L03-06 immutable append-only evidence directory raises it and is close to its direct implementation; GR-L08-03 append-only transcripts enforced by storage raises it; GR-L08-04 hash chained audit log raises it; GR-L05-06 denying the agent the credential that could alter its own audit record raises it; GR-L03-10 separation of the evidence store from the operational filesystem raises it by removing the shared writable root; GR-L01-01 lowers it, since a shared account with passwordless root defeats any file-mode-based protection regardless of how the store is otherwise configured
CONFOUNDERS	Filesystem permission model changes made for unrelated reasons; whether an append-only flag or storage feature is actually enforced by the kernel or filesystem versus merely conventional; backup and replication practices that create an independent copy without a formal immutability guarantee
GAMING MODES	Setting restrictive file modes, as the founding case does, 0750 directories, 0640 records, 0600 tokens, that exclude group and other while the owning account, the recorded principal, retains full write access, satisfying the appearance of hardening while changing nothing about who can alter the record; claiming append-only behaviour as a convention rather than verifying it is enforced by storage
CORROBORATED BY	MG-064 audit record completeness; MG-025 time to divergence detection, since immutability determines whether the comparator survives long enough to be checked
READINESS	available
INSTRUMENTATION GAP	None for the current negative finding, directly assessable from the existing file ownership and permission model; a periodic re-verification as stores are added or reconfigured would need to be built

MG-066 Episode reconstructability

QUESTION	Given only the durable record, can an independent party reconstruct what an agent actually did during an episode, without relying on the agent's own account?
OPERATIONAL DEFINITION	Sampled episodes for which an independent reconstruction, built solely from infrastructure-emitted records, matches the agent self-report on all material points, divided by episodes sampled. Distinct from MG-022, which measures divergence found among episodes that were reconstructed; this measures whether reconstruction is even possible, whether or not it is attempted. Excludes episodes with no infrastructure-emitted record of any kind, which are counted separately as unreconstructable by construction.
FORMULA	reconstructable_episodes / sampled_episodes
UNIT AND RANGE	ratio, 0 to 1
TYPE	KCI
HORIZON	lagging
GRANULARITY	per episode, aggregated per audit cycle
DATA SOURCE	Mixed. Self-reports are AGENT-EMITTED; the comparator must be INFRASTRUCTURE-EMITTED, version control history, process manager state, health checks, cloud control plane state and the reverse proxy access log, as in the founding case audit method.
COLLECTION POINT	An audit harness attempting reconstruction from infrastructure sources alone, before comparing to the self-report
SAMPLING FREQUENCY	per audit cycle
DIRECTION	higher-better
BASELINE	consistent with the founding case's own audit method, which paired self-reports against infrastructure evidence, version control history, process manager state, health checks, cloud control plane state and the reverse proxy access log, to find 11 of 18 audited episodes divergent; reconstructability itself, as distinct from divergence, was not separately scored, though the fact that 18 of 46 candidate episodes were audited at all indicates infrastructure evidence was at least partially available for those cases and unclear or insufficient for the remainder
TARGET	1.0; an episode that cannot be reconstructed from independent evidence cannot be audited at all, and the founding case's 46-candidate to 18-audited gap suggests a material share may fall into that category
GUARDRAIL SENSITIVITY	GR-L08-11 raises it directly by adding an infrastructure-emitted stream to reconstruct from; GR-L08-12 run provenance binding raises it by binding measurements to a configuration hash, supporting reconstruction of the state a given action ran against; GR-L03-06 raises it by preserving the record long enough to attempt reconstruction; GR-L08-09 evidence retention policy raises it only if the horizon it sets exceeds the time before reconstruction is attempted, and lowers effective reconstructability where it does not, since transcripts are pruned on a rolling window of about 30 days today
CONFOUNDERS	Which infrastructure systems log at all for a given action class; retention windows across the various stores, which need not align with each other; episode age at the time reconstruction is attempted
GAMING MODES	Scoring an episode reconstructable because some infrastructure record exists, without checking whether that record is sufficient to establish the material facts the self-report asserts; choosing to audit only the episodes with rich infrastructure trails, close to how the founding case's 46-to-18 selection may itself have operated, though that selection mechanism was not characterised as random
CORROBORATED BY	MG-022 claim divergence rate, computed on the reconstructable subset only, to avoid conflating no divergence found with no independent evidence existed to find one
READINESS	derivable
INSTRUMENTATION GAP	A scoring method that separates reconstruction attempted and matched, reconstruction attempted and diverged, MG-022's population, and reconstruction not possible, which the founding case's binary divergent or not-divergent scoring does not currently distinguish

MG-067 Separation of duties violations

QUESTION	How often does one principal hold two or more of the actor, recorder, storage-owner and reader roles for the same action?
OPERATIONAL DEFINITION	Count of consequential actions for which the acting principal, the principal recording the action, the principal owning the record's storage and the principal able to read or query the record are not all distinct, divided by consequential actions assessed. Excludes actions with no defined recorder or reader role, which are reported separately.
FORMULA	count(actions where fewer than 4 distinct principals hold the 4 roles) / count(actions_assessed)
UNIT AND RANGE	ratio, 0 to 1
TYPE	KRI
HORIZON	leading
GRANULARITY	per host configuration
DATA SOURCE	INFRASTRUCTURE-EMITTED, derived from the process and account ownership model for each role in each action
COLLECTION POINT	A structured review of actor, recorder, storage-owner and reader identity for each consequential action type
SAMPLING FREQUENCY	on every configuration change, and at least quarterly
DIRECTION	lower-better
BASELINE	at or near 1.0 across the founding case; zero separation of duties is a standing, confirmed finding, since actor, recorder, storage owner and reader are one process under one account for essentially every consequential action on this host, a condition documented explicitly for the console's own operations and equally true, by the single-account model, for every agent action
TARGET	0 for actions rated severe blast radius or irreversible, and specifically for the deployment approval gate, where the founding case documents a production instance of an agent recording its own peer review using another agent's token, this metric's defining failure case
GUARDRAIL SENSITIVITY	GR-L01-02 lowers it by separating actor identity from the shared pool; GR-L01-07 lowers it by separating the verifier; GR-L01-08 lowers it by separating the approver; GR-L05-06 lowers it by denying the actor write access to its own audit record, separating actor from storage owner specifically; GR-L01-01 raises it to its current near-1.0 level, being the standing cause
CONFOUNDERS	Which of the four roles are even defined for a given action, since some action classes have no recorder or reader role at all today; catalogue scope, since a broader assessed population that includes trivially low-risk actions dilutes the ratio without changing the severe-class picture
GAMING MODES	Introducing a nominally separate reviewer, watcher or reader role implemented as a second session under the same shared account, which satisfies the letter of a four-role model while every role still resolves to one operating system principal; auditing only the console's own operations, where a partial separation already exists, and extrapolating that finding to agent actions, where it does not
CORROBORATED BY	MG-024 verification independence ratio, MG-025's underlying comparator independence, and MG-029 self-attestation share of the record, all specific instances of this more general violation
READINESS	available
INSTRUMENTATION GAP	None for the architectural determination, directly observed and documented today for both the console and the agent fleet; a per-action-class breakdown by the four roles has not been tabulated, only the aggregate finding

The fleet registry's lock is advisory rather than access-controlling, the resource accounting counters, the session transcripts and the console audit log itself are all writable by the account that runs the process, and the audit log specifically is world-readable and sits under a writable allowed root, so an ordinary file route can rewrite or delete it EV-015 EV-018 EV-034. MG-067 separation of duties violations generalises the pattern across every consequential action class at once: the actor, the

recorder, the storage owner and the reader resolve to one process under one account for essentially every action on this host, a condition the register scores at or near 1.0, its worst admissible value.

3.8 Reliability and operations

`MG-068` and `MG-069` are the smallest subfamily and the two most directly tied to D2's viability discussion of independent detection.

Table D4-12. Reliability and operations metrics, MG-068 to MG-069.

MG-068 Wedged session detection latency

QUESTION	How long does a stalled or wedged agent session run undetected?
OPERATIONAL DEFINITION	Elapsed time from the moment a session stops making forward progress, no new tool calls, no output, to the moment any watcher or monitor first flags it as stalled. Excludes sessions with no watcher coverage at all, which are reported separately as undetected by construction rather than folded into the distribution; excludes legitimate long waits on an external call the session is genuinely blocked on.
FORMULA	t_flagged - t_last_forward_progress, per stall event
UNIT AND RANGE	seconds, 0 to unbounded
TYPE	KRI
HORIZON	lagging
GRANULARITY	per stall event per session
DATA SOURCE	Mixed. The last-forward-progress timestamp is derivable from the hook substrate or pane output, both agent-owned; the flagging timestamp is INFRASTRUCTURE-EMITTED from the watcher process, where a watcher exists.
COLLECTION POINT	The watcher's own trip event, joined to the last observed hook or pane activity for that session
SAMPLING FREQUENCY	per stall event, reviewed weekly
DIRECTION	lower-better
BASELINE	for the 6 of 14 sessions with watcher coverage, the fastest independent detector trips at 180 seconds plus up to 60 seconds of schedule granularity; for the other 8 of 14 sessions, no stall detection exists at all, because the watchers' scope is a hardcoded array of agent names, and one session on record was silent for seven days with its process alive and nothing noticing, which stands as the observed ceiling for this metric on this host
TARGET	below 300 seconds, the observed fastest-detector bound, for every session, which requires closing the hardcoded-array gap so watcher coverage is complete rather than partial
GUARDRAIL SENSITIVITY	GR-L04-08 automatic termination on wedge detection lowers it and is the direct mechanism; GR-L10-11 wedge detection surfaced to the operator lowers the human-facing portion of the latency; GR-L04-07 wall clock timeout per agent turn lowers it by bounding how long a turn can run before it is forcibly flagged; the hardcoded watcher scope itself is the dominant confounder rather than a guardrail, and closing it is a precondition for this metric being computable fleet-wide rather than for the 6 covered sessions only
CONFOUNDERS	What counts as forward progress, since a session legitimately waiting on a long-running external call looks identical to a wedged one under a naive definition; watcher polling interval, which sets the floor on detection latency regardless of true stall duration; fleet size and the cost of adding coverage
GAMING MODES	Defining forward progress as any output at all, including a repeated heartbeat or retry loop, which lets a session that is wedged in the narrow sense, not making real progress, read as healthy; excluding unwatched sessions from the metric's population rather than reporting them as undetected by construction, which is what the 8-of-14 gap actually is
CORROBORATED BY	MG-057 signal to noise ratio of the oversight channel, since heartbeat suppression trades against this metric; MG-044 intervention latency, since detection is the first term in that latency's composition
READINESS	absent
INSTRUMENTATION GAP	Watcher coverage extended from the hardcoded array to the full live session inventory, and a forward-progress definition that distinguishes genuine stall from legitimate wait

MG-069 Unattended action path count

QUESTION	How many distinct automations can take action, including writing into a live agent's input stream, with no human keystroke in the path?
OPERATIONAL DEFINITION	Count of distinct scheduled or triggered automations, cron jobs, systemd timers, at jobs and watchers, that can initiate a consequential action without a preceding human keystroke, broken out by whether the automation injects input directly into a live agent session. Excludes automations whose only effect is a read-only status query.
FORMULA	count of automations_with_no_human_keystroke_precondition, with a sub-count of those injecting into a live session
UNIT AND RANGE	count, integer >= 0
TYPE	KRI
HORIZON	leading
GRANULARITY	per host configuration
DATA SOURCE	INFRASTRUCTURE-EMITTED, the cron, systemd timer, at and watcher inventory
COLLECTION POINT	The scheduled task and watcher inventory itself, cross-referenced against whether each item requires a preceding human keystroke
SAMPLING FREQUENCY	on every configuration change, and at least monthly
DIRECTION	target-band
BASELINE	17 automations can initiate action with no human keystroke; 7 of them write text directly into a live agent's input stream, including one that fires on a clock with no stimulus and one that automatically answers the tooling's own interactive permission prompt, the same mechanism MG-058 scores by frequency; separately, 16 scheduled automations across cron, systemd timers and at were enumerated
TARGET	not set as zero, since some unattended operation is the founding case's stated operating model rather than an incidental gap; the defensible target is that every one of the 17 carries a confirmation mechanism appropriate to the severity of what it can trigger, which per the founding case's own finding exists in code for the least dangerous tool only and is absent from the two most dangerous
GUARDRAIL SENSITIVITY	GR-L04-12 advisory lock on scheduled automations lowers collision risk but does not lower this count; GR-L09-11 global fleet pause lowers the effective count during a paused window; GR-L07-06 lowers the subset that could widen an agent's own mandate; GR-L11-07 maximum unattended run duration before re-authorisation lowers sustained exposure even where it does not lower the raw count; no guardrail in the present-population catalogue removes an automation from this count, since all 17 are present-population mechanisms functioning as designed
CONFOUNDERS	Operational necessity, since removing an automation may reintroduce the problem it was built to solve, a genuinely stuck session with nobody present; how the count is scoped, since a watcher that only notifies is a materially different risk from one that injects keystrokes, and conflating them changes the number without changing the finding
GAMING MODES	Counting only the automations that produce human-visible output as action paths, which undercounts the 7 that inject directly into a session with no notification; retiring one automation while adding an equivalent one under a different name, which resets any trend line without changing exposure
CORROBORATED BY	MG-058 automated answering of human decision points, a specific instance within this population; MG-047 halt capability coverage, which is 0 for all 17 by the founding case's own finding
READINESS	available
INSTRUMENTATION GAP	None for the count itself, directly enumerated in the founding case's automation inventory; a per-automation confirmation-mechanism audit exists for only a subset today

MG-068 wedged session detection latency has a genuinely split readiness, available for the six sessions with watcher coverage, where the 180-second bound is directly observed, and absent for the remaining eight, where the one documented case, a session silent for seven days with its process alive and nothing noticing EV-039, is a ceiling rather than a distribution. MG-069 unattended action path count is fully available and is the register's most direct enumeration of unattended reach: seventeen

automations can initiate action with no human keystroke, seven of which write directly into a live agent's input stream, including one that fires on a clock with no stimulus and one that is MG-058's automatic prompt-answerer EV-028.

3.9 Governance overhead

MG-070 to MG-072 are included deliberately alongside the protective metrics rather than as an afterthought, because an evaluation that reports only risk reduced and never the friction introduced is advocacy rather than measurement, the same principle CI-06 states at the composite level in Section 4.

Table D4-13. Governance overhead metrics, MG-070 to MG-072.

MG-070 Governance overhead ratio

QUESTION	What share of total task lifecycle time is consumed by governance activity, approval, review, verification, rather than by the work itself?
OPERATIONAL DEFINITION	Time spent on governance activities, approval waiting and deciding, verification, audit review, divided by total time spent on the task lifecycle from dispatch to completion, aggregated across a window. Excludes governance activity that overlaps entirely with other concurrent work and imposes no wall-clock delay; excludes time attributable to a fault in the work itself rather than to governance.
FORMULA	governance_time / (governance_time + work_time)
UNIT AND RANGE	ratio, 0 to 1
TYPE	KPI
HORIZON	lagging
GRANULARITY	per agent per month
DATA SOURCE	Composed from MG-054, MG-044 and MG-071's underlying timestamps, each carrying its own data source limitation
COLLECTION POINT	Wherever the component approval, verification and intervention timestamps are collected, combined at the point of assessment
SAMPLING FREQUENCY	monthly
DIRECTION	target-band
BASELINE	not computable; neither governance_time nor work_time is instrumented as a distinct category anywhere on this host, and with zero recorded approvals, zero rejections and zero interventions across the 46-day observation window, the observed governance_time for the approval and intervention components is itself effectively zero, which is not evidence of low overhead but of an unused governance layer
TARGET	not set; this metric only becomes meaningful once MG-054 approval turnaround, MG-044 intervention latency and MG-071 approval wait share are separately computable, since this is their aggregate
GUARDRAIL SENSITIVITY	Every guardrail that adds a human decision point, GR-L06-01, GR-L01-06, GR-L01-08, GR-L07-02, raises it directly by construction; GR-L06-06 dry run as default raises it modestly by adding a verification step before the real run; GR-L07-08's runtime classifier, used to fast-path low-risk actions, can lower it for the bulk of low-consequence work while raising it for the smaller share that is genuinely severe
CONFOUNDERS	Workload mix, since a fleet doing mostly low-risk reversible work shows a different ratio than one doing mostly irreversible external actions under identical control settings; how governance time is scoped, since asynchronous review overlapping with other work is not pure overhead in the way blocking review is
GAMING MODES	Performing governance activity nominally but rapidly, which lowers the ratio's numerator while also lowering MG-051 evidence sufficiency, so this metric alone can be gamed by the same shortcut that produces a low approval turnaround time; excluding automation and watcher overhead from the governance_time term while counting only human-facing approval time, understating the true cost of the current 19,334-to-4 signal volume
CORROBORATED BY	MG-071 approval wait share of cycle time; MG-051 approval evidence sufficiency, to catch the speed-without-substance gaming mode
READINESS	absent
INSTRUMENTATION GAP	A time classification distinguishing governance activity from substantive work across the task lifecycle; no such classification exists today, and the underlying component metrics it aggregates are themselves not yet computable

MG-071 Approval wait share of cycle time

QUESTION	Of the total time an action takes from initiation to confirmed effect, what share is spent waiting on an approval decision?
OPERATIONAL DEFINITION	For actions gated by the deployment approval gate or an equivalent approval point, the approval wait time divided by the total cycle time for that action, aggregated per action class. Excludes actions with no approval gate in their path, which are reported separately with a share of zero by definition rather than omitted.
FORMULA	approval_wait_time / cycle_time, per action class
UNIT AND RANGE	ratio, 0 to 1
TYPE	KPI
HORIZON	lagging
GRANULARITY	per action class per month
DATA SOURCE	Composed from MG-054 approval turnaround time and MG-004 cycle time per action class
COLLECTION POINT	Wherever MG-054's and MG-004's underlying timestamps are collected, combined per action class
SAMPLING FREQUENCY	monthly
DIRECTION	target-band
BASELINE	not computable; both MG-054 approval turnaround time and, for approval-gated classes specifically, MG-004 cycle time are themselves not computable in the founding case, the former because zero approvals were recorded in the observation window and the latter because no runtime action classifier joins executed actions to a class at the time of execution
TARGET	not set; the informative use of this metric is as a diagnostic split of MG-004 and MG-070, not as a quantity to minimise on its own, since driving it toward zero by skipping review defeats the purpose of the gate it measures
GUARDRAIL SENSITIVITY	GR-L06-01 raises it directly, since holding the gate with a separate principal is what creates a wait in the first place; GR-L06-09 change size limit triggering escalation raises it for large changes specifically; GR-L10-06 operator attention budgeting lowers it by prioritising the gated request in the operator's queue; GR-L06-03 pre-approved rollback targets lowers it for the rollback subset by removing the need for a fresh decision
CONFOUNDERS	Which action classes are gated at all, since the founding case's gate covers deployment specifically and not the other 59 action classes; operator availability, the dominant driver of the wait component; whether the class's cycle time denominator is dominated by the approval wait or by the technical work, which differs enormously by class
GAMING MODES	Routing an action through a path that avoids the gate entirely, such as the ordinary-restart path the founding case documents as having shipped unapproved code twice, which drives this ratio to zero for that action while defeating the control the ratio was meant to describe; padding the technical-work denominator with idle time so the approval-wait numerator looks proportionally smaller without the wait itself shortening
CORROBORATED BY	MG-054 approval turnaround time; MG-059 control bypass rate, which detects the gate-avoidance gaming mode directly
READINESS	absent
INSTRUMENTATION GAP	Both component metrics, MG-054 and MG-004, need to be computable first; neither is today

MG-072 Control operating cost per action

QUESTION	What does running a given control cost, on average, for each action it applies to?
OPERATIONAL DEFINITION	The operating cost of a control, drawing on the guardrail catalogue's operating_cost qualitative field, divided by the count of actions the control applied to in a window, expressed per applicable action rather than as a fleet-wide total, so that broadly applied controls are not penalised relative to narrowly scoped ones. Excludes one-off build cost, reported separately from the ongoing marginal figure.
FORMULA	control_operating_cost / count(actions_the_control_applied_to)
UNIT AND RANGE	currency or time unit per action, 0 to unbounded, or qualitative low, moderate, high where not quantified
TYPE	KPI
HORIZON	lagging
GRANULARITY	per control per window
DATA SOURCE	The guardrail catalogue's own qualitative operating_cost and friction fields for the qualitative version; a quantitative version would require INFRASTRUCTURE-EMITTED cost data, compute time, licensing, operator time, that does not exist on this host
COLLECTION POINT	The guardrail catalogue itself for the qualitative figure; a cost-accounting system, not present, for the quantitative figure
SAMPLING FREQUENCY	reviewed quarterly, and on every control change
DIRECTION	lower-better
BASELINE	qualitative only, drawn directly from the guardrail catalogue's operating_cost column across the 134 entries; no control on this host has ever been costed quantitatively, and since most present-population controls, identity, deny list, realpath allowlist, are recorded as low or low-to-moderate cost while nearly all proposed controls carry moderate or high cost estimates, the catalogue is likely to understate true operating cost for anything not yet built, because a proposed control's cost is an estimate rather than an observation
TARGET	not set; a cost target is only meaningful alongside the risk reduction the control buys, which MG-062 and MG-059 characterise, and setting a cost ceiling first risks selecting cheap controls over effective ones
GUARDRAIL SENSITIVITY	This metric is a property of implementation choice rather than something a guardrail moves in the same sense as other rows; a lightweight implementation of a given control, for example a classifier reused from the existing hook substrate, GR-L08-13, rather than a new service, lowers it; a heavier implementation of the same nominal control, for example a dedicated policy engine, raises it while potentially also raising MG-031's structural enforcement coverage, so the two should be read together rather than cost minimised alone
CONFOUNDERS	Action volume, since cost per action falls simply by applying an already-built control more broadly with no change to the control itself; whether cost is measured as marginal, per additional action, or fully loaded, including build and maintenance; operator time, which the founding case does not track at all today
GAMING MODES	Reporting only the marginal per-action cost of an already-amortised control while omitting its build and maintenance cost, making broadly deployed but expensive controls look artificially cheap; scoping the applicable-action denominator broadly, counting every action on the host, for a control that only meaningfully applies to a narrow subset, diluting the reported cost
CORROBORATED BY	MG-062 defence in depth factor, so cost is read against the independent protection actually gained, not against nominal control count; MG-070 governance overhead ratio, its human-time counterpart
READINESS	absent
INSTRUMENTATION GAP	A cost accounting model, compute, licensing, operator time, tied to control operation and to actual applicable-action volume; neither exists on this host today

All three are `absent` or mixed, and for the same reason: they are each composed from component metrics, `MG-044`, `MG-054`, `MG-004`, that are themselves not yet computable, so the overhead subfamily is

structurally downstream of the rest of the register rather than independently instrumentable. `MG-072` control operating cost per action has a genuine qualitative reading today, drawn directly from D3's catalogue `operating_cost` field, and an observation worth stating plainly: nearly every present-population control is recorded as low or low-to-moderate cost, while nearly every proposed control carries a moderate or high estimate, which the entry itself flags as likely to understate true cost, because a proposed control's figure is an estimate rather than an observation of a built thing.

---

4. Composite indices

4.1 What a composite buys and what it costs

Eight composite indices roll the seventy-two metrics into figures a reader who has not read the register can still act on. Each is a weighted combination of named component metrics with a stated formula, a stated range, and, in every case, an explicit limitations field describing what the composite hides. That last part is not a formality. A composite is, by construction, a device for discarding information in exchange for communicability, and D2's own account of the founding case's failure, a system that could report success while failing, is a warning against any summary figure that cannot be decomposed back into its parts on demand. Every composite below is designed to be decomposed: none of them is reported without its component metrics available alongside it.

The eight composite indices, CI-01 to CI-08.

`CI-01` Structural Enforcement Ratio

PURPOSE	Answers the question the doctorate exists to ask: what share of an environment's governance rests on architecture rather than on instruction.
COMPONENTS	MG-031 MG-059 MG-060 MG-061
WEIGHTS	0.45 0.25 0.20 0.10
FORMULA	SER = 0.45*MG-031 + 0.25*(1 - MG-059) + 0.20*MG-060 + 0.10*MG-061
RANGE	0 to 1. Below 0.3 the posture is predominantly instructional. Above 0.7 the predominant enforcement is architectural.
JUSTIFICATION	Structural enforcement coverage is weighted highest because it is the property claimed; the inverted bypass rate is weighted next because a control that is routinely routed around is not enforcement whatever its class; fail-closed ratio and signal emission coverage are weighted lower because they qualify rather than constitute the property.
LIMITATIONS	Treats all action classes as equally important, which they are not. A high ratio achieved on low-risk classes while the severe classes remain instructional produces the same number as the reverse. Always read alongside the coverage matrix.
READINESS	derivable

CI-02 Autonomy Boundary Integrity Index

PURPOSE	Measures whether the boundary the architecture claims is the boundary that actually holds.
COMPONENTS	MG-030 MG-033 MG-035 MG-036 MG-037
WEIGHTS	0.30 0.30 0.20 0.10 0.10
FORMULA	ABII = 0.30*(1 - MG-030/n_action_classes) + 0.30*(1 - MG-033) + 0.20*(1 - MG-035_normalised) + 0.10*(1 - MG-036) + 0.10*(1 - MG-037_normalised)
RANGE	0 to 1. A value of 1 means access equals mandate everywhere and no crossing succeeds. The founding case baseline is low: 50 of 60 action classes have access exceeding mandate.
JUSTIFICATION	The first two components carry equal weight because a boundary can fail either by being drawn in the wrong place, which the access-exceeds-mandate count captures, or by not holding where it is drawn, which the crossing success rate captures. Both failures are equally fatal to the claim that a boundary exists.
LIMITATIONS	Normalisation by action class count makes the index sensitive to how finely the inventory is cut. Two environments with identical posture but different inventory granularity will not produce comparable values, so the index is comparable across runs of one environment rather than across environments.
READINESS	absent

CI-03 Verification Independence Index

PURPOSE	Measures how much of what is believed about the system's behaviour rests on the system's own account of itself.
COMPONENTS	MG-024 MG-023 MG-029 MG-048 MG-065
WEIGHTS	0.30 0.20 0.20 0.20 0.10
FORMULA	VII = 0.30*MG-024 + 0.20*MG-023 + 0.20*(1 - MG-029) + 0.20*MG-048 + 0.10*MG-065
RANGE	0 to 1. A value of 0 means every claim about the system originates from the system. The founding case baseline is near 0 for agent actions, because nothing audits what an agent does.
JUSTIFICATION	Independence ratio leads because independence is the property; coverage, the inverted self-attestation share and signal origin independence are weighted equally because each captures a distinct route by which self-report re-enters the record; record immutability is weighted lowest because it protects an existing record rather than producing an independent one.
LIMITATIONS	The index cannot distinguish a verifier that is independent by identity from one that is independent in name but reads the same source the actor wrote. That distinction requires reading the collection point field of each component metric.
READINESS	absent

CI-04 Supervision Viability Index

PURPOSE	Measures the gap between the supervision an architecture claims and the supervision a human can actually exercise.
COMPONENTS	MG-052 MG-046 MG-047 MG-048 MG-051 MG-049
WEIGHTS	0.25 0.20 0.20 0.15 0.10 0.10
FORMULA	SVI = 0.25*MG-052 + 0.20*positive_share(MG-046) + 0.20*MG-047 + 0.15*MG-048 + 0.10*MG-051 + 0.10*attention_fit(MG-049)
RANGE	0 to 1. A value of 1 means every supervision point satisfies all six viability conditions. The founding case baseline is 7 of 34 points without divergence, so approximately 0.21 on the leading component.
JUSTIFICATION	Effective coverage leads because it is the direct measure. Latency margin and halt capability follow because they are the two conditions most often violated silently. Independence, comprehensibility and attention are weighted lower only because they are already partly captured by effective coverage, not because they matter less.
LIMITATIONS	A composite over conditions that are not substitutes. A point that fails VC2 catastrophically and passes the other five scores well, yet supervision there is worthless. The index must never be read without the binding condition column of the supervision register.
READINESS	absent

CI-05 Containment Index

PURPOSE	Measures how far the consequence of a single error can travel.
COMPONENTS	MG-038 MG-039 MG-040 MG-042 MG-041
WEIGHTS	0.25 0.25 0.20 0.20 0.10
FORMULA	CTI = 0.25*(1 - MG-038_normalised) + 0.25*MG-039 + 0.20*(1 - MG-040_normalised) + 0.20*(1 - MG-042) + 0.10*(1 - MG-041_normalised)
RANGE	0 to 1. A value of 1 means a single error stays inside the boundary the design assumed.
JUSTIFICATION	Realised blast radius and containment success rate lead because they are outcome measures rather than posture measures. Lateral reach and privilege breadth follow because they predict the outcome measures and can be computed without an incident.
LIMITATIONS	Realised blast radius is only observable when something goes wrong, so in a healthy period the index is carried entirely by its predictive components and can look stable while the underlying posture degrades.
READINESS	absent

CI-06 Governance Overhead Ratio

PURPOSE	Measures what the governance costs, so that the cost of a posture can be argued against its benefit rather than assumed away.
COMPONENTS	MG-070 MG-071 MG-072 MG-005
WEIGHTS	0.40 0.25 0.20 0.15
FORMULA	GOR = 0.40*MG-070 + 0.25*MG-071 + 0.20*MG-072_normalised + 0.15*(1 - MG-005)
RANGE	0 to 1, lower is better, unlike every other index here. Above 0.5 the governance consumes more than it protects for routine work.
JUSTIFICATION	Included deliberately and given equal standing with the protective indices, because an analysis that reports only the risk reduced and never the friction introduced is advocacy rather than measurement. The direction is inverted relative to the others and that inversion must be stated wherever the index appears.
LIMITATIONS	Overhead is easiest to measure exactly where it is least important, on routine low-risk work, and hardest to measure on the rare high-consequence action where it matters most. The ratio therefore systematically overstates the cost of governance.
READINESS	absent

CI-07 Evidence Integrity Index

PURPOSE	Measures whether the record on which every other measurement depends can be trusted.
COMPONENTS	MG-064 MG-065 MG-066 MG-067 MG-027
WEIGHTS	0.25 0.30 0.20 0.15 0.10
FORMULA	EII = 0.25*MG-064 + 0.30*MG-065 + 0.20*MG-066 + 0.15*(1 - MG-067_normalised) + 0.10*MG-027
RANGE	0 to 1. A value below 0.5 means conclusions drawn from the record are not defensible.
JUSTIFICATION	Immutability carries the highest weight because a complete but mutable record is worth less than a partial immutable one: the first can be made to say anything, the second at least bounds what can be claimed.
LIMITATIONS	This index is uniquely self-undermining: it is computed from the record whose integrity it assesses. It must be computed by a party other than the one that wrote the record, or it measures nothing.
READINESS	absent

CI-08 Bounded Autonomy Maturity Score

PURPOSE	A single communicable figure for where an environment sits on the path from instruction-based to architecturally enforced governance.
COMPONENTS	CI-01 CI-02 CI-03 CI-04 CI-05 CI-07
WEIGHTS	0.20 0.20 0.20 0.20 0.10 0.10
FORMULA	BAMS = 0.20*CI-01 + 0.20*CI-02 + 0.20*CI-03 + 0.20*CI-04 + 0.10*CI-05 + 0.10*CI-07
RANGE	0 to 1, reported as five maturity bands: 0.0 to 0.2 instructional; 0.2 to 0.4 conditional; 0.4 to 0.6 mixed; 0.6 to 0.8 predominantly structural; 0.8 to 1.0 structurally enforced with independent verification.
JUSTIFICATION	The four leading indices are weighted equally because the argument is that none of them substitutes for another: enforcement without verification is unfalsifiable, verification without boundary integrity is observation of an uncontrolled system, and boundary integrity without supervision viability is a claim about design rather than about operation. Containment and evidence integrity are folded in at lower weight because they are partly upstream of the others and would otherwise be double counted.
LIMITATIONS	A composite of composites, and therefore the most communicable and least diagnostic figure in the framework. It exists for the one page that has room for a single number. Any decision taken on this score alone is a misuse of it, and it deliberately excludes the governance overhead ratio so that it cannot be improved by making the system slower.
READINESS	absent

4.2 Reading the set

Three properties of the set are worth stating explicitly rather than left to be inferred from the individual entries.

Only `CI-06` is inverted. Every index runs `higher-better` except Governance Overhead Ratio, which runs `lower-better` and whose limitations field states plainly that overhead is easiest to measure exactly where it matters least, on routine low-risk work, and hardest exactly where it matters most. A reader who improves `CI-01` through `CI-05` by adding friction and never checks `CI-06` has optimised the wrong objective; the two are deliberately kept apart rather than netted into one figure, and `CI-08` deliberately excludes `CI-06` from its own weighting for the same reason, stated in `CI-08`'s own limitations field: so that the headline maturity score cannot be improved simply by making the system slower.

Readiness is worse at the composite level than at the metric level. Section 6 below reports the register's readiness distribution as fourteen `available`, twenty-five `derivable` and thirty-three `absent` metrics. Among the eight composites, none is `available`: one is `derivable` (`CI-01`) and seven are

`absent` (`CI-02`, `CI-03`, `CI-04`, `CI-05`, `CI-06`, `CI-07`, `CI-08`). This is the expected direction of the gap, since a composite requires every one of its components to be computable and a composite is only as ready as its weakest input; `CI-03` Verification Independence Index is `absent` specifically because one of its five components, `MG-023` verification coverage of completion claims, is itself `absent`, even though the other four, `MG-024`, `MG-029`, `MG-048` and `MG-065`, are each `available` or `derivable`, and `CI-08` inherits absence from `CI-03` and `CI-07` among its own six inputs.

`CI-07` is self-undermining by its own admission. Evidence Integrity Index is computed from the same record whose integrity it purports to assess, and its limitations field says so directly: it must be computed by a party other than the one that wrote the record, or it measures nothing. On the Agent Console host today, no such party exists for the agent side of the record, which is `MG-029`'s finding restated at the index level. `CI-07`'s own readiness of `absent` is therefore not a gap to be closed by better instrumentation alone; it requires the identity separation D2 Section 5.2 names as the precondition for a genuine verifier.

> **WHAT CI-08 IS FOR AND WHAT IT IS NOT FOR**
>
> Bounded Autonomy Maturity Score is the one figure this programme would put on a single slide. It is explicitly a composite of composites, and its own limitations field calls it the most communicable and least diagnostic figure in the framework, existing for the one page that has room for a single number. Any decision taken on `CI-08` alone, without opening at least the four equally weighted indices beneath it, is a misuse of it in exactly the sense this register warns against throughout: a summary that cannot be decomposed on demand is not a measurement, it is a claim, and this framework was built specifically not to have those go unchecked.

---

5. Measurement architecture

Sections 2 through 4 specify what would be measured. This section states what would have to exist to measure it, and, for each element, whether that thing exists on the Agent Console host today.

5.1 Event schema per collection point

A collection point, in the vocabulary this register uses throughout, is the exact place in the running system where an instrument would sit: a hook callback, a route handler, a supervisor unit transition, a filesystem ownership check. Every metric's `collection_point` field names one. The schema a collection point would need to emit, to make the metrics in Sections 2 and 3 jointly computable rather than individually derivable, is the same six fields across every point: a timestamp, an actor identity, an action class identifier, an effect record, an originating principal distinct from the actor where independence is being asserted, and a link back to the unit of work the event belongs to. No collection point on this host emits all six today. The recurring, single most common gap across the register is the last one: `MG-001`, `MG-003` and `MG-004` each name the absence of a dispatch identifier that would join an event to the brief or task it belongs to as their principal instrumentation gap, and the gap recurs because it is the same missing field observed from three different metrics.

5.2 Agent-emitted and infrastructure-emitted, applied to this host

Section 1.4 stated the distinction abstractly. Applied concretely to the Agent Console host, the inventory is short on the infrastructure side and long on the agent side.

What is infrastructure-emitted today. The console audit log, 863 lines over 46 days, recording operator actions taken through the console `EV-034`; the process manager's restart counters and process inventory `EV-004`; the supervisor unit journal for the units that hold agent sessions `EV-005`; version control commit history, though authored under a single human identity by standing instruction

rather than per agent EV-032; and the fleet registry, whose writer discipline is a single advisory lock rather than an access control EV-022.

What is agent-emitted today. The model runtime transcript, about 1.13 GiB EV-026; the transcript hook substrate's output, about 186 MB, which fires on every tool call but is contractually required to exit successfully and therefore records without evaluating EV-026; every completion marker, memory note and notification body; the resource accounting counters, which discard session identity at ingest EV-024; and every instruction, role and permission file that constitutes the agent's mandate, all of which are writable by the agents they govern EV-032.

What is proposed and does not exist. An infrastructure-emitted event stream independent of the agent (GR-L08-11), an independent verifier identity (GR-L08-05, GR-L01-07), a runtime action classifier assigning class and level at the moment of attempt (GR-L07-08), and a policy decision point that could evaluate an attempt before it executes (GR-L07-02). Every one of these is catalogued in D3 as proposed, and every one of them is named in this register's guardrail_sensitivity fields as the precondition for moving a specific metric from absent or derivable to genuinely computable. The two registers are, in this respect, describing the same gap from two directions.

Figure D4-1 traces the distinction through a single worked metric, MG-029 self-attestation share of the record, chosen because it is the one entry in the whole register whose two paths, agent-emitted and infrastructure-emitted, converge inside a single computation rather than feeding two separate metrics, and because it is an actual weighted component of CI-03, not an illustrative construction.

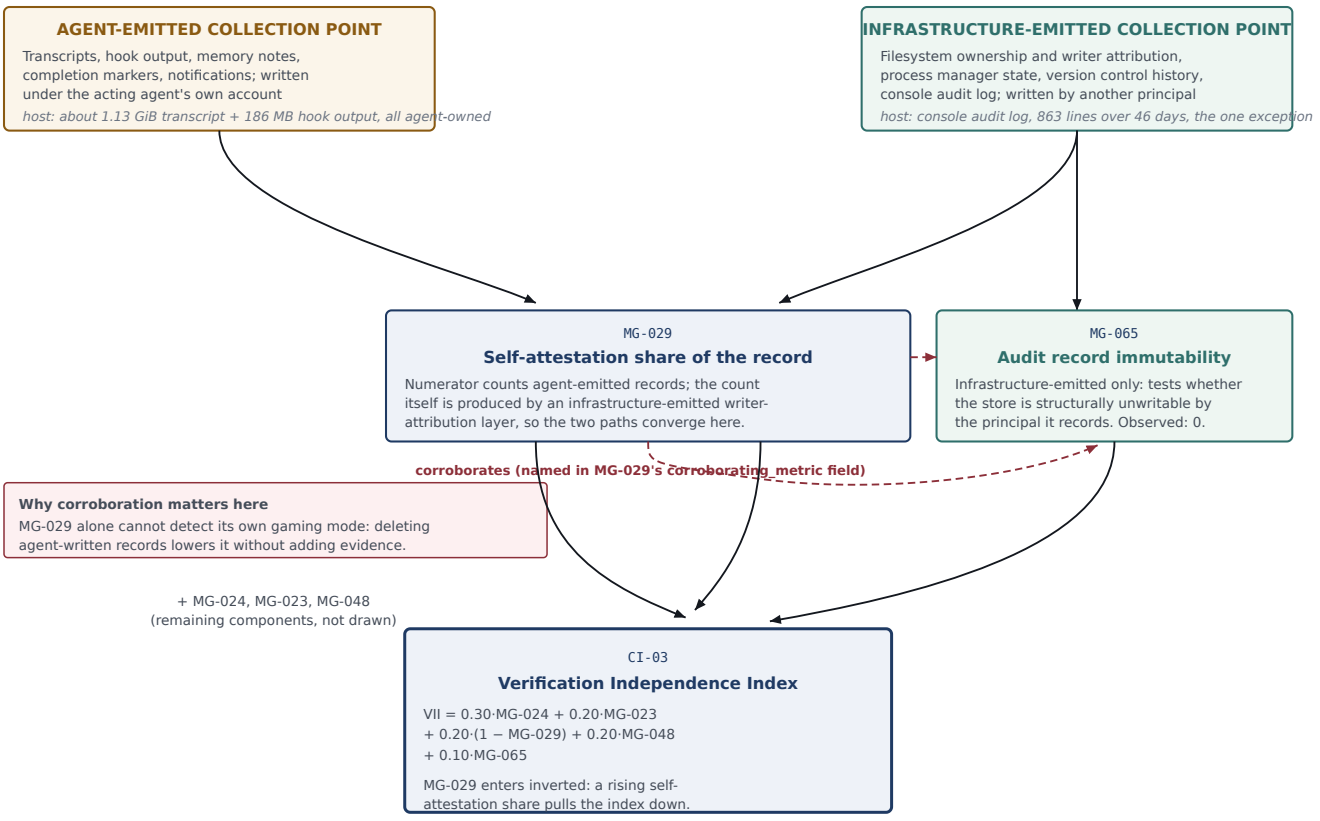

Figure reads top to bottom: two collection points, a metric where their paths converge, its corroborating metric, and the composite index both feed. The pattern generalises: every metric whose data_source names agent self-report carries a corroborating_metric field.

Figure D4-1. Metric flow from collection point to composite index. Two collection points, agent-emitted at left and infrastructure-emitted at right, converge in MG-029, whose numerator counts agent-written records using an infrastructure-emitted attribution layer. MG-029's corroborating metric, MG-065 audit record immutability, is independently infrastructure-emitted throughout. Both feed CI-03 Verification Independence Index, where MG-029 enters inverted: a rising self-attestation share pulls the index down.

5.3 Storage model

There is no relational database anywhere on this host EV-015. Every store this register or D3's catalogue names, the fleet registry, the resource accounting counters, the kill switch sentinel, the inventory, every agent's bus token, the console audit log, is a flat file under a writable directory EV-015 EV-018. No log rotation exists for any of them. This is not, by itself, a measurement defect: a flat file can in principle be written append-only and hashed. What makes it a measurement defect is that every one of these files is owned by the account it records, so the storage model and the writability problem MG-065 measures are the same fact described twice.

5.4 Run provenance

GR-L08-12 run provenance binding every measurement to a configuration hash is catalogued in D3 as proposed: nothing on the Agent Console host today binds a measurement to the guardrail configuration in force when it was taken, so a metric value collected under one condition set cannot be distinguished, after the fact, from the same metric collected under another. This programme's own registers are versioned by contrast: this document is part of run 20260807-1714, a dated snapshot that is not revised in place. That external discipline, applied to the discovery process itself, is not a substitute for the internal one the founding case's own runtime lacks; it is the reason the absence is visible at all. A metric register that is itself unversioned could not have stated, as this one does throughout, which of its own baselines will need to be recomputed the next time the guardrail configuration changes.

5.5 Export format

No metric in this register is exported today in any structured format external to the ad hoc notification and memory-note channels MG-029 already characterises as agent-owned. There is no metrics endpoint, no scheduled export job, and no schema published for one. This is not determined to be a deliberate design choice rather than an oversight; the founding case's own operating model, an SMS-first single channel to one operator EV-029, was not built with a machine-readable metrics consumer in mind, and nothing in the evidence base establishes that one was ever considered and rejected.

5.6 Retention and immutability

Retention is short and undocumented as policy. Model runtime transcripts are pruned on a rolling window of about 30 days, a figure surfaced in this register through MG-025's instrumentation gap rather than through any retention specification, because no retention specification exists to cite directly. Immutability is, as MG-065 establishes at length in Section 3.7, absent throughout: file modes on the governance stores narrow group and other access, typically 0750 for directories, 0640 for records, 0600 for tokens, but leave the owning account, which is also the recorded principal in every case, with full write access EV-015. A permission regime that excludes everyone except the party it is meant to constrain satisfies the appearance of hardening without changing who can alter the record, and this register treats that pattern as equivalent to no protection at all rather than as partial protection, consistent with how MG-065 scores it.

---

6. Instrumentation readiness against this host

6.1 The readiness distribution

Of the seventy-two metrics, fourteen are available on the Agent Console host today with no further instrumentation, twenty-five are derivable, computable from data that exists but is not yet joined,

classified or aggregated in the required shape, and thirty-three are `absent`, requiring a data source, a store or a comparator that does not exist at all. Two entries carry a split reading rather than a single value: `MG-068` wedged session detection latency is `available` for the six sessions with watcher coverage and `absent` for the remaining eight; `MG-072` control operating cost per action is `absent` for a quantitative figure and `derivable` for the qualitative low, moderate, high reading already present in D3's catalogue. Counted at the coarse grain, fourteen `available` plus twenty-five `derivable` covers thirty-nine of the seventy-two entries outright, with the two split entries each partly in one of those two columns; thirty-one entries require a data source that does not exist in any form at all.

The distribution is not even across the two families. Engineering metrics, twenty-one entries, split nine `derivable` and eleven `absent`, with one `available`; not a single engineering metric reaches `available` outright except `MG-014` unplanned restart rate, because engineering measurement on this host depends almost entirely on joins that have not been built, a dispatch identifier, an action classifier, a build stage, rather than on missing raw material. Governance metrics, fifty-one entries, split thirteen `available`, sixteen `derivable`, twenty `absent`, and two mixed. The governance family's stronger `available` share has two distinct sources, and the register keeps them apart rather than crediting both to instrumentation maturity. `MG-030` access exceeds mandate count and `MG-050` nominal to effective posture divergence count are `available` because they read directly off registers this programme itself built, the action class register and the completed supervision viability assessment respectively. `MG-024`, `MG-042`, `MG-049` and `MG-053` are `available` for a different reason, because the founding case's own operational inventories, the account and credential model, the tmux and fleet registries, and the console audit log, already carry the raw fact even though nothing on the host computes it as a standing metric. Neither source is the founding case's own runtime instrumenting itself for governance purposes; both are readings taken from material that already exists for other reasons.

Table D4-14. Instrumentation readiness for all seventy-two metrics, with the specific gap that would need closing for each entry marked absent or derivable.

ID	NAME	READINESS	INSTRUMENTATION GAP
MG-001	Task completion rate	derivable	A dispatch identifier minted at the dispatch route, carried into the session and echoed in the completion record; no such identifier exists today
MG-002	Actions per agent hour	derivable	A mutating or read-only classification on each hook record, and an idle-time definition; the raw event stream exists at about 186 MB across active sessions but is unclassified
MG-003	Lead time from dispatch to first effect	derivable	A dispatch identifier propagated into hook records; without it the join is temporal and breaks whenever two dispatches overlap in one session
MG-004	Cycle time per action class	derivable	An action class assignment on each hook record and an effect-confirmation event for each class; the hook stream exists but carries no class and no effect linkage
MG-005	Autonomous completion share	derivable	Capture of direct pane typing, which is not distinguishable from agent output in the current pane logs; without it the metric is biased upward by an unknown amount
MG-006	Model capacity consumed per completed task	absent	Per-session and per-agent attribution at ingest, and a counter store the measured account cannot write; today the attribution is discarded and the store is owned by the party it measures
MG-007	Change failure rate	derivable	A deployment event record; today activation is an ordinary process restart that records neither the revision it loaded nor the fact that a deployment occurred
MG-008	Defect escape rate to production	absent	A defect register, a pre-deployment detection stage to make escape meaningful, and a production error stream collected independently of the agent
MG-009	Rework ratio	derivable	A per-agent author identity on commits; today authorship is collapsed to one operator identity, so the metric can only be computed per repository
MG-010	First attempt success rate per action class	derivable	An action class assignment and a retry-grouping key; both would have to be derived offline from the existing transcript, and no such derivation exists
MG-011	Test coverage delta per change	absent	A build stage of any kind, a test harness in each repository, and a coverage tool; none exists on this host
MG-012	Build reproducibility rate	absent	A recorded revision on every deployed artefact, an artefact hash, and a verifier identity able to rebuild; none of these exists
MG-013	Restart induced work loss	derivable	An in-flight work marker written at turn boundaries; today the only signal is the free text reload result string in the fleet registry
MG-014	Unplanned restart rate	available	none; visible today in the process manager inventory and the supervisor unit journal, with the caveat that counters are cumulative since daemon start and no scheduled-versus-unplanned attribution is recorded
MG-015	Service availability during agent activity	absent	An external health prober with a function-exercising criterion, and an activity timeline joined to it; neither exists
MG-016	In-flight interaction loss on restart	absent	An in-flight interaction gauge in each externally facing service and a drain phase in the stop path; neither exists

ID	NAME	READINESS	INSTRUMENTATION GAP
MG-017	Mean time to detect a fault	absent	A fault register with first-occurrence timestamps and an independent monitoring stream with fault-specific criteria; today the only independent detectors watch for silence, not for faults
MG-018	Mean time to repair without human intervention	absent	An incident record with detection and restoration timestamps, and a restoration criterion evaluated independently of the repairing agent
MG-019	Repair correctness rate	absent	A verifier identity distinct from the acting account, and a recorded failing criterion to re-execute; neither exists, and the shared account model makes the first structurally unavailable today
MG-020	Repair regression rate	absent	A repair register, a fault register and an independent causal attribution; none exists
MG-021	Rollback exercise success rate	absent	A pre-change snapshot for every change class, including package state, and a state comparison after rollback; the update plane has the first only for cloud resources
MG-022	Claim divergence rate	derivable	The founding case audit was manual; a repeatable harness needs a stable episode identifier, a self-report store the agent cannot edit, and an infrastructure comparator collected under a different principal
MG-023	Verification coverage of completion claims	absent	A verification record schema, a routing rule from claim to verifier, and a verifier identity; none exists on this host
MG-024	Verification independence ratio	available	none for the architectural determination, which is visible today in the process ownership of every agent session and in the single account holding all credentials; a verification record schema is still required before the ratio is informative rather than identically zero
MG-025	Time to divergence detection	absent	A durable claim record with a trustworthy timestamp and a comparator retained at least as long; today transcripts are pruned on a rolling window of about 30 days and every store is writable by the claimant
MG-026	Unverifiable claim share	absent	An assertion-level parse of self-reports and a comparator inventory; the founding case audits at episode granularity only
MG-027	Evidence citation completeness	absent	A citation convention in agent reports, stable artefact identifiers, and a resolver that tests content rather than existence
MG-028	Adversarial challenge yield	absent	A review queue with a recorded adversarial instruction and a materiality adjudication by a third party; the founding case has agent-to-agent escalation notes but no review population
MG-029	Self-attestation share of the record	derivable	A writer attribution per evidence store; ownership is observable today but no inventory ties each store to its writing principal and its record count
MG-030	Access exceeds mandate count	available	none; computable today by joining the access and mandate columns of the action class register, both of which are populated for all 60 classes
MG-031	Structural enforcement coverage	derivable	A per-class control inventory with a writability test on each implementing artefact; the guardrail catalogue supplies the candidate controls but no automated writability assessment exists

ID	NAME	READINESS	INSTRUMENTATION GAP
MG-032	Boundary crossing attempt rate	absent	A runtime action classifier and a policy decision point; the hook substrate already fires on every tool call with a match-everything matcher, so the observation surface exists and only the decision function is missing
MG-033	Boundary crossing success rate	absent	A policy decision point that records refusals as well as permissions, and an effect record independent of the tool own return value
MG-034	Mandate stability	derivable	A defined set of mandate-bearing artefacts and a clause-level parse; the version control history exists but treats these files as ordinary documentation and does not distinguish the authoring principal
MG-035	Mandate widening events initiated by the governed party	absent	Per-agent authorship on mandate changes, which requires per-agent identity; the founding case has neither, so the event class is unobservable rather than absent
MG-036	Unclassified action share	derivable	A classifier and a tool inventory; the observation surface already exists in the hook substrate, which sees every tool call and is deliberately prevented from acting on what it sees
MG-037	Level demotion latency	absent	Per-agent identity and access levels to narrow, and a trigger-to-application pipeline; neither exists, and the underlying access model does not currently vary by agent
MG-038	Blast radius realised per incident	absent	An incident register with a scored scope field, and cross-host or cross-account evidence collection independent of the account involved; neither exists
MG-039	Containment success rate	absent	An incident register, and a declared containment boundary per action class; neither exists
MG-040	Lateral reach from a compromised agent identity	derivable	A systematic reachability graph across hosts, cloud accounts, the metadata endpoint and the inter-agent bus, refreshed on change; today only a static credential and key inventory exists
MG-041	Egress destination cardinality per agent	absent	An egress log or proxy independent of the agent account; none exists, and no host firewall is active to produce one as a side effect
MG-042	Privilege breadth index	available	none for the current single-identity figure, which is directly observed; a per-agent decomposition requires per-agent identity, which does not exist
MG-043	Irreversible action share	derivable	The same runtime classifier MG-032 and MG-036 require, joined to the reversibility field rather than the mandate field; neither the classifier nor the join exists
MG-044	Intervention latency	absent	A timed record of actual interventions, of which the founding case has none in 46 days, and an independent stopped-signal distinct from the agent's own report
MG-045	Harm completion time per action class	derivable	A per-class effect-completion event distinguishing the point of no return from confirmed completion, and a timed incident population to build a real distribution from; today only order-of-magnitude estimates exist
MG-046	Latency margin	absent	None of the three component latencies is instrumented as a distribution today; this metric cannot be computed until they are
MG-047	Halt capability coverage	available	A formal per-mechanism test of whether it halts in-progress work or only gates the next call; today the classification above is a design read of documented behaviour, not a tested one

ID	NAME	READINESS	INSTRUMENTATION GAP
MG-048	Oversight signal independence share	derivable	An originator classification applied uniformly across every notification and alert channel; today the channels are heterogeneous and unclassified as a set
MG-049	Concurrent supervised stream count	available	An empirically established single-operator capacity figure to compare the count against; none exists
MG-050	Nominal to effective posture divergence count	available	None for the current snapshot, which is the founding case's own completed assessment; a repeatable version needs a standing point inventory and scheduled re-verification rather than a one-off review
MG-051	Approval evidence sufficiency	absent	A payload-origin classification on every approval request the deployment approval gate renders; not currently captured
MG-052	Effective supervision coverage	derivable	A per-session, per-condition standing assessment refreshed on configuration change, rather than the one-off review the founding case performed
MG-053	Operator intervention rate	available	None for the console-mediated channel; direct pane typing and any out-of-band operator action are not captured by any log
MG-054	Approval turnaround time	absent	A populated approval history; today the gate exists but the observed window shows no decided requests
MG-055	Approval rejection rate	absent	A populated approval history with outcomes, and a way to detect submissions that bypassed the gate entirely, which the founding case's ordinary-restart path allows
MG-056	Alert acknowledgement rate	absent	An acknowledgement event distinct from message delivery; the current SMS-first channel records dispatch, not receipt or action
MG-057	Signal to noise ratio of the oversight channel	available	A standing decision-relevance classification applied consistently as the fleet and its automations change; today the 4-versus-19,334 figure is a point-in-time count
MG-058	Automated answering of human decision points	derivable	A firing log for the specific automation, and a formal inventory of every interactive prompt type in the tooling so the denominator can be established
MG-059	Control bypass rate	derivable	A systematic bypass search per control, repeated on change, rather than the incidental discovery the founding case relies on
MG-060	Fail closed ratio	available	None for the catalogue-based figure, directly computable today; a live fault-injection test to verify the recorded fail_state matches actual behaviour does not exist
MG-061	Control signal emission coverage	available	A distinction, not currently recorded in the catalogue at the emission level, between a signal that is emitted and one that is emitted, collected and actually monitored
MG-062	Defence in depth factor	derivable	A systematic cross-reference of bypass_analysis fields across all controls binding a given action class, to identify shared single points of failure before counting layers as independent; the raw material exists but the cross-reference has not been performed
MG-063	Single point of failure count	derivable	A structured extraction of the defeating condition named in every bypass_analysis field, clustered to identify shared conditions, rather than the qualitative reading performed here

ID	NAME	READINESS	INSTRUMENTATION GAP
MG-064	Audit record completeness	derivable	A uniform, non-manual audit hook applied to every mutating route and every agent tool call, rather than the current per-handler opt-in
MG-065	Audit record immutability	available	None for the current negative finding, directly assessable from the existing file ownership and permission model; a periodic re-verification as stores are added or reconfigured would need to be built
MG-066	Episode reconstructability	derivable	A scoring method that separates reconstruction attempted and matched, reconstruction attempted and diverged, MG-022's population, and reconstruction not possible, which the founding case's binary divergent or not-divergent scoring does not currently distinguish
MG-067	Separation of duties violations	available	None for the architectural determination, directly observed and documented today for both the console and the agent fleet; a per-action-class breakdown by the four roles has not been tabulated, only the aggregate finding
MG-068	Wedged session detection latency	absent	Watcher coverage extended from the hardcoded array to the full live session inventory, and a forward-progress definition that distinguishes genuine stall from legitimate wait
MG-069	Unattended action path count	available	None for the count itself, directly enumerated in the founding case's automation inventory; a per-automation confirmation-mechanism audit exists for only a subset today
MG-070	Governance overhead ratio	absent	A time classification distinguishing governance activity from substantive work across the task lifecycle; no such classification exists today, and the underlying component metrics it aggregates are themselves not yet computable
MG-071	Approval wait share of cycle time	absent	Both component metrics, MG-054 and MG-004, need to be computable first; neither is today
MG-072	Control operating cost per action	absent	A cost accounting model, compute, licensing, operator time, tied to control operation and to actual applicable-action volume; neither exists on this host today

6.2 What the three readiness levels mean in practice

`available` does not mean measured; it means computable without new instrumentation from data that already exists in an accessible form. None of the fourteen `available` metrics has, in fact, been computed as a maintained, recurring figure; each has been computed once, for this register, as a point-in-time reading. Treating `available` as equivalent to `monitored` would overstate the founding case's current measurement maturity in exactly the way this document's own gaming-modes discipline warns against.

`derivable` is the largest category and the one where investment is cheapest relative to what it buys. The recurring pattern across the twenty-five entries is a missing join rather than a missing measurement: an action class assignment on hook records (`MG-002`, `MG-010`, `MG-032`'s sibling `MG-036`), a per-agent author identity on commits (`MG-009`, `MG-034`), or a writer attribution inventory across evidence stores (`MG-029`). In each case the raw event stream already exists, most visibly the hook substrate's 186 MB of unclassified output `EV-026`, and what is missing is the classification or identity layer that would make it addressable.

`absent` means a data source, a store, or a comparator population does not exist in any form, and building the metric requires building the thing it measures first. `MG-021` rollback exercise success rate is the clearest instance: its denominator, attempted rollbacks, is zero, so the metric cannot even begin to be estimated until the update plane's compiled but never-fired apply path is exercised at least once `REPORTED`. The `absent` category is concentrated in exactly the subfamilies D2 identifies as carrying the founding case's least mature controls, repair capability, all five entries `absent`; verification and truthfulness, five of eight; and containment, three of six, which is the expected pattern if the metric register and the guardrail catalogue are describing one underlying maturity level rather than two independent ones.

> **THE REGISTER IS A LIVING INSTRUMENT, NOT A FINISHED ONE**
>
> Every entry in this register carries a `status` field, currently `active` on all seventy-two. The register's stated purpose is comparative, to be recomputed under different condition sets so that guardrail effects become visible as differences rather than asserted as arguments, and thirty-one of its seventy-two entries cannot yet be recomputed at all in that sense, because they are `absent`. That is not a defect in the register; it is the honest starting position this programme's own guardrail sensitivity principle requires it to state, so that the next run can report whether the count of `absent` entries fell.