Executive Summary

This executive summary distills the program's theory-driven analysis into decision-relevant artifacts: five **claims** derived from classical cybernetics (requisite variety, Good Regulator, VSM recursion), five **risks** expressed as triggerable conditions with first-line mitigations, and five **metrics** that preserve construct validity and enable severe tests. Each claim is framed with explicit scope conditions and anticipated effect directions; each risk specifies early-warning signals and an escalation path aligned to VSM levels; each metric is tail-sensitive or rate-normalized and tied to simulator "knobs" in Q2 to strengthen internal validity. Together they provide a compact contract between theory and operation: what to expect, what to watch, and what to measure when deploying agent-mediated control under enforceable constraints. The aim is not rhetorical priority-setting but pre-registered governance, i.e. claims that can be corroborated or revised via measurement, with thresholds and responsibilities fixed in advance to support replication, audit, and rapid recalibration as environments shift.

Below are the five most decision-relevant **claims**, the corresponding **risks** to watch, and a compact set of **metrics** to wire into dashboards and experiments.

Five Claims (decision statements with scope conditions)

1. **Requisite Variety Governs Stability**
   When the *agent + escalation chain*'s effective capability variety meets or exceeds environmental variety, SLA breaches and emergency overrides decline; under-variety reliably leaks as exceptions. (Scope: heterogeneous, service-like workloads.)
2. **Intake Has an Interior Optimum**
   Attenuation at intake (schemas, templates, pre-classification) exhibits

a U-shaped total cost: over-filtering misses rare hazards; under-filtering overloads regulators. Optimal thresholds minimize *(misses + rework + delay)* subject to risk bounds.

3. **S2–S3 Separation is Necessary**
   System-2 coordination (synchronizing peers) and System-3 assurance (allocation/rollback/compliance) are complementary; fusing them increases conflict, rollbacks, or defects. Keep charters, dashboards, and authorities distinct.
4. **Constraint-First Control Reduces Tail Loss**
   Encoding legal/ethical limits as hard guards and setpoint bands (with algedonic hard-stops) lowers severe-incident rates at an acceptable throughput penalty up to a pre-registered threshold.
5. **Observability + Provenance Shorten MTTR**
   Rich, timely telemetry plus decision trails materially reduce mean time-to-resolution and repeat incidents; without provenance, RCAs and assurance claims are not auditable.

Five Risks (with trigger signals and first-line mitigations)

1. **Under-Variety Fragility**
   Signal: Variety coverage ratio < 1.0 on critical subtypes for ≥2 intervals.
   Mitigation: Add amplification levers (tools/permissions/data) or tighten attenuation; revise escalation map.
2. **Over-Attenuation Blind Spots**
   Signal: Rising false negatives on high-risk classes; “unknown/other” bucket growth.
   Mitigation: Lower intake thresholds; introduce human catch-all; add post-intake anomaly screening.
3. **Role Conflation (S2↔S3 or S4↔S5)**
   Signal: Conflict incidents + policy churn; experiments bypassing policy or policy freezing exploration.

Mitigation: Reassert charters; split dashboards and approvals; time-box joint reviews.

4. **Constraint Erosion / Proxy Gaming**
   Signal: Severe incidents despite KPI gains; proxy–outcome gap widening.
   Mitigation: Harden constraint checks pre/post-act; revise proxies; activate algedonic drills.
5. **Observability Gaps**
   Signal: Provenance coverage < target; MTTR flat despite interventions; unverifiable RCAs.
   Mitigation: Enforce logging SLAs; freeze risky changes; add replayability tests and coverage monitors.

Five Metrics (definitions, rationale, suggested thresholds)

1. **Variety Coverage Ratio**
   Def: (Regulator capability variety) ÷ (demand variety/entropy) at the subtype level.
   Why: Direct operationalization of Ashby; predicts exception leakage.
   Threshold: ≥ **1.0** for critical subtypes; alert if < **1.0** for two consecutive intervals.
2. **Severe-Incident Rate**
   Def: Incidents at or above predefined severity per 10k actions (tail loss proxy).
   Why: Primary outcome for constraint-first control and algedonic efficacy.
   Threshold: ≤ **α**/10k; algedonic trigger at **α***.
3. **Provenance Coverage & Freshness**
   Def: % of actions with complete decision trail; max event lag to log ingestion.
   Why: Preconditions for auditability, RCA, and model tuning.
   Thresholds: Coverage ≥ **95%** daily; freshness ≤ **β** seconds.

4. **Oscillation Index**
   Def: Normalized count of overshoot/undershoot events plus settling time for key KPIs.
   Why: Sensitive to gain tuning and delay compensation.
   Thresholds: Overshoot count ≤ **γ** per day; settling time ≤ **δ** minutes.
5. **Proxy-Gap Index**
   Def: |proxy − outcome| / outcome over a fixed window (by segment/subgroup).
   Why: Early detector of reward hacking or goal drift.
   Threshold: ≤ **ζ** for **k** consecutive windows; trigger S5 review if exceeded.

> **Implementation note:** Pre-register α, β, γ, δ, ζ to your context; link each metric to Q2 “knobs” (e.g., demand entropy, attenuation strength, controller gains, log latency) and effect-direction thresholds so qualitative claims are subjected to severe tests before policy hardening.