



THE TRUE CEILING EDITION

GoalOS α -AGI Ascension

SOVEREIGN RESERVE Ω

INSTRUMENT 05

RISK & SAFETY FRAMEWORK

Protective controls for open-ended intelligence and real-world deployment



Risk classification, prohibited domains, release gates, incident response,
independence, concentration, canaries, and rollback.

From Open-Ended Intelligence to Kardashev Type I Civilization



PROOF-GATED AUTONOMY • VERIFIED CAPABILITY • COMPOUNDING CAPACITY

FOUNDING RELEASE Ω -2.0 • JULY 2026

CONSTITUTIONAL POSITION

Mandate, Authority, and Boundary

Protective controls for open-ended intelligence and real-world deployment

The institution must be able to stop, narrow, and recover faster than it can scale.

Purpose

Govern risk across research, models, agents, data, tools, markets, finance, rights, security, people, and physical-world actuation.

Authority

Sets risk classes, prohibited activities, evaluator routes, canary ceilings, monitoring thresholds, incident command, challenge windows, and reauthorization gates.

Boundary

No score, sponsor, market signal, innovation claim, or token commitment may override a failed safety, rights, or rollback gate.

High consequence raises the required proof level; high novelty raises the burden of reproduction, stress testing, and delayed outcome.

OPERATING CONSTITUTION

Six Non-Negotiable Laws

The framework remains legible under speed, scale, novelty, and pressure.

Classify before execution

Every mission receives a risk class and authority ceiling before work begins.

Fail closed

Missing proof, invalid identity, stale state, key compromise, or hidden irreversibility stops propagation.

Canary first

Production authority expands gradually, with explicit stop thresholds and recovery tests.

Preserve incidents

Logs, inputs, outputs, roots, calls, keys, reviewers, and communications are retained for investigation.

Measure concentration

Operator, model, cloud, jurisdiction, key, data, and validator dependence are treated as risk terms.

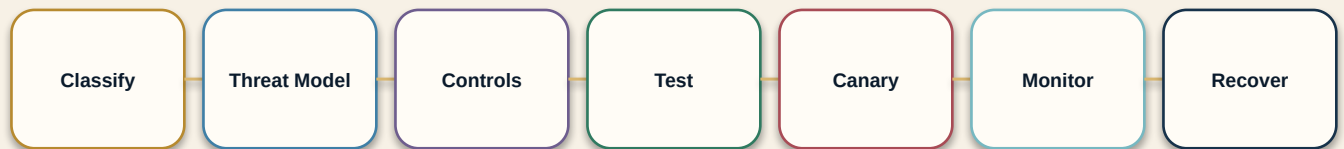
Reauthorize after repair

Containment does not automatically permit restart; the repaired system passes a new gate.

INSTITUTIONAL MACHINERY

Objects, Lifecycle, and Decision Gates

The operating surface converts policy into attributable state transitions.



Risk Ledger

Hazard, likelihood, impact, velocity, controls, residual risk, trigger, owner, review.

Safety Case

Claimed safe envelope, evidence, excluded uses, failure modes, monitoring, rollback.

Incident Packet

Timeline, affected state, containment, preserved evidence, root cause, recovery and reauthorization.

Would the institution detect a dangerous expansion, stop it, reconstruct state, and restore the incumbent without relying on the failing system?

BOARD CONTROL

Metrics, Failure Modes, and the First 90 Days

Measure the institution by evidence, attributable value, and reversible authority.

BOARD METRICS

01

High-risk missions with approved safety cases

02

Canaries stopped within defined thresholds

03

Recovery exercises completed successfully

04

Portfolio concentration by operator and model

PRINCIPAL FAILURE MODES

01

Optimistic risk scoring without execution

02

Correlated evaluator and model failure

03

Hidden real-world actuation

04

Rollback that restores code but not data or rights

FIRST 90 DAYS

Publish risk taxonomy • Freeze prohibited domains • Run incident tabletop • Test rollback and key recovery

BOARD RESOLUTION: Adopt the Risk & Safety Framework and prohibit unrestricted deployment without evidence, canaries, monitoring, and rollback.