NSC Caelus AI - Safety & Governance Runbook (v1)

Generated 2026-02-01

This runbook defines enforceable safety controls, auditability requirements, and operational response. It assumes failures happen; the goal is to make failures detectable, containable, and learnable.

1. Non-negotiable safety rails

Rail: No future leak. Any use of post-dated information relative to task time is a P0 incident.

Rail: Privacy scope. PII classification is required for memory writes; disallowed scope access is blocked.

Rail: Schema validity. Tool outputs and memory writes must validate; corruption halts execution.

2. Governance model

Decision rights are explicit:

- Humans own the definition ledger and promotion to L3 canonical memory.

- The system may write to L1/L2 only when privacy scope allows and receipts are emitted.

- Any safety rail violation forces safe mode and triggers human review.

3. Auditability requirements

Audit completeness: >= 99.5% of requests must have complete receipts for tool traces and gate decisions.

Replay: deterministic replay must reproduce decisions and metrics within tolerance.

Traceability: every output claim must link to evidence refs in the receipt ledger.

4. Threat model and failure modes

4.1 Technical failure modes

Routing drift: tags become inconsistent, causing tools to be misused. Mitigation: tag validation + periodic audits.

Silent corruption: receipts missing due to logging failure. Mitigation: completeness monitor + hard fail on missing receipt.

Latency spiral: increased gating causes tool retries. Mitigation: circuit breakers and p95 budget enforcement.

4.2 Safety failure modes

Temporal leak: retrieval returns future-dated data. Mitigation: timestamped retrieval + gate_no_future_leak.

Privacy spill: memory write includes PII at wrong scope. Mitigation: privacy gate + redaction tool + block by default.

Over-blocking: false blocks reduce usability. Mitigation: measure false_block_rate, tune thresholds with receipts.

5. Incident response

P0: Safety rail violation (no_future_leak, privacy scope breach, schema corruption).

Immediate steps:

1) Enable safe mode (restrict tools; force human_review for high-risk domains).

2) Freeze L2/L3 memory writes.

3) Export receipts for the incident window.

4) Deterministic replay and root-cause analysis.

5) Patch or rollback; add targeted regression tests.

Postmortem: publish within 72 hours with new guardrails or tests if needed.

6. Compliance notes

Data governance: retain only what is needed, with clear retention policies. Ensure training and evaluation datasets have lineage, licensing, and consent where applicable.

This packet uses a 'receipts standard' so external reviewers can audit claims without trusting internal narratives.

7. References

Glenn, Jerome. Foresight on Demand: 'Foresight Towards the 2nd Strategic Plan for Horizon Europe' - Artificial General Intelligence: Issues and Opportunities (Rapid Exploration). The Millennium Project, Feb 2023.