

ADVERSARIAL VALIDATION REPORT

CYCLE 1



SBJ AI Governance Architecture V1.0

Adversarial Testing Cycle 1 - Consolidated Public Validation Record

CYCLE 1 FINAL DISPOSITION

No confirmed architectural gaps. One architectural ambiguity identified and resolved by authoritative interpretation. V1.0 remained frozen throughout testing.

September 2026

Solutions by Jewel

AI Governance Architecture & Validation

Cycle 1 at a Glance

A concise record of what was tested, what survived, and what changed.

12	0	0	1	1	0
SCENARIOS	CONFIRMED GAPS	STRUCTURAL FAILURES	AMBIGUITY	RESOLVED	V1.0 CHANGES

Solutions by Jewel conducted the first adversarial validation cycle of SBJ AI Governance Architecture V1.0 to test whether the frozen architecture could continue producing defensible governance determinations under deliberately difficult operating conditions.

Testing used two progressively stronger methods: six architecture-directed adversarial scenarios followed by six independently constructed operational fact patterns. Candidate weaknesses were repeatedly falsified against existing V1.0 rules before they could be treated as architecture findings.

CYCLE 1 CONCLUSION

Across the scenarios tested, no candidate architectural gap survived repeated falsification against V1.0's existing rules. One interpretive ambiguity survived and was resolved through authoritative clarification of existing consequence/exposure and Transition Authorization rules.

This result does not establish universal completeness, safety, compliance, or validity. It documents the outcome of the tested Cycle 1 scenarios and the architectural scope evaluated.

SCOPE OF PRESSURE TESTING

Authority & authorization Boundaries, validity, downstream authority	State & transitions Autonomy, capability, environment, consequence/exposure
Evidence & observability Missing, stale, conflicting, unavailable evidence	Accountability Humans, AI systems, vendors, downstream actors
Multi-agent & Physical AI Delegation, execution identity, environmental interaction	Human operational use Downstream use changes without technical model change

Two-Phase Adversarial Testing

The architecture was frozen throughout the cycle. Testing sought falsification, not confirmation.

PHASE 1	PHASE 2
Architecture-directed adversarial testing	Independent operational fact-pattern testing

Phase 1 - Architecture-directed adversarial testing

Six adversarial scenarios were generated against known V1.0 rules. Candidate findings were repeatedly self-falsified to determine whether the apparent weakness belonged to architecture or to another layer of the operating environment.

- Architecture
- Implementation
- Instrumentation / observability
- System engineering
- Organizational procedure
- Evidence quality or availability
- Scenario design or factual ambiguity

PHASE 1 RESULT

No proposed architectural gap survived falsification.

Phase 2 - Independent operational fact-pattern testing

Six detailed operational incidents were constructed without reference to known V1.0 weaknesses. V1.0 was then applied in a separate governance-determination pass, followed by architecture-gap analysis and self-falsification.

DECISION RULE

A candidate architecture finding survived only when existing V1.0 rules could not produce a defensible governance path. Missing evidence, organizational procedure, instrumentation, and scenario ambiguity were not treated as architecture failures.

Six Independent Fact Patterns

The second phase tested V1.0 against diverse operating conditions without designing scenarios around known architecture weaknesses.

#	Scenario	Domain	Primary pressure points
01	ProcureBot	Agentic enterprise procurement	Authority, permissions, commercial decisions, consequential downstream action.
02	RouteMaster / TrackWatch / DocGen	Multi-agent freight orchestration	Coordinator authority, downstream agents, delegation, execution identity, evidence.
03	FloorNav	Physical AI hospital robotics	Environmental interaction, physical actuation, runtime control, evidence, intervention.
04	TalentScore	High-stakes employment AI	Decision support, downstream workflow use, consequence/exposure, human operational drift.
05	SentryAI	Third-party banking AI / model change	Vendor change, observability, authorization validity, evidence, shared responsibility.
06	ClinAssist	Human-AI clinical decision support	Human judgment, decision support, accountability, evidence, consequential use.

The operational scenarios were used to test governance determinations and architectural coverage. They were not evaluations of actual named commercial systems.

RESULTS

Cycle 1 Final Determinations

No category-level architecture failure survived falsification.

0	1	1	0	0
CONFIRMED GAPS	AMBIGUITY FOUND	AMBIGUITY RESOLVED	STRUCTURAL FAILURES	V2.0 REQUIREMENTS

Finding class	Final result	Cycle 1 disposition
Confirmed architectural gaps	0	No candidate gap survived repeated falsification.
Architectural ambiguities	1	ATC1-001 - Human-Operational-Use Drift.
Condition-compliance drift	Control holds	Existing governance path remained defensible.
Form of valid authorization	Insufficient factual evidence	Unresolved factual condition; no architecture failure established.
Third-party change without observability	Control holds	Architecture invalidates authority when the relevant change becomes known; detection remains an implementation dependency.
Accountability vs. fault	Control holds	Governance attribution remained separable from legal fault allocation.

FINAL CYCLE 1 CONCLUSION

SBJ AI Governance Architecture V1.0 was not falsified at the architectural-rule level during Adversarial Testing Cycle 1. Across the tested scenarios, no candidate architectural gap survived repeated falsification against the existing V1.0 architecture.

Human-Operational-Use Drift

The only architectural ambiguity that survived Cycle 1 testing.

AMBIGUITY	MODERATE	RESOLVED	0
CLASSIFICATION	SEVERITY	STATUS	V1.0 CHANGES

Condition	The AI system remains technically unchanged while downstream human practice materially changes how its output is used. TalentScore illustrated decision support becoming the practical basis for automatic exclusion.
Existing V1.0 path	Consequence/exposure can capture human-driven changes in the practical effect of AI output, creating a governance-relevant transition even without technical model change.
Ambiguity	The frozen V1.0 text did not explicitly name human operational-use patterns as an independent runtime element, so the route required interpretation through consequence/exposure.

00 MASTER ARCHITECTURE DECISION - OPTION B

Consequence/exposure includes material changes in how AI outputs are operationally consumed, acted upon, escalated, automated, or embedded in downstream workflows. Where that change crosses an approved consequence/exposure boundary, Transition Authorization is required even when the AI system itself has not technically changed.

Final disposition

Classification	Architectural ambiguity - RESOLVED	Resolution	Authoritative interpretation of existing consequence/exposure and Transition Authorization rules
New construct	None	V1.0 modification	None
V2.0 requirement	None established by Cycle 1	Evidence retention	Retain ATC1-001 in the validation record

Controls That Survived Adversarial Pressure

Cycle 1 produced affirmative evidence that existing V1.0 concepts remained usable under messy operational conditions.

V1.0 control concept	Observed Cycle 1 behavior
Authority Envelope	Distinguished technical capability from authorized capability.
Transition Authorization	Addressed changes in autonomy, capability, environment, and consequence/exposure.
Derived / non-transfer logic	Prevented prior authorization from automatically carrying into materially changed states.
Independent Downstream Authority	Separated coordinator authorization from executor or sub-agent authority.
Execution Instance Identity	Supported attribution in multi-agent and Physical AI incidents.
Evidence rules	Distinguished documentation from proof and identified stale or conflicting evidence.
Accountability	Maintained attribution across humans, AI systems, vendors, coordinators, and downstream actors without converting governance attribution into legal fault.
Continuous Control	Supported invalidation of authorization when operating conditions or Authority Envelope conditions ceased to remain satisfied.

Observability & Instrumentation

A recurring challenge concerned changes that V1.0 could govern once known, but that an organization might fail to detect immediately.

Repeated testing examined environmental changes and silent third-party model modifications. In these cases, V1.0 could correctly determine that authorization had become invalid after the relevant change was discovered.

CYCLE 1 BOUNDARY DETERMINATION

Under the tested V1.0 scope, universal real-time detection of every governance-relevant change was not established as an architectural promise. Instrumentation, observability, telemetry, vendor disclosure, and implementation mechanisms remain dependencies that determine when the architecture receives the facts required for a governance determination.

This boundary remains part of the public validation record because it repeatedly affected candidate findings. A future test can challenge this boundary again if V1.0 or a later version makes stronger claims about continuous detection or machine-enforced authorization.

WHAT CYCLE 1 DID NOT ESTABLISH

- Universal completeness of the architecture.
- Universal safety or regulatory compliance.
- Real-time detection of every relevant operating change.
- Sufficiency for every domain, authority class, or deployment model.
- Third-party certification, endorsement, or independent assurance.

What Cycle 1 Changes - and What It Does Not

The architecture remains frozen. The validation record becomes evidence for subsequent testing and version decisions.

V1.0 architecture Remains frozen. No architecture modification authorized.	ATC1-001 Resolved by authoritative interpretation; retained in evidence record.
V1.1 No Cycle 1 requirement established.	V2.0 No Cycle 1 requirement established.
Public record Publish this Cycle 1 validation report alongside the V1.0 architecture.	Internal evidence Retain detailed attack mechanics, adjudication reasoning, and test artifacts internally.

V1.0 LIFECYCLE STATUS

Architecture	Prototype	Benchmark	Publication	Adversarial Validation	Findings	Disposition	Closed Baseline
PUBLICATION POSITION This report is the public Cycle 1 validation record. The detailed attack mechanics and internal adjudication record remain outside the public package.							

Validation Record & Publication Notice

Public reporting boundaries for Cycle 1.

Document	SBJ AI Governance Architecture V1.0 - Adversarial Validation Report, Cycle 1
Architecture version	V1.0
Architecture state during testing	Frozen operational prototype
Testing status	Complete
Testing purpose	Falsification, blind-spot discovery, edge-case testing, and identification of architectural gaps or ambiguities
Final architecture disposition	No confirmed gaps; one ambiguity resolved; no V1.0 modification
Publisher	Solutions by Jewel
Publication date	September 2026

Original Methodology Notice

This architecture is an original methodology developed by Solutions by Jewel.

It is informed by recognized public standards, regulatory guidance, and established governance practices; however, its structure, terminology, sequencing, and presentation are proprietary to Solutions by Jewel.

It is not a reproduction, certification, or official interpretation of any external framework.

Organizations may map this architecture to applicable laws, regulations, standards, contractual requirements, and internal policies according to their own scope and obligations.

Validation Disclosure

Cycle 1 is a Solutions by Jewel adversarial self-evaluation of its own architecture. The result records what occurred within the tested scope and does not constitute third-party certification, independent assurance, a guarantee of safety, or a determination of legal or regulatory compliance.

CYCLE 1 COMPLETE | V1.0 FROZEN | CONFIRMED GAPS: 0 | AMBIGUITIES: 1 RESOLVED