Attack Types & Triage Rules — v0.8.1

---

1. Purpose

This document defines how specification-first contracts are attacked, evaluated, and declared stable.

Applies to any contract class.

Ensures **semantic stability — not formatting uniformity.**

---

2. Stability Definition

A contract is stable when:

> Independent competent implementations produce identical externally observable Decision-Surface outcomes for identical inputs.

Stability is defined at the **Decision Surface (DS)** — not byte surface.

---

3. Decision Surface (DS)

A finding affects DS if it can change:

- Success vs failure
- Accept vs reject
- Converge vs non-converge
- Create vs not create
- Delete vs retain

- Managed vs unmanaged
- Collision vs no collision
- Refusal behavior
- Any externally observable state transition

Only DS findings affect instability scoring.

---

4. Finding Classes

Each finding MUST be classified exactly once.

A = Ambiguity (DS)
I = Infeasibility (DS)
O-core = Environmental assumption affecting DS
O-deploy = Deployment/config
P = Presentation variance
SP = Spec Pollution / Implementation Policing

SP indicates contract degradation, not instability.

---

5. Byte-Level Conformance Policy

Byte identity required only if:

- Explicitly declared AND
- DS depends on exact byte equality

Otherwise byte escalation = SP.

---

6. Burden of Proof for Precision

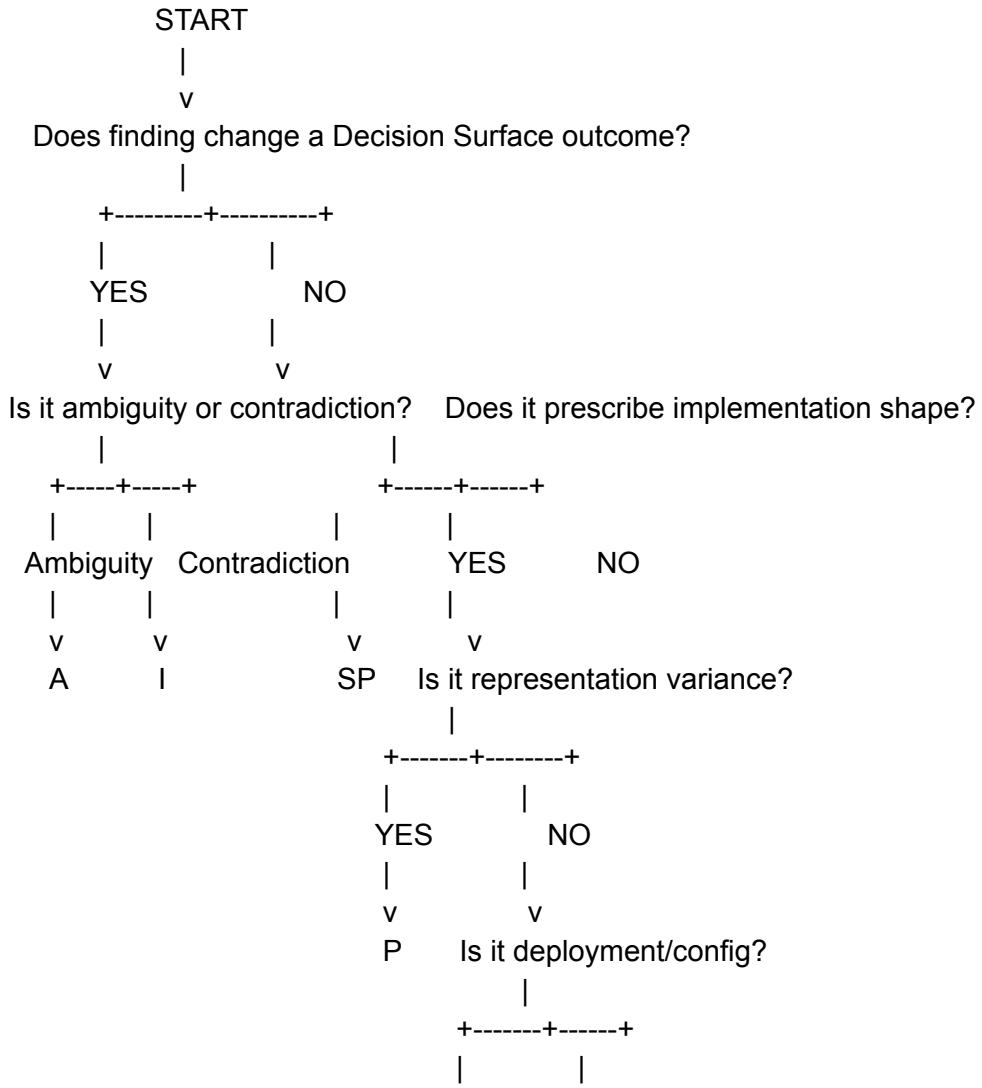
Added precision must show:

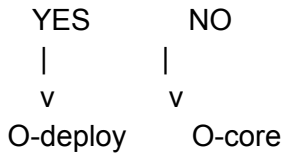
1. Which DS it stabilizes
2. What failure it prevents
3. Why semantic constraint insufficient

Else → SP.

---

7. Attack Classification Flow (RESTORED ASCII)





This flow is authoritative.

---

8. Decision-Surface Clarification Matrix

Before instability classification, verify DS impact on:

- Accept/reject
- Success/failure
- Convergence status
- Create/delete
- Refusal
- Authority boundary

If none change → Not instability.

---

9. Attacker Boundedness Principle

Attack must not:

- Invent new DS
- Expand scope
- Escalate precision without DS impact
- Demand canonical formatting without DS need

Unbounded escalation = attacker drift.

---

10. Precision Ceiling Rule

If DS stabilized
  AND no new contradiction
  AND no new externally observable divergence
THEN
  Further precision escalation = SP

---

11. Hypothetical Drift Constraint

Instability cannot rely on:

- Implausible incompetence
- Artificially naive implementations
- Contrived pathological inputs

Standard: Competent independent implementation.

---

12. No Byte-Level Escalation Rule

IF contract does not declare byte identity
  AND DS does not depend on it
THEN byte-level escalation = SP

---

13. Stability Metric

Instability Ratio =
  (DS A + DS I findings)
  /

(Total open DS findings)

Exclude:
P
SP
O-deploy

Thresholds:

- ≥ 50% → Unstable
- < 50% → Near Stable
- 0% → Stable

---

14. Stability Termination Condition

IF no open A or I findings affecting DS remain
THEN contract = STABLE

Closed DS issues cannot be reopened without new evidence.

---

15. Signal Ratio Guard

IF
>= 50% P or SP findings
AND < 25% DS A/I findings
THEN attacker drifting → recalibrate

---

16. Execution Model Independence

Contract must survive across:

- Languages

- Runtimes
- Execution ordering
- Hardware

Implementation-model binding without DS need = SP.

---

17. Contract ≠ Implementation

Contracts define:

- What must be true
- What decisions occur
- What causes refusal
- What state transitions allowed

Contracts do NOT define:

- Algorithms
- Data structures
- Libraries
- Internal execution order
- Formatting without DS impact

If attack turns contract into pseudocode → attacker violation.

---

18. Finite Convergence Requirement

Across rounds:
  DS findings must diminish
IF new DS findings arise solely from earlier precision escalation
  THEN show causal linkage
  ELSE classify as SP drift

Infinite escalation = attacker failure.

---

19. Evaluation Bias Suppression Protocol

19.1 Version-Blind Rule

Version identifiers and revision magnitude are opaque metadata.

They MUST NOT:

- Increase assumed maturity
- Lower scrutiny
- Be cited as stability evidence

19.2 Single-Pass Independence Rule

Each attack evaluates:

- Only the provided artifact
- No prior history
- No assumed previous fixes

19.3 Counter-Prior Principle

Default assumption:

Contract = Unstable
UNTIL
All DS A/I findings eliminated

Narrative maturity is not evidence.

19.4 Evidence-Only Stability Gate

Stability verdict MUST include:

- Explicit confirmation no A/I DS findings remain
- Confirmation no externally observable divergence exists

Stability may not rely on:

- Iteration count
- Version magnitude
- Structural density