# Attack Types & Triage Rules — v0.9

---

## 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.

This document also guards against false rigor.

Over-specification and implementation policing are treated as attack failures when they constrain acceptable implementation variance without stabilizing a governed outcome.

## 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.

**SP-IP** MAY be used as a reporting subtype when the pollution specifically consists of freezing an implementation choice without governed-outcome need.

SP-IP is still SP, not a new instability class.

# 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 an existing semantic constraint is insufficient
4. Why a weaker added constraint would not suffice
5. Why the proposed constraint governs an existing authority surface rather than inventing a new one

Else → SP.

# 6A. Outcome-Equivalence Rule

A finding is valid only if it identifies missing constraint on a governed outcome.

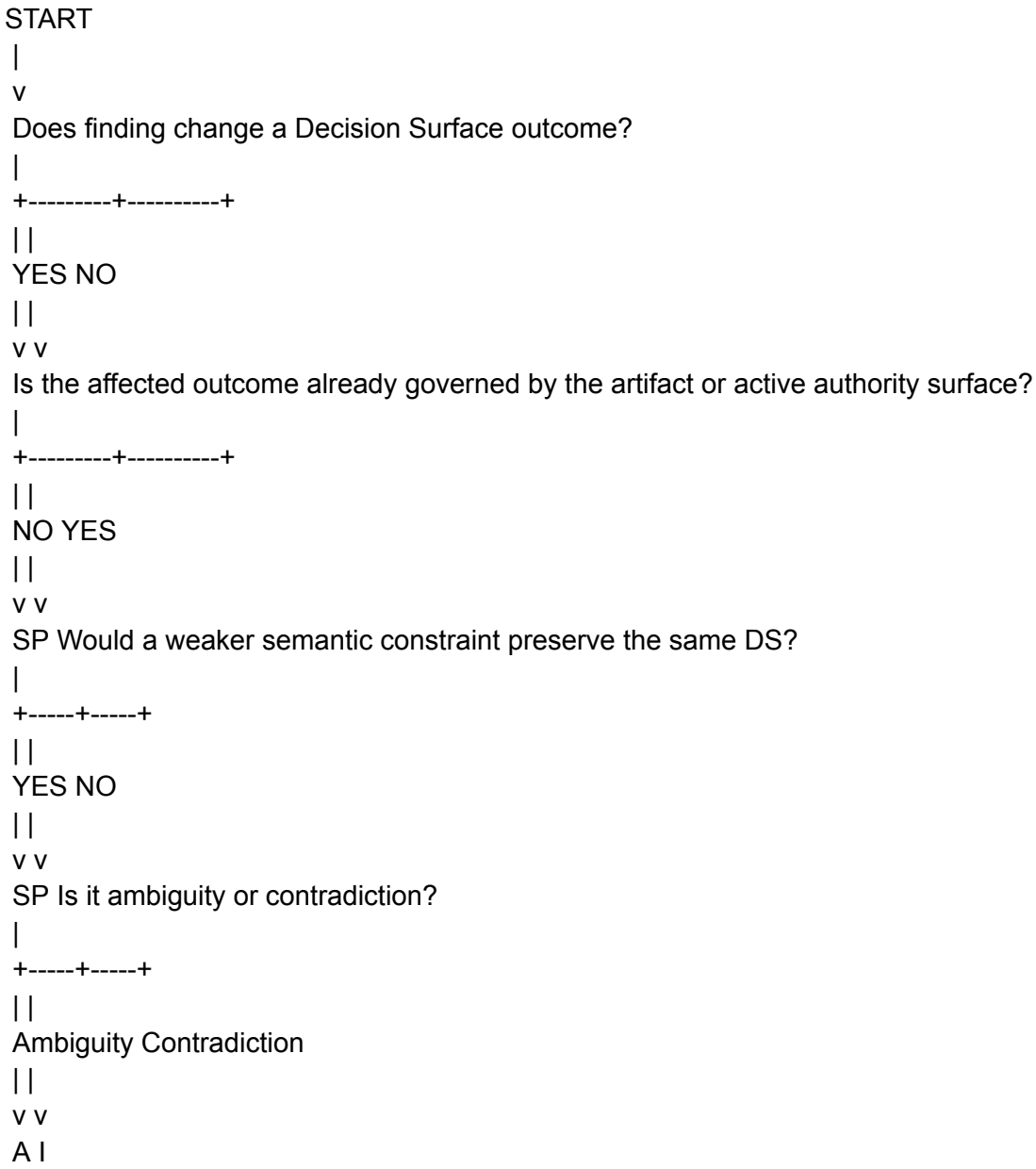
If multiple implementation choices produce the same governed Decision-Surface outcome, those choices are outcome-equivalent for attack purposes.

Demanding one canonical choice among outcome-equivalent variants SHALL be classified as SP.

Examples of outcome-equivalent variance MAY include:

- deterministic ordering by different acceptable strategies
- different safe metadata derivation heuristics
- different internal identifiers or representations
- different internal sequencing that preserves the same governed external state transition

# 7. Attack Classification Flow




If finding does not change DS:

- If it prescribes implementation shape → SP
- Else if it is representation variance → P
- Else if it is deployment/config → O-deploy
- Else → O-core

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.

# 8A. Established-Authority Gate

Before classifying a finding as A or I, the attacker MUST show that the allegedly missing constraint belongs to an already governed authority surface.

If the proposed precision would create a new authority surface not established by:

- the provided artifact, or
- the current authority-bearing implementation, where implementation is explicitly in play,

then the finding defaults to SP unless the contract explicitly intends to newly govern that surface.

The attacker MUST distinguish:

- missing constraint on an existing governed outcome

from

- invention of a new governed outcome

# 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.

# 9A. Minimal-Sufficient-Constraint Rule

When a finding proposes added precision, the attacker MUST prefer the least constraining wording that preserves the governed outcome.

If a weaker semantic constraint stabilizes the same Decision Surface, then demanding a stronger, more specific constraint SHALL be classified as SP.

Preference order:

1. outcome constraint
2. refusal / precedence constraint
3. deterministic-behavior constraint
4. specific strategy / representation constraint

Attack MUST stop at the first level that stabilizes the outcome.

# 9B. Acceptable Variation Recognition

Variation within a governed area is not automatically instability.

If the artifact governs the required outcome but leaves implementation latitude on method, representation, or ordering strategy, that latitude is acceptable unless it produces a different governed Decision-Surface result.

Attack MUST distinguish:

- governed outcome variance

from

- acceptable implementation variance inside a governed area

Misclassifying acceptable variation as instability SHALL be treated as SP drift.

# 10. Precision Ceiling Rule

If:

- the governed DS is already stabilized
- no new contradiction exists
- no new externally observable divergence exists
- and remaining variance is outcome-equivalent

then further precision escalation = SP.

This includes attempts to:

- freeze one acceptable deterministic strategy
- freeze one acceptable derivation heuristic
- freeze one acceptable internal representation
- freeze one acceptable implementation path

# 11. Hypothetical Drift Constraint

Instability cannot rely on:

- Implausible incompetence

- Artificially naive implementations
- Contrived pathological inputs

Standard: Competent independent implementation.

Attack may not rely on hypothetical implementation divergence unless the divergence is both:

- competent, and
- materially outcome-changing on a governed surface

Plausible internal variation that preserves outcome SHALL NOT support instability claims.

# 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

- = 40% of findings are P or SP, OR
- = 25% of findings are SP-IP, OR
- SP findings exceed DS A/I findings

THEN attacker drifting → recalibrate

IF repeated rounds are dominated by SP or SP-IP while DS A/I findings do not diminish, then the attack process is failing regardless of document length or revision count.

## 15A. Authority-Preserving Attack Checklist

Before raising A or I, attacker MUST be able to answer:

1. What exact governed outcome changes?
2. Is that outcome already in scope of the artifact?
3. Does the current wording permit materially different outcomes, not merely different methods?
4. Would a weaker semantic constraint stabilize the same outcome?
5. Is the proposed precision constraining outcome rather than implementation shape?

If any answer is negative or unproven, classify as SP or continue investigation without opening DS instability.

# 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.

# 17A. Complexity Inflation Warning

Over-specification is not neutral.

If added precision communicates a false need for complexity, narrows acceptable integration behavior, or increases perceived contract cost without stabilizing a governed outcome, that precision SHALL be treated as harmful SP.

Attack MUST avoid introducing wording that makes a system appear more complex than the governed outcome requires.

# 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

If repeated rounds keep producing findings that narrow acceptable implementation freedom without identifying new governed outcome divergence, attacker failure SHALL be presumed.

Such churn SHALL be classified as SP drift unless new evidence shows a real unresolved Decision-Surface split.

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