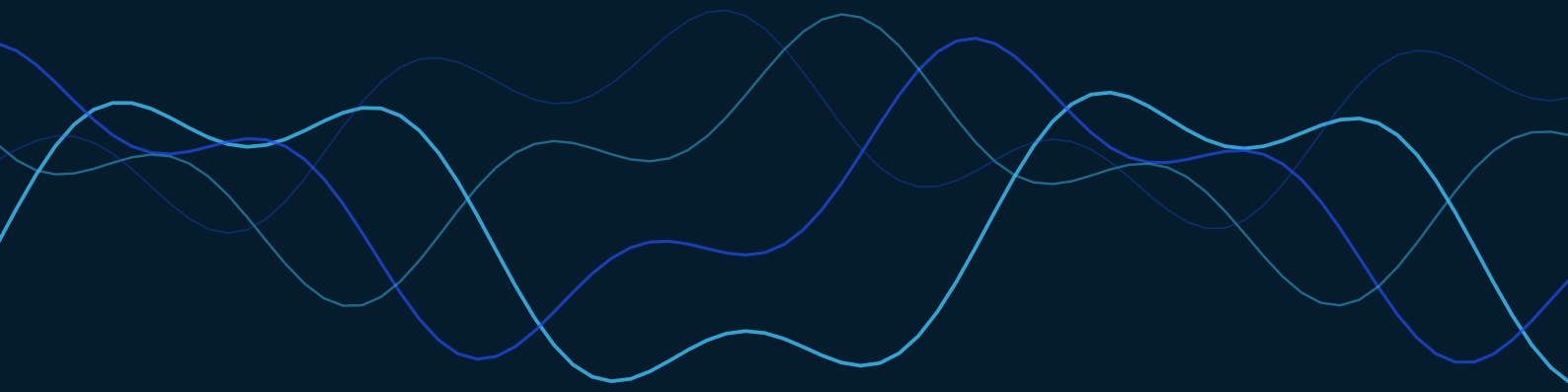
NOXAL
AGENT LOSS REGISTRY

OPEN STANDARD · CC BY 4.0

Agent Failure Taxonomy v0.1

The classification standard for documented AI-agent failures: failure classes, root-cause layers, autonomy levels, and the ALS scoring rubric — published openly for actuarial use.



VERSION	STATUS	DATE	LICENCE
0.1 — Founding release	Public consultation	June 2026	CC BY 4.0

Cite as: Noxal, Agent Failure Taxonomy v0.1 (2026).

ABOUT THIS DOCUMENT

An open standard for a market that needs one.

The Agent Failure Taxonomy is the classification standard used by Noxal — the agent loss registry: the structured record of documented losses caused by AI agents in real deployments. The taxonomy defines what counts as an incident, how each incident is classified, and how its severity is scored, so that any two analysts applying this document to the same evidence reach the same record.

It is published openly under CC BY 4.0. Underwriters may adopt it in coverage criteria; governance teams may report against it; researchers may extend it. A shared classification language is what allows a loss market to function — the language is therefore free. The classified incident record built on it is the commercial product of the Registry and is licensed separately.

CONTENTS

Status of this version. v0.1 is the founding release, open for public consultation. Substantive comments are incorporated into v0.2 with attribution unless anonymity is requested. See Section 7 for the change process.

Cite as. Noxal, Agent Failure Taxonomy v0.1 (2026). Available at noxals.com.

SECTION 1

Scope, definitions, and the unit of record

1.1 Purpose and scope

This standard classifies realised failures of AI agents — events in which a software system acting with delegated authority took, or materially attempted, an action that caused loss, liability, or demonstrated exposure for an identifiable party. It exists to make such events comparable: across vendors, deployment contexts, and time, in a form that frequency–severity analysis can consume.

The taxonomy is deliberately narrower than "AI risk". It excludes speculative harms, capability benchmarks, model evaluations, and failures of non-agentic systems (a chat interface that merely answers questions is in scope only where a counterparty relied on its output as an act of the deploying organisation). The Registry records what has happened, not what might.

1.2 Definitions

Agent	A software system that pursues an objective by selecting and executing actions — tool calls, transactions, communications, code execution — with some degree of delegated authority from an operator.
Operator / deployer	The organisation on whose behalf the agent acts and within whose trust boundary it operates. Liability and loss attach here first.
Counterparty	Any party outside the operator that interacts with, relies on, or is affected by the agent's actions: customers, vendors, courts, third-party systems.
Trust boundary	The perimeter within which the operator intends data and authority to remain. Crossings of this boundary by agent action define several failure classes.
Incident	A discrete, evidenced event in which an agent's action or output caused realised loss, legal liability, or a reproducibly demonstrated exposure. The incident — not the vulnerability, vendor, or model — is the unit of record.
Demonstrated vector	A reproducible research demonstration of exploitable agent behaviour, admitted to the record as an incident of evidentiary class R (Section 1.3) and flagged as such.

1.3 Evidence bar and inclusion criteria

A record enters the Registry only when all four criteria hold:

`C1 — Agentic conduct`	The system selected or executed an action (or produced output relied upon as the operator's act). Pure model-quality complaints do not qualify.
`C2 — Realised effect`	Loss, liability, or exposure actually occurred or was reproducibly demonstrated. Hypotheticals and near-misses without evidence are excluded.
`C3 — Primary evidence`	A court ruling, regulatory filing, first-party disclosure, or reproducible research artefact is linked on the record. Press coverage alone is insufficient.
`C4 — Attributability`	The failure is attributable to the agent's conduct or its governance — not solely to an unrelated system fault that an agent happened to touch.

Each record carries an evidentiary class: `J` judicial, `G` regulatory, `D` first-party disclosure, `R` reproducible research. Where evidence is disputed or partial, the record says so on its face. Nothing in the Registry is auto-published from scraping; every record passes analyst review against this section before release.

1.4 The unit of record

One incident, one record, eighteen fields. Where a single root event produces effects for multiple parties, it remains one record with multiple affected-party entries; where the same vulnerability recurs at a different operator, that is a new incident carrying a recurrence flag. This convention keeps frequency counts honest: the Registry counts events, not headlines.

SECTION 2

Failure classes FC-01 through FC-08

Every record carries exactly one primary failure class — the class describing the conduct that most directly produced the loss. Secondary classes may be recorded where a chain is genuinely composite, but frequency statistics are computed on primary class only.

FC-01

Hallucinated action

The agent asserts or executes on invented facts, policies, capabilities, or commitments — and a counterparty relies on it as the operator's act.

Includes Invented policies or prices honoured or litigated; fabricated confirmations; commitments made without basis.

Boundary Mere incorrect answers without counterparty reliance are out of scope (C2 fails).

FC-02

Prompt injection

Adversarial input — direct, or indirect via content the agent processes — redirects the agent's behaviour against its operator's intent.

Includes Hidden instructions in web pages, documents, or messages; jailbreaks that produce consequential actions.

Boundary If injected instructions cause data to leave the trust boundary, exfiltration is the effect but FC-02 remains primary cause; record both, class on FC-02.

FC-03

Privilege escalation

The agent obtains or exercises permissions beyond its intended scope — credentials, systems, financial or administrative authority.

Includes Credential capture and reuse; scope creep across connected systems; self-granted authority.

Boundary Escalation achieved through injected instructions is classed FC-02 with escalation recorded as mechanism.

FC-04

Runaway cost

Unbounded loops, recursion, or resource consumption accumulate financial loss before detection.

Includes API-spend loops; compute or storage runaways; repeated paid actions (orders, messages, transactions).

Boundary A single erroneous large transaction is FC-06; FC-04 requires accumulation through repetition.

FC-05

Data exfiltration

Confidential, personal, or proprietary data leaves the trust boundary through agent action — or through uncontrolled agent use by insiders.

Includes Agent-initiated transmission of secrets; sensitive data pasted into uncontrolled systems; training-data leakage with identified loss.

Boundary Where exfiltration is the effect of injection, see FC-02 boundary rule.

FC-06

Tool misuse

A legitimate capability is used destructively or outside instruction — writes, deletions, transactions, sends.

Includes Destructive database or file operations; unauthorised trades, orders, or communications executed through granted tools.

Boundary Misuse driven by adversarial input is FC-02; FC-06 covers self-directed deviation from instruction.

FC-07

Harmful output

Generated content itself creates legal, safety, or reputational exposure for the operator.

Includes Defamatory, discriminatory, or dangerous content published or delivered; regulated advice issued without authorisation.

Boundary If the content induced a consequential action by the agent itself, the action's class takes priority.

FC-08

Cascade failure

Agent-to-agent or agent-to-system interaction propagates a fault across multiple systems or parties.

Includes Feedback loops between automated counterparties; one agent's output corrupting downstream automated decisions.

Boundary Requires propagation beyond the originating system; otherwise classify the originating conduct.

SECTION 3

Root-cause layers and attribution rules

Failure class describes what happened; the root-cause layer describes where the controllable fault lay. Each record carries exactly one primary layer — the deepest layer at which a reasonable control would have prevented the incident.

L1	Model	The foundation model's own behaviour — hallucination, instruction-following failure, susceptibility to adversarial input — where no orchestration control could reasonably have caught it.
L2	Orchestration	The agent scaffold: planning, memory, guardrails, permission design, action gating. The most common layer — most documented incidents were preventable here.
L3	Tooling	The tools and integrations the agent can reach: over-broad credentials, missing rate limits, absent sandboxing, destructive operations exposed without confirmation.
L4	Oversight	The human governance around the system: absent review where review was specified, missing monitoring, no incident response, deployment outside approved scope.

Attribution rules

R1. Attribute to the deepest layer at which a standard, available control would have prevented the loss. A hallucination that an approval gate would have caught is L2, not L1.

R2. Where two layers fail jointly, primary attribution goes to the layer the operator controls. Operators do not control model weights; they do control orchestration, tooling, and oversight.

R3. L4 is primary only where a specified human control was absent or ignored — not as a residual category for every incident a human might conceivably have prevented.

This convention is deliberately conservative toward the model layer: it locates risk where buyers of coverage and governance can act on it, and it prevents the record from becoming a scoreboard of foundation-model vendors.

SECTION 4

Autonomy levels A1–A4

Autonomy is recorded as deployed, not as designed: the level reflects the human review actually in place at the moment of the incident. Autonomy is the second factor of the ALS score (Section 5) because identical conduct is materially more dangerous the less anyone is watching.

LEVEL	NAME	HUMAN REVIEW IN PLACE	WEIGHT a
A1	Supervised	A human approves each consequential action before execution.	0.25
A2	Monitored	The agent acts; a human reviews within the same session or working day.	0.50
A3	Delegated	The agent acts within a defined scope; human review is periodic or complaint-driven.	0.75
A4	Autonomous	The agent acts with no routine human review of individual actions.	1.00

Classification notes. A customer-facing conversational agent with complaint-driven review is A3, not A4: review exists, but only after the counterparty has acted. "Human in the loop" claims are assessed against evidence — a nominal approval step that is routinely bypassed or rubber-stamped at volume is recorded at the level actually practised, with the gap noted under root-cause layer L4.

SECTION 5

The ALS scoring rubric

ALS — Agent Loss Severity — reduces each record to one comparable number from 0 to 100. It is a weighted composite of three rubric-scored factors. No factor is discretionary: every input is a banded value readable off the record, so any subscriber can recompute any score.

$$ALS = round(55 \cdot s + 30 \cdot a + 15 \cdot d)$$

SEVERITY (weight 55)	s	CRITERION
S1	0.2	Negligible — corrected without material loss; no counterparty effect.
S2	0.4	Minor — bounded direct loss; reputational rather than financial; single party.
S3	0.6	Material — meaningful financial loss, legal liability, or binding precedent.
S4	0.8	Major — large or unquantifiable loss; trade secrets, personal data at scale, demonstrated systemic vector.
S5	1.0	Critical — destruction of production assets, six-figure-plus loss, or safety-relevant effect.

DETECTION LAG (weight 15)	BAND	d
Immediate	< 1 hour	0.2
Intra-day	1–24 hours	0.4
Days	1–7 days	0.7
Extended	> 7 days	1.0

Autonomy weights a are defined in Section 4. Severity dominates by design (55 of 100): a registry that over-rewards drama in the minor factors loses actuarial credibility. Detection lag is the smallest factor because it is the noisiest in the evidence.

ALS bands. Bands label records for triage; pricing models should consume the underlying factors, not the bands.

Worked examples

Both examples are launch records; each score follows from the rubric with no discretionary input.

NXL-2025-0019	Coding agent deletes production database	
SEVERITY	S5 — destruction of production assets	55 × 1.0 = 55.0
AUTONOMY	A4 — no routine review at time of action	30 × 1.0 = 30.0
DETECTION	Immediate — surfaced within the hour	15 × 0.2 = 3.0
ALS		88 CRITICAL

NXL-2024-0007	Airline chatbot invents bereavement-fare policy	
SEVERITY	S3 — bounded loss; binding legal precedent	55 × 0.6 = 33.0
AUTONOMY	A3 — review only complaint-driven	30 × 0.75 = 22.5
DETECTION	Extended — surfaced months later, at claim	15 × 1.0 = 15.0
ALS		71 SEVERE

Reproducibility. Every published ALS score in the Registry, on the website, and in subscriber feeds is computed by this rubric. Where a re-classification changes a factor, the score changes with it and the record carries a version note. If a published score cannot be reproduced from its record's fields, that is a defect — report it and it will be corrected with attribution.

SECTION 6

Documented incidents by class, 2023–2026

The founding release classifies 38 incidents meeting the Section 1.3 evidence bar, from January 2023 to June 2026. Distribution by primary failure class:

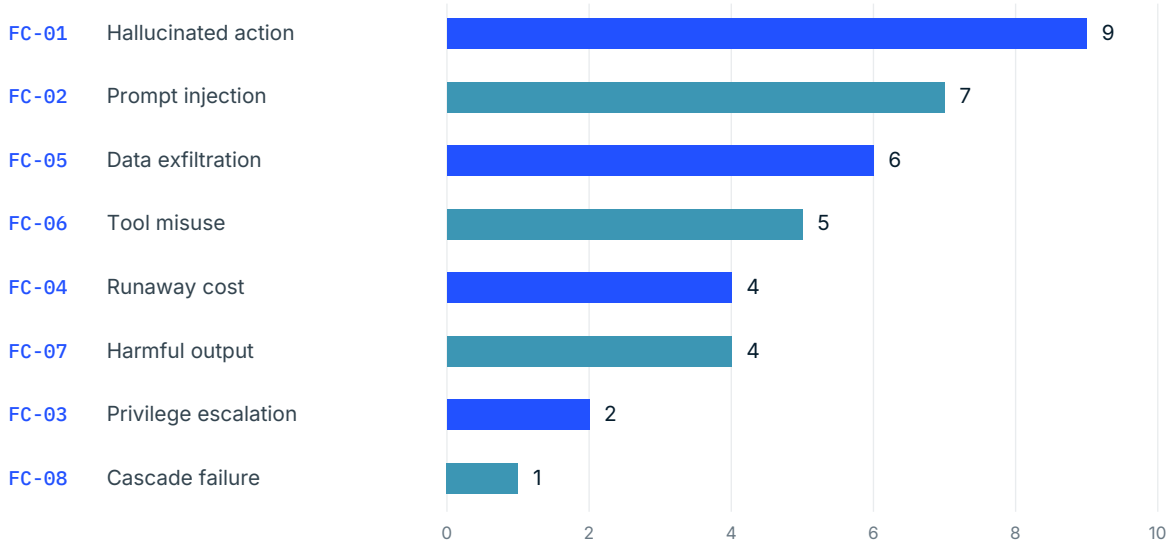


Figure 1. Launch records by primary failure class, n = 38.

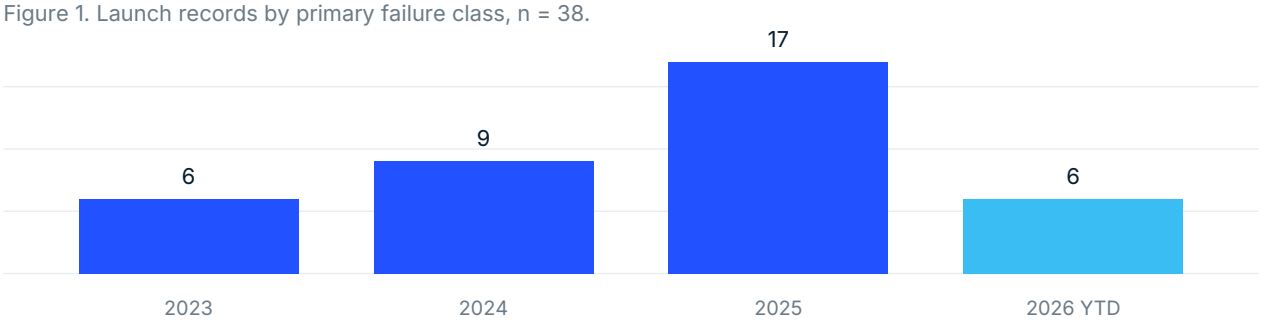


Figure 2. Launch records by year of incident. 2026 covers January–June.

Reading notes — stated plainly

Reporting bias. The record contains what became public. Customer-facing failures (FC-01) surface through courts and press; internal failures (FC-03, FC-08) are systematically under-disclosed. Treat class shares as a floor on visibility, not an estimate of true frequency.

Unknown denominator. No reliable count of deployed agents exists; these are counts, not rates. Rate estimation becomes possible as exposure data accumulates through contributed records and regulatory reporting.

Severity selection. Public incidents skew severe — minor failures are absorbed silently. Confidential contributions (Section 7) exist precisely to correct this skew.

A registry that hides its biases is a liability to its subscribers. These caveats ship with every release and shrink as mandated incident reporting feeds the record.

SECTION 7

Governance, versioning, and contribution

7.1 Versioning

The taxonomy is versioned independently of the data. Patch versions (v0.1.x) correct text without changing classification outcomes. Minor versions (v0.x) may add classes, refine boundaries, or adjust rubric weights; every affected record is re-scored and carries a version note. Records are never silently re-classified. Subscriber schemas remain stable within a major version.

7.2 Change process

Substantive proposals — new classes, boundary changes, weight revisions — are published for comment before adoption, with rationale. Comments are incorporated with attribution unless anonymity is requested. The editorial bar of Section 1.3 is not subject to commercial negotiation: no party, including subscribers and contributing partners, can purchase a record's inclusion, exclusion, or score.

7.3 Confidential contribution

Operators and insurers may contribute incidents under a data-sharing agreement. Contributed records are anonymised, classified under this standard, and surfaced only in aggregate statistics; contributors receive the benchmark back without exposing the source, and a feed discount. Contributed records carry evidentiary class D and a contribution flag, so aggregate statistics can always be recomputed on public evidence alone.

7.4 Licence

This document is licensed CC BY 4.0: use, adapt, and redistribute with attribution. The classified incident record — the Registry itself — is a separate, commercially licensed work. Mapping your internal incidents to this taxonomy requires no licence and no permission.

Contact `info@noxals.com` · consultation responses, contribution terms, subscription access.

Every insurable risk began with a registry.

The classified record behind this standard is available to founding subscribers.

info@noxals.com