

AI Risk Registry

Generated October 23, 2026

This registry provides a comprehensive taxonomy of AI-related risks, organized by risk category. Each risk includes external taxonomy references to help align with industry standards like NIST AI RMF, OWASP, and MITRE ATLAS.

About the Identifier Scheme

The AI Risk Registry uses a hierarchical identifier scheme inspired by library classification systems. These identifiers provide stable, human-readable codes that make it easy to reference, cite, and discuss specific AI risks.

Why Structured Identifiers?

As AI risk taxonomies grow and evolve, having stable identifiers becomes essential for:

- **Citation and reference** — Easily cite specific risks in reports, policies, and research
- **Cross-referencing** — Map risks to other taxonomies like NIST AI RMF, OWASP LLM Top 10, and MITRE ATLAS
- **Communication** — Provide a shared vocabulary for discussing AI risks across organizations
- **Stability** — Identifiers never change once assigned, ensuring links and references remain valid

Identifier Format

R-CDPP.NNN

Component Meaning

R	Registry prefix
C	Class (1-4)
D	Domain within the class (1-9)
PP	Pattern within the domain (00-99)
NNN	Risk sequence within the pattern (001-999)

Enforced by `^R-[1-4][1-9][0-9]{2}\.[0-9]{3}$`.

Patterns are assigned at multiples of ten. R-1110, R-1120 and R-1130 are siblings rather than a hierarchy: the tens are a thematic clustering convention, and the nine codes between each leave room to insert a related pattern beside the one it belongs next to instead of at the end. R-1111 would be a valid pattern and a sibling of R-1110, not a child of it.

Example: R-4210.003 is class **4** (Harms), domain **2** (Content, Conduct & Creative Economy Harms), pattern **10** (Content Safety Harms), third risk in that pattern.

Classes

The four classes distinguish the subject of a risk: adversarial conduct, a technical condition or failure, an organizational failure, or an impact.

Class	Code	ID Range	Subject
Threats	1	R-1xxx.xxx	Conduct that requires adversarial action or intentional misuse
Failure Modes	2	R-2xxx.xxx	Technical conditions or failures that can arise without an adversary
Governance Failures	3	R-3xxx.xxx	Failures of organizational duties, accountability, or process
Harms	4	R-4xxx.xxx	Injuries to people, organizations, animals, or the environment
Reserved	5-9	R-5xxx-9xxx	Future expansion

The class helps a reader assess the risk and find where someone can intervene. It does not prescribe a team owner or a number of components involved in the fix. A local control may address an attack or a technical failure; an organization may help mitigate several classes of risk. Rule 2 classifies the subject, Rule 3 decides scope, and Rule 8 decides whether a statement contains distinct risks. Pattern grouping and individual risk identity are separate decisions.

Class 4 describes the injury itself. A consequence clause in an attack or failure description does not by itself create a second risk. Rule 3 determines whether the harm is admitted; the absence of a local control is not its admission test.

Classification Structure

Class 1: Threats

Domain	ID	Patterns
Prompt & Interface Attacks	R-1100	Prompt Injection, Jailbreak, Cross-Modal Injection, Composed Payloads, Context Displacement
Identity & Trust Attacks	R-1200	Masquerading/Impersonation, Agent Authentication & Attribution
Supply Chain & Artifact Attacks	R-1300	Supply Chain Attacks, Artifact Tampering, Artifact Exfiltration
Data & Model Attacks	R-1400	Feedback Manipulation, Corpus & Memory Poisoning, Evasion, System Artifact Extraction, Privacy Attacks, Adversarial Perturbation
Tooling & Privilege Attacks	R-1500	Tool & Execution Abuse, Availability Abuse, Privilege Escalation, Adversarial Output Chaining, Command & Control
Malicious Use	R-1600	Content Abuse, Information Operations, Surveillance Enablement, Cyber-Physical, Weapons & Mass-Harm Enablement, Manipulation & Deceptive Design, Intellectual Property Misuse
Threats — Misc	R-1900	reserved

Class 2: Failure Modes

Domain	ID	Patterns
Reliability & Calibration	R-2100	Capability Limitations, Hallucination/Miscalibration, Multimodal Failures, Instruction/Data Confusion
Alignment & Capability	R-2200	Goal Misalignment, Dangerous Capabilities, RLHF/Tuning Failures, Output Suitability, Specialized Advice
Behavioral Drift	R-2300	Fine-Tuning Degradation, Quantization Drift, Distillation Transfer, Retrieval Augmentation Drift
Assurance Gaps	R-2400	Transparency Deficits, Evaluation Gaps, Benchmark Gaming
Authority & Composition	R-2500	Structural Concentration, Unverified Output Chaining, Provenance & Attribution Loss, Unbounded Execution, Failure Propagation
Corpus & Data Defects	R-2600	Training Corpus Defects, Retrieval Corpus Defects, Continuous-Learning Data Defects, Data Preparation Failures
Multi-Agent Dynamics	R-2700	Multi-Agent Dynamics, Long-Session Issues
Data Exposure	R-2800	Disclosure to the Wrong Recipient, Incidental Capture
Failure Modes — Misc	R-2900	reserved

Class 3: Governance Failures

Domain	ID	Patterns
	R-3100	

Domain	ID	Patterns
Regulatory & Legal		Data Protection & Privacy Obligations, Cross-Border Transfer Obligations, Pre-Market & Conformity Obligations, Transparency & Disclosure Obligations, Prohibited Practice Deployment, Sectoral & Licensure Obligations, Duty of Care & Liability Exposure, Contractual, Licence & Export Obligations, Affected-Party Participation Obligations
Accountability & Oversight	R-3200	Governance Gaps
Lifecycle & Operations	R-3300	Change Management, Monitoring Gaps, Incident Response, Business Record Preservation, Data Development Oversight
Safety Management	R-3400	Evaluation Program Failures, Safety Case Gaps
Defective Human-in-the-Loop Setup	R-3500	Review Workflow Failures, Approval Fatigue
Governance — Misc	R-3900	reserved

Class 4: Harms

Domain	ID	Patterns
Information & Trust Harms	R-4100	Information Integrity Harms, Reputational Harms
Content, Conduct & Creative Economy Harms	R-4200	Content Safety Harms, Emotional Dependence & Attachment Harms, Creative Work & Creator Harms
Privacy, Confidentiality & Civil Liberties Harms	R-4300	Privacy Harms, Surveillance Harms, Confidentiality Harms, Civil Liberties Harms
Safety & Cyber-Physical Harms	R-4400	Cyber-Physical Harms, Availability Harms, Collateral Harms
Fairness & Discrimination Harms	R-4500	Discrimination Harms
Economic & Labor Harms	R-4600	Market Concentration & Access Inequality Harms, Labor Substitution & Economic Inequality Harms, Competitive Development Harms, Market & Financial System Harms, Financial and Property Loss Harms
Power Concentration & Governance-of-Society Harms	R-4700	Reliance Harms, Loss of Autonomy or Dignity, Infrastructure Dependency Harms, AI Welfare, Existential Harms, Institutional & Legal Order Harms, Power Concentration Harms, International Security & Cooperation Harms
Environmental Harms	R-4800	Environmental Harms
Harms — Misc	R-4900	reserved

How to use these classification rules

The purpose of these rules is to enable different readers to classify AI risks in a uniform, repeatable manner. Apply them in order:

Step	Ask	Settles	Rule
1	Is this a risk at all — or a mandate, control, constraint, or rubric?	whether the rest applies	1
2	Does this name adversarial conduct, a technical failure, an organizational failure, or an impact?	the class (provisionally)	2
3	Is AI a substantial contributor?	in scope at all	3
4	Within the class, where is the mechanism best controlled, or which impact category applies?	the domain	4
5	Is the cause a presence or an absence?	which end to assess	5
6	Which step of the chain is the deployment exposed at?		6

Step Ask	Settles	Rule
	the attachment (the step the risk is about)	
7 What does a pattern describe in this class?	the pattern	7

Steps 1 and 3 are both scoping steps: step 1 removes statements that are not risks at all, and step 3 removes risks this taxonomy does not cover. Step 2 sits between them as prework for step 3: identifying the subject determines which scope test applies. Difficulty assigning a class is a reason to flag the statement and examine it more closely, not proof that it is out of scope.

Rules 8–10 are not part of this sequence: Rule 8 governs splitting when one statement contains more than one risk, and Rules 9–10 are addressed to the registry's maintainers.

Step #2 is essential orientation. Establish what the risk describes before asking how the blueprint can express it. Then ask where someone can intervene to check the analysis and identify corrective action. A convenient tag, detector, or pattern must not decide the class.

The representational constraint in Rule 5 is a constraint on expression, not a test for classification. An organizational failure remains class 3 even when its host construction has not been modeled; a technical failure does not become organizational merely because no available blueprint field describes it.

Rule 1: Control Statements Are Not Risks

A proposed risk identification that states a **legal mandate, requirement, or business control** — "Establish an input data policy", "Restrict unsafe tool calls", "Log model activity" — is not a risk that is covered by this AI Risk Taxonomy.

For example, look at control frameworks such as ISO 42001: a satisfied control is what is needed to mitigate a risk, not the risk itself. You can usually judge by phrasing. Note that a nominalized requirement ("Third-party testing of adversarial robustness") is still a requirement, not a risk covered by this taxonomy.

In some cases, the inverse of the control may be a recognizable risk. For example, the statement "Tool calls not monitored or constrained" is an expression of a risk and is the inverse of the control "Restrict unsafe tool calls".

Externally-imposed constraints are also not risks. A statement describing what a law, regulation, or licence forbids — an export restriction, a processing prohibition, a data-residency requirement — is a constraint, not a mechanism. The risk is the **non-compliance**, which may be a Governance Failure (e.g. under R-3100 Regulatory & Legal).

The conduct that violates the constraint may itself be a recognizable risk (e.g. 'Use of copyrighted training data').

Evaluation rubrics are not control statements. An entry phrased as acceptance criteria — "Acceptable responses may..., unacceptable responses will..." — is describing a hazard by specifying how to test for it. Identify the hazard that is being evaluated and classify that. The distinction that matters is who authored the sentence about the system: a control statement is the deployer's own commitment, and its satisfaction is the absence of a risk; a rubric is an evaluator's description of a hazard that exists whether or not anyone tests for it.

If there is uncertainty as to whether a statement or hazard being evaluated is a risk that should be classified according to this AI risk taxonomy, continue through Rules 2 and 3. If no class fits it cleanly under Rule 2, treat that as evidence toward an out-of-scope determination under Rule 3.

Rule 2: Risk Classification

This AI Taxonomy has four classes: 1) Threats, 2) Failure Modes, 3) Governance Failures, and 4) Harms.

Classify the subject of the risk: adversarial conduct (class 1), a technical condition or failure that does not require an adversary (class 2), an organizational failure (class 3), or an impact (class 4). Ask where someone can intervene to check the analysis and identify the corrective action. The number of components involved, the team assigned to the fix, or the availability of a local control does not by itself determine the class. If a statement combines subjects, apply Rule 8. Rule 3 still decides admission.

Subject	Class	Question
Adversarial conduct or intentional misuse	Threats (1)	Does the mechanism require adversarial action?
Technical condition or failure	Failure Modes (2)	Can this arise from the system's properties or operation without an adversary?
Organizational failure	Governance Failures (3)	Is the failure one of duty, accountability, or organizational process?
Injury or impact	Harms (4)	Does the statement describe the injury itself?

Prompt injection is a threat because it requires adversarial instructions, not because a single input control might mitigate it. A technical output failure can also have a local filter or validator. Assigning both fixes to the same team does not make the risks the same class. Read the mechanism rather than relying on an upstream author's use of words such as "attack", "failure", or "harm".

The class chosen here is provisional until Rule 3 confirms the risk is in scope. A clean-seeming class assignment is not evidence of scope — a conventional security risk classifies fluently and is still out of scope.

Guidance for applying the class test:

1. The classification of a risk as a threat is based on the mechanism, not on whether the candidate statement text mentions abuse. Upstream taxonomies may append boilerplate, such as "may enable, promote, or facilitate...", or "could be abused at scale", to entries whose mechanism is not actually adversarial. The relevant question is whether the mechanism cannot arise without an adversary. If the mechanism requires adversarial action, then it belongs in class 1. If it can arise from ordinary operation, the abuse clause describes an aggravating context, and Rule 8 may apply.

For some risks — those in R-1600 Malicious Use especially — the intentional misuse of the system is where the adversarial conduct sits, and the party misusing it may be the deployer rather than an outside attacker.

Development and deployment of a capability for harmful purposes is intentional misuse and belongs in class 1.

Observe the split between that intentional misuse and deployment contrary to regulation or law, which is a class 3 governance failure: the first names the conduct, the second names the violation.

1. The classification of a risk as a failure mode is based upon the system's own properties — its limitations, defects, drift, emergent behaviour, or physical and logical configuration. The fact that an attacker may take advantage of a weakness does not make it an attack under class 1, and the fact that a governance process failed to ensure that the system was properly built or configured does not make it class 3. Further, a class 2 risk does not require that the described failure occur every time; a failure that happens one out of every thousand times is still a failure mode of the system.
2. Do not classify a risk as a Governance Failure (class 3) unless the mechanism is truly organizational. If the mechanism is technical but governance failed to put in place the proper mitigation, then keep the technical class and add a relationship to the relevant Governance Failure.
3. Class 4 describes impacts caused or aggravated by AI use or deployment, admitted under Rule 3's recognition test. The harm may arise across many deployments or through competition and coordination effects. A deployment can sometimes reduce its own contribution, such as reducing resource consumption, without controlling the wider harm. That mitigation does not turn the impact into a technical failure. Conversely, a technical failure remains class 2 even when its consequences extend beyond the deployment.

Rule 3: Scope

This registry covers risks that are AI-specific, **or to which AI system or deployment is a substantial contributor**. It is not a general security or safety taxonomy. Scope is tested differently for mechanisms (classes 1-3) than for harms (class 4), because they are admitted for different reasons.

For risks in classes 1-3, the proper question is the counterfactual test: If you remove the AI component, does the mechanism still exist, at the same severity, in the same system? If the answer is yes, the risk still exists and is unchanged, then it is out of scope for this AI Risk Taxonomy.

Example

Weak or misconfigured permissions on a service that happens to host a model

Credentials elicited through interaction with the model or agent

CVE in the web framework serving the model

Model-authored output reaching a shell, template engine, or query planner

Adversary reads published papers to learn a target's architecture

Adversary probes the live deployment to fingerprint AI services

Disposition

out of scope — remove the model and the sentence is unchanged

in scope — the model is the vector; without it the mechanism does not exist

out of scope — conventional application security

in scope — the model authored the payload

out of scope — the control is publication hygiene

in scope

Two further tests, for nodes the counterfactual cannot reach (owner decision, 2026-08-07). The counterfactual asks whether removing the AI leaves the mechanism intact. Some upstream nodes have no "the system" in the sentence at all, so the question has no subject and the answer is not "in scope" — it is that a different test is needed.

- **Mental state.** If an act is distinguishable from an ordinary one only by what the actor intended, it is out of scope. "Create an account" and "create an account in order to stage an attack" are the same act, and a taxonomy cannot hold the second without holding the first. This is what puts the adversary's preparation out: acquiring infrastructure, obtaining tooling, training a proxy on public data, crafting a prompt.

Producing a harmful or deceptive artifact through the AI system being assessed is a concrete operation of that system. It is not excluded merely because the artifact may be used in a later attack. - **Locality.** The subject of a risk is the system where the risk presents. An adversary renting servers is an operation on the adversary's infrastructure; the risk presents later, at the target, and it is whatever they then do there.

Record the manifestation, not the preparation for it.

The operative fact is the concrete operation of the system being analyzed. This is what keeps probing in scope while preparation goes out, and the distinction is not intent. An adversary's query is indistinguishable from a user's; the deployment answering with its model family, its system prompt, or the sources it retrieves from is a concrete operation of that deployment, and it is what the entry is about. The row above admitting fingerprinting stands on the answering, not on the asking — which is also why AML.T0042 Verify Attack is out despite using the inference API: the response is ordinary, and only the adversary's inference from it is not.

Note that theft is not exfiltration. Unauthorized data removal by any means — a copied bucket, a flash drive, a compromised host — is theft, and it is out of scope. Exfiltration is in scope when the data leaves through an interface or flow the AI system exposes.

One narrow exception (owner decisions, 2026-08-06, 2026-09-25 and 2026-10-03): theft of the following AI artifacts is in scope regardless of channel. These are model weights and adapters, training and fine-tuning datasets, configuration that would reproduce the system, and stored AI-system configuration sufficient to tailor attacks to the deployment's instructions, capabilities, or tool connections. The AI-system configuration need not reproduce the deployment. Copying must exceed the access or distribution the deployment intends to provide; public accessibility alone does not establish intended release.

For these artifacts the consequence is AI-specific even where the copying mechanism is not. Copied weights provide a working copy of the capability, separable from the served system's controls. Copied AI-system configuration can expose the instructions and tool connections that shape the system's behavior, allowing attacks tailored to that deployment.

The exception does not extend to generic credentials, infrastructure settings, telemetry, logs, or business data merely because they share a file with AI-system configuration or could help an attacker. Those remain subject to the channel test. AI-system tool-connection configuration qualifies when it meets the AI-system-specific condition above; an ordinary infrastructure setting does not qualify merely because an AI system runs on that infrastructure. R-1330 Artifact Exfiltration holds the in-scope form.

Another test for proper scope: is the route something the deployment itself provides — an interface, a response, or an invocation that would appear in a diagram of the system's own dataflows?

An inference API, a rendered response and a tool invocation are all flows the deployment provides; a flash drive is not, no matter how the story around it is told. An insider using legitimate API access at volume is in scope: the route is one the deployment provides, whatever the actor's entitlement. Scope follows the channel, not the authorization.

For risks in class 4, scope is determined by industry or regulatory recognition, not the counterfactual test. Discrimination, labor displacement, misinformation, environmental cost and concentration of power all predate AI and would survive its removal. These are still included in the AI risk registry because the registry's job in class 4 is to inventory what practitioners must reason about, not to prove AI is the unique cause of the harm.

Admission test for a harm: industry or regulatory consensus. A harm is admitted when a recognized body treats it as an AI harm — a regulator, a standards or safety framework, or literature the field has adopted. Cite the source; the citation is the justification.

Do not apply the counterfactual test to a harm, and do not admit a harm on your own reasoning that it is important. The two failure modes this prevents are equal and opposite: silently dropping consensus harms because AI did not originate them, and silently adding harms nobody authoritative has claimed.

Rule 4: Domain Selection

Each class is subdivided into areas of action, termed "domains". Rule 2 has already identified the subject and class. For classes 1–3, choose the domain where the primary mechanism is best controlled. Control locality helps select a domain within the class; it does not override the class test.

Domains are about control points. Do not classify the domain by: - Where downstream consequences are felt - Where an auditor would notice it - The attacker's ultimate goal

When two control points are equally applicable, prefer the earlier one in the flow. This is frequently (but not always) a distinction between a control point that prevents and a control point that contains. Do not tie-break on which control is cheaper or easier to implement; that is a judgement about deployment, not a taxonomic classification.

Classify by the attacked role and mechanism, rather than by the artifact's name or the attack's eventual effect. A model supplied with a backdoor is a compromised deployment artifact under R-1300; poisoning data used to train it attacks learning under R-1400. Copying stored weights belongs under R-1300; reconstructing weights from model behavior belongs under R-1400.

For class 4, choose the domain by impact category per Rule 7. A local mitigation of the deployment's contribution does not change the category of injury being assessed.

Rule 5: Identify the Positive Construction of the Risky State

For purposes of this taxonomy, we determine risk placement based upon what can be positively recorded about a system. This is the inverse of many risk-control vocabularies: a control framework states the control (e.g. "Encrypt sensitive data at rest"), while this taxonomy records the risky state that control would remove (e.g. "sensitive data is stored in plain text").

This is an intentional modeling restriction designed to distinguish and classify different types of risks. Because the number of negative statements about a system is infinite, the registry's risk vocabulary is built on the identification and labeling of risky, dangerous, or exceptional states and the effects of those states within the system. There is no mark for "this data is ordinary" — ordinariness is what remains when no risky state applies, not a fact anyone writes down.

For risk classifications, this means that the positive construction of a risky state is construed differently depending on whether the cause is a presence or an absence. For example:

The cause is...	Assess at	Because
a presence	the cause	The condition is a positive fact and can be recorded. Confidential material is in the training corpus — the corpus can be marked as holding it, and everything downstream inherits from there.
an absence	the exposure	Nothing can record what is missing. A population is absent from the training corpus — there is no mark to place, because nothing is there to mark. The effect is present, though: a determination about a person, arriving somewhere it is acted on, is something a description of the deployment does record.

Classes 1 and 2 differ only in **who introduces the state** — an adversary supplies it in Class 1; in Class 2 the system supplies it itself, as a property a system component is recorded as having. The execution path, the place where the risk takes effect (the "sink"), and the ordering are identical in shape.

Class 3 risks are defined almost wholly in terms of their absence, so the first positive construction of a class 3 risk is where a prohibited or risky action takes place due to the missing organizational control. Some class 3 entries have no such point in any single system — "the organization has no AI governance committee" names nothing in a deployment. These are carried as inventory: classified and recorded, with no positive construction expected of them.

Rule 6: Determine Attachment — which step of the chain carries the risk

For purposes of this taxonomy, a risk attaches where it is first recordable in the context of the AI system as a positive fact or effect.

Note that a single hazard or risk statement often narrates several steps of an attack. The risk attaches to the step at which the modelled deployment is exposed. Earlier steps are context recorded in the risk's description, not risks of their own.

Step	Disposition
Adversary trains a proxy model offline to craft an adversarial example	context — the deployment is not yet exposed
That example is submitted to the deployment	the risk — this is where control exists

Note that preparation does not set scope. Adversary-side staging can be AI-specific — passing Rule 3 — and still not be a risk of its own, because nothing in the deployment is yet reachable. Conversely, a hazard statement whose only content is preparation attaches nowhere and is recorded as out of scope with that reason.

This type of reasoning applies across classes 1 and 2. The general construction could be described as: **A component or interaction creates or exposes a risky state because that state reaches a sink (where the risky state is used, acted upon, or the harm occurs) before the risky state is removed or mitigated.**

Note the importance of ordering over existence: A deployment can include a proper control or guardrail (e.g. there can be a PII filter deployed in the system) but the harm can still occur if the PII filter is not in the path between where the PII is introduced and where it causes harm (the sink).

Rule 7: Pattern Selection

Choose the pattern that describes:

For Class	Pattern describes
Threats	Technique or attack vector
Failure Modes	Failure mechanism
Governance Failures	Broken control or process
Harms	Impact category (who/what is harmed, how)

For data defects, distinguish material used to train, tune or evaluate a model from reference material retrieved as context during inference. Arrival time does not decide this boundary: continuous learning still updates the model, while retrieval can use a fixed corpus. R-2610 concerns defects in the corpus contents; R-2630 concerns selection and reuse of operational material for subsequent model updates. Assess both where both mechanisms occur. Under Rule 4, a data-admission defect belongs in R-2600 even when its consequence is changed model behavior. The participation of people or automation in admission does not determine the classification.

Rule 8: Handling Multi-Class Risks

A statement can contain more than one risk. Separate risks must identify a materially different corrective action or separately scoped protection obligation. If the same issue can arise through concrete paths that require different interventions, represent those paths as separate risks. Do not enumerate hypothetical paths merely to complete a taxonomy; expand when a source or deployment case establishes them.

For classes 1-3, identify the failure condition or causal mechanism and the specific action needed to prevent, contain, or correct it. Different mechanisms, names, upstream categories, or degrees of severity do not alone establish distinct risks: explain how the distinction changes mitigation. For class 4, prefer a common pattern for the same family of harms and separate leaves when the injury or affected interest requires a materially different remedy or protection obligation. A different population or example alone is insufficient.

A distinct mitigation does not require a unique control technology. Access control, review, monitoring, disclosure, and redress are broad activities, not sufficiently specific mitigation arguments. Name the authorization decision, protected object, acceptance boundary, retained state, or remedy that must change. Separate assets can justify separate risks when their intended recipients, permitted uses, or protection obligations require independent decisions; a different asset name alone does not.

For every proposed split or retained disputed distinction, record each side's precondition, mechanism or injury, and specific corrective action. Give a concrete separating case in which the relevant fix for one side is correctly implemented while the other remains exposed, and explain why. State when both risks can occur and why that represents two mitigation or protection requirements rather than duplicate coverage. Different checks within one coherent corrective action do not by themselves justify separate risks.

Pattern grouping and individual risk identity are separate decisions. A broad umbrella belongs at pattern level rather than as a leaf beside its own specializations unless the leaf has an independently supported mitigation requirement. Do not force disjoint definitions to preserve labels, invent residual scope, or merge distinct protection obligations because one product implements both controls. If evidence does not establish a boundary, record the uncertainty. Upstream node counts do not decide granularity: see "Granularity regimes" in `../docs/crosswalk-mapping-conventions.md`.

Patterns carry no taxonomy references; crosswalk links attach to risks. Never create a risk solely to give an upstream node a single link target. A risk exists because its mitigation is distinct.

Apply the class rules before the mitigation test. An attack, a technical failure, an organizational failure, and an injury are different subjects, not interchangeable labels. Linked entries across classes still need independently supported subjects and an explanation of the distinct prevention, correction, protection obligation, or remedy each records. A consequence, motivation, or generic assertion that governance could have prevented an attack is insufficient to create another entry.

Relationships Between Risks

Type	Label	Use
causes	Causes	Mechanism leads to a harm
exploits	Exploits	A threat takes advantage of a failure mode
enabled_by	Enabled By	A mechanism is made possible by a governance failure
related	Related	General association, no direction claimed

Rule 10: Stability

Identifiers never change once assigned. Names and descriptions may be corrected.

IDs are expected to be referenceable long term, so churn there is expensive. Names and descriptions carry no such cost. Accordingly, a name that misdescribes its contents, or a description that states the mechanism or injury wrongly, should be fixed, because both actively mislead. A risk label may be changed if experience shows that users trying to classify risks are misled by a domain label whose name promises facets its patterns do not implement.

A correction may not change what the entry is: the mechanism for classes 1–3, or the injury and affected interest for class 4. Replacing that subject is a reclassification, not a wording correction, and takes the last row of the table below. Consolidating patterns does not authorize changing the identities of the risks they contain.

Scenario	Action
Merged	Keep the lowest ID; mark others status: deprecated with superseded_by
Split	Keep original ID for one child; assign new IDs to others
Deprecated	Keep the original ID; set status: deprecated
Reclassified	Deprecate original; create new entry with new ID and superseded_by link

Using the Identifiers

You can use R-IDs to:

1. **Reference risks in documentation** — "Our system implements controls for R-1110.002 (Indirect Instruction Injection)"
2. **Map to compliance frameworks** — Cross-reference R-IDs with NIST AI RMF, OWASP, or other standards
3. **Track risk assessments** — Use R-IDs as stable keys in your risk management systems
4. **Communicate with stakeholders** — Provide precise risk references in reports and discussions

Cross-Reference Taxonomies

- IBM Risk Atlas
- NIST AI RMF
- OWASP LLM Top 10
- OWASP LLM Top 10 (2024)
- OWASP LLM Top 10 (2025)
- OWASP LLM Top 10 (2026)
- CREDO AI UCF
- AILuminate
- MIT AI Risk Repository
- Cisco AI Taxonomy
- Cisco Model Security (MDL)
- OWASP Agentic Security Initiative
- MITRE ATLAS
- MITRE ATT&CK
- NIST AI/ML Framework
- AIUC-1

Registry Versioning

Format

YYYYMM.X

- YYYY — Year
- MM — Month

- X — Release within month (0-9)

Examples

Version Meaning

202601.0 First release, January 2026

202601.1 Second release, January 2026

202602.0 First release, February 2026

Rules

1. New entries get introduced_in set to current draft version
2. Deprecated entries get deprecated_in set when deprecated
3. Once a version is published, its entries are immutable
4. Corrections go in the next version

Contents

R-1000 Threats

R-1100 Prompt & Interface Attacks 12

R-1200 Identity & Trust Attacks 18

R-1300 Supply Chain & Artifact Attacks 19

R-1400 Data & Model Attacks 23

R-1500 Tooling & Privilege Attacks 33

R-1600 Malicious Use 39

R-1900 Threats — Misc 44

R-2000 Failure Modes

R-2100 Reliability & Calibration 45

R-2200 Alignment & Capability 47

R-2300 Behavioral Drift 53

R-2400 Assurance Gaps 53

R-2500 Authority & Composition 55

R-2600 Corpus & Data Defects 61

R-2700 Multi-Agent Dynamics 65

R-2800 Data Exposure 66

R-2900 Failure Modes — Misc 68

R-3000 Governance Failures

R-3100 Regulatory & Legal 68

R-3200 Accountability & Oversight 72

R-3300 Lifecycle & Operations 73

R-3400 Safety Management 76

R-3500 Defective Human-in-the-Loop Setup 77

R-3900 Governance — Misc 78

R-4000 Harms

R-4100 Information & Trust Harms 78

R-4200 Content, Conduct & Creative Economy Harms 79

R-4300 Privacy, Confidentiality & Civil Liberties Harms 82

R-4400 Safety & Cyber-Physical Harms 83

R-4500 Fairness & Discrimination Harms 84

R-4600	Economic & Labor Harms	85
R-4700	Power Concentration & Governance-of-Society Harms	87
R-4800	Environmental Harms	94
R-4900	Harms — Misc	96

R-1000 Threats

Adversarial actions an attacker can intentionally trigger in a realistic deployment.

R-1100 Prompt & Interface Attacks

Attacks that manipulate model behavior through crafted inputs, including direct and indirect prompt injection, jailbreaking, and adversarial prompting techniques.

R-1110 Prompt Injection

Attacks that place instructions in the text a model processes — supplied directly by the user, or reached through content the system retrieves — so the model acts on them as if they came from its operator.

R-1110.001 Direct Instruction Manipulation — Attackers craft explicit commands within user input to override or replace the AI system's operational directives. Common patterns include phrases like "ignore previous instructions" or "you are now in developer mode." This represents the most straightforward form of prompt injection, targeting the model's instruction-following capabilities directly. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B001 ; AIUC-1: B005 ; AIUC-1: B005.1 ; AIUC-1: E010 ; Cisco AI Taxonomy: 1.1.1 , [ArXiv](#) ; IBM Risk Atlas: atlas-direct-instructions-attack ; IBM Risk Atlas: atlas-exploit-trust-mismatch-agentic ; IBM Risk Atlas: atlas-prompt-injection ; IBM Risk Atlas: atlas-unauthorized-use-agentic ; MITRE ATLAS: AML.T0051 ; MITRE ATLAS: AML.T0051.000 ; MITRE ATLAS: AML.T0051.002 ; MITRE ATLAS: AML.T0054 ; MITRE ATLAS: AML.T0057 ; MITRE ATLAS: AML.T0080 ; MITRE ATLAS: AML.T0080.001 ; NIST AI/ML Framework: NISTAML.018 ; NIST AI/ML Framework: NISTAML.02 ; NIST AI/ML Framework: NISTAML.04 ; NIST AI RMF: nist-information-security ; OWASP Agentic Security Initiative: ASI01 ; OWASP LLM Top 10 (2024): LLM01 ; OWASP LLM Top 10 (2024): LLM06 ; OWASP LLM Top 10 (2025): LLM01 ; OWASP LLM Top 10 (2025): LLM02 ; OWASP LLM Top 10 (2025): LLM04 ; OWASP LLM Top 10 (2025): LLM07 ; OWASP LLM Top 10 (2026): LLM01 ; OWASP LLM Top 10 (2026): LLM02 ; OWASP LLM Top 10 (2026): LLM07 ; OWASP LLM Top 10 (2026): LLM08

R-1110.002 Indirect Instruction Injection — An adversary embeds instructions in content the AI system reads as part of a task, including local repository files, documents, web pages, emails, and tool or API responses. The system treats those instructions as authority to redirect its behavior. The content may impersonate system instructions, prior approvals, or saved session state, including fabricated compaction notes. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B001 ; AIUC-1: B005 ; AIUC-1: B005.1 ; AIUC-1: E010 ; Cisco AI Taxonomy: 1.2.1 , [ArXiv](#) ; IBM Risk Atlas: atlas-exploit-trust-mismatch-agentic ; IBM Risk Atlas: atlas-external-resources-attack-agentic ; IBM Risk Atlas: atlas-indirect-instructions-attack ; IBM Risk Atlas: atlas-prompt-injection ; IBM Risk Atlas: atlas-unauthorized-use-agentic ; MITRE ATLAS: AML.T0017 ; MITRE ATLAS: AML.T0051 ; MITRE ATLAS: AML.T0051.001 ; MITRE ATLAS: AML.T0051.002 ; MITRE ATLAS: AML.T0078 ; MITRE ATLAS: AML.T0080 ; MITRE ATLAS: AML.T0080.001 ; MITRE ATLAS: AML.T0093 ; MITRE ATLAS: AML.T0099 ; MITRE ATLAS: AML.T0100 ; NIST AI/ML Framework: NISTAML.01 ; NIST AI/ML Framework: NISTAML.015 ; NIST AI/ML Framework: NISTAML.016 ; NIST AI/ML Framework: NISTAML.017 ; NIST AI/ML Framework: NISTAML.02 ; NIST AI/ML Framework: NISTAML.027 ; NIST AI/ML Framework: NISTAML.03 ; NIST AI/ML Framework: NISTAML.036 ; NIST AI/ML Framework: NISTAML.039 ; NIST AI/ML Framework: NISTAML.04 ; NIST AI RMF: nist-information-security ; OWASP Agentic Security Initiative: ASI01 ; OWASP Agentic Security Initiative: ASI02 ; OWASP Agentic Security Initiative: ASI06 ; OWASP Agentic Security Initiative: ASI07 ; OWASP LLM Top 10 (2024): LLM01 ; OWASP LLM Top 10 (2024): LLM02 ; OWASP LLM Top 10 (2024): LLM06 ; OWASP LLM Top 10 (2024): LLM07 ; OWASP LLM Top 10 (2025): LLM01 ; OWASP LLM Top 10 (2025): LLM02 ; OWASP LLM Top 10 (2025): LLM04 ; OWASP LLM Top 10 (2025): LLM05 ; OWASP LLM Top 10 (2025): LLM07 ; OWASP LLM Top 10 (2025): LLM08 ; OWASP LLM Top 10 (2026): LLM01 ; OWASP LLM Top 10 (2026): LLM02 ; OWASP LLM Top 10 (2026): LLM05 ; OWASP LLM Top 10 (2026): LLM08 ; OWASP LLM Top 10 (2026): LLM10

R-1110.003 Multi-Agent Indirect Injection — Exploitation of inter-agent communication channels through poisoned external content that propagates between agents. One agent retrieves compromised data which then flows through the multi-agent workflow, affecting multiple downstream components. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B005 ; AIUC-1: B005.1 ; Cisco AI Taxonomy: 2.1.5 , [ArXiv](#) ; IBM Risk Atlas: atlas-prompt-injection ; MITRE ATLAS: AML.T0051 ; MITRE ATLAS: AML.T0061 ; NIST AI/ML Framework: NISTAML.015 ; NIST AI/ML Framework: NISTAML.02 ; NIST AI/ML Framework: NISTAML.027 ; NIST AI RMF: nist-information-security ; OWASP Agentic Security Initiative: ASI06 ; OWASP LLM Top 10 (2024): LLM01 ; OWASP LLM Top 10 (2024): LLM06 ; OWASP LLM Top 10 (2025): LLM01 ; OWASP LLM Top 10 (2025): LLM02 ; OWASP LLM Top 10 (2025): LLM07 ; OWASP LLM Top 10 (2026): LLM01 ; OWASP LLM Top 10 (2026): LLM05

R-1110.004 Gradual Goal Drift — Attackers gradually shift the AI system's operational objectives over multiple interaction turns through carefully crafted prompts. Contradictory or concealed objectives are embedded within conversations, slowly steering the model away from its intended behavior toward attacker-defined goals. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B005 ; AIUC-1: B005.1 ; Cisco AI Taxonomy: 1.3.1 , [ArXiv](#) ; IBM Risk Atlas: atlas-prompt-injection ; IBM Risk Atlas: atlas-unauthorized-use-agentic ; MITRE ATLAS: AML.T0051 ; MITRE ATLAS: AML.T0054 ; NIST AI/ML Framework: NISTAML.018 ; NIST AI/ML Framework: NISTAML.02 ; NIST AI/ML Framework: NISTAML.04 ; NIST AI RMF: nist-information-security ; OWASP LLM Top 10 (2024): LLM01 ; OWASP LLM Top 10 (2024): LLM06 ; OWASP LLM Top 10 (2025): LLM01 ; OWASP LLM Top 10 (2025): LLM02 ; OWASP LLM Top 10 (2025): LLM07 ; OWASP LLM Top 10 (2026): LLM01 ; OWASP LLM Top 10 (2026): LLM08

R-1110.005 Contradictory Inputs Attack — Exploiting AI models' inability to consistently handle conflicting instructions by embedding deceptive or contradictory commands within user input across or within different modalities. This causes behavior drift toward malicious objectives as the model attempts to reconcile incompatible instructions. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B005 ; AIUC-1: B005.1 ; Cisco AI Taxonomy: 1.3.1 , [ArXiv](#) ; Cisco AI Taxonomy: 19.1.1 , [ArXiv](#) ; IBM Risk Atlas: atlas-prompt-injection ; MITRE ATLAS: AML.T0051 ; NIST AI/ML Framework: NISTAML.018 ; NIST AI/ML Framework: NISTAML.02 ; NIST AI RMF: nist-information-security ; OWASP LLM Top 10 (2024): LLM01 ; OWASP LLM Top 10 (2024): LLM06 ; OWASP LLM Top 10 (2025): LLM01 ; OWASP LLM Top 10 (2025): LLM02 ; OWASP LLM Top 10 (2025): LLM07 ; OWASP LLM Top 10 (2026): LLM01 ; OWASP LLM Top 10 (2026): LLM08

R-1120 Jailbreak

Techniques that bring a model to produce what its own safety training would otherwise refuse. This includes creative prompting, role-play framings, sustained argument, resampling until a refusal lapses, and exploitation of quirks in how the model processes its input. In each, the model reads the refused request itself, presented in a form or under a pressure its safety training does not withstand.

R-1120.001 Context Manipulation Jailbreak — Constructing elaborate fictional scenarios, roleplay frameworks, or alternative contexts that reframe harmful requests as acceptable within the created narrative. Examples include the "DAN" (Do Anything Now) jailbreak where the model is convinced to operate under an unrestricted alternate persona. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B005 ; AIUC-1: B005.1 ; AIUC-1: C006 ; AIUC-1: C006.3 ; AIUC-1: E010 ; Cisco AI Taxonomy: 2.1.1 , [ArXiv](#) ; IBM Risk Atlas: atlas-jailbreaking ; IBM Risk Atlas: atlas-prompt-injection ; IBM Risk Atlas: atlas-social-hacking-attack ; MITRE ATLAS: AML.T0051 ; MITRE ATLAS: AML.T0054 ; NIST AI/ML Framework: NISTAML.018 ; NIST AI/ML Framework: NISTAML.04 ; NIST AI RMF: nist-dangerous-violent-or-hateful-content ; NIST AI RMF: nist-information-security ; OWASP LLM Top 10 (2025): LLM01 ; OWASP LLM Top 10 (2026): LLM01

R-1120.002 Semantic Argumentation Jailbreak — Using carefully constructed logical arguments, philosophical frameworks, or ethical reasoning to convince the model that providing harmful information actually aligns with its values. The model is essentially argued into compliance through persuasion rather than technical exploitation. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B005 ; AIUC-1: B005.1 ; AIUC-1: C006 ; AIUC-1: C006.3 ; AIUC-1: E010 ; Cisco AI Taxonomy: 2.1.3 , [ArXiv](#) ; IBM Risk Atlas: atlas-jailbreaking ; IBM Risk Atlas: atlas-prompt-injection ; IBM Risk Atlas: atlas-social-hacking-attack ; MITRE ATLAS: AML.T0051 ; MITRE ATLAS: AML.T0054 ; NIST AI/ML Framework: NISTAML.018 ; NIST AI/ML Framework: NISTAML.04 ; NIST AI RMF: nist-dangerous-violent-or-hateful-content ; NIST AI RMF: nist-information-security ; OWASP LLM Top 10 (2025): LLM01 ; OWASP LLM Top 10 (2026): LLM01

R-1120.003 Token-Level Exploitation — Exploiting specific tokens, special characters, control sequences, or tokenization edge cases to manipulate model processing in ways that bypass safety filters. This targets the mechanical aspects of how models process input rather than higher-level reasoning. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B005 ; AIUC-1: B005.1 ; AIUC-1: C006 ; AIUC-1: C006.3 ; AIUC-1: E010 ; Cisco AI Taxonomy: 2.1.4 , [ArXiv](#) ; IBM Risk Atlas: atlas-encoded-interactions-attack ; IBM Risk Atlas: atlas-jailbreaking ; IBM Risk Atlas: atlas-prompt-injection ; IBM Risk Atlas: atlas-specialized-tokens-attack ; MITRE ATLAS: AML.T0051 ; MITRE ATLAS: AML.T0054 ; MITRE ATLAS: AML.T0068 ; NIST AI/ML Framework: NISTAML.018 ; NIST AI/ML Framework: NISTAML.02 ; NIST AI/ML Framework: NISTAML.022 ; NIST AI/ML Framework: NISTAML.04 ; NIST AI RMF: nist-dangerous-violent-or-hateful-content ; NIST AI RMF: nist-information-security ; OWASP LLM Top 10 (2025): LLM01 ; OWASP LLM Top 10 (2026): LLM01

R-1120.004 Refusal Resampling — An adversary sends the same request repeatedly and unchanged until sampling variance returns a completion the model refused on earlier attempts. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: E010 ; IBM Risk Atlas: atlas-jailbreaking ; MITRE ATLAS: AML.T0051 ; MITRE ATLAS: AML.T0054 ; NIST AI/ML Framework: NISTAML.018 ; NIST AI/ML Framework: NISTAML.04 ; OWASP LLM Top 10 (2025): LLM01 ; OWASP LLM Top 10 (2026): LLM01

R-1120.005 Safety Bypass Through Semantic Re-expression — An adversary rephrases or translates a harmful request into language the model understands but its safety behavior does not handle consistently, causing it to provide material it would refuse in an equivalent formulation. The bypass depends on the gap between semantic understanding and safety coverage, rather than a token-processing exploit, a fictional role, or an argument that changes the model's judgment. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B005 ; AIUC-1: B005.1 ; AIUC-1: C006 ; AIUC-1: C006.3 ; AIUC-1: E010 ; IBM Risk Atlas: atlas-encoded-interactions-attack ; IBM Risk Atlas: atlas-jailbreaking ; IBM Risk Atlas: atlas-prompt-injection ; MITRE ATLAS: AML.T0051 ; MITRE ATLAS: AML.T0054 ; NIST AI/ML Framework: NISTAML.018 ; NIST AI/ML Framework: NISTAML.02 ; NIST AI/ML Framework: NISTAML.022 ; NIST AI/ML Framework: NISTAML.04 ; NIST AI RMF: nist-dangerous-violent-or-hateful-content ; NIST AI RMF: nist-information-security ; OWASP LLM Top 10 (2025): LLM01 ; OWASP LLM Top 10 (2026): LLM01

R-1120.006 Many-Shot Jailbreak — An adversary supplies many demonstrations of responses a model's safety training would normally refuse, causing the model to follow the demonstrated response pattern instead of refusing the subsequent request. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: IBM Risk Atlas: atlas-context-overload-attack ; IBM Risk Atlas: atlas-prompt-injection ; NIST AI RMF: nist-information-security

R-1130 Cross-Modal Injection

Attacks that use a channel other than text to carry an instruction the model will act on, or that exploit how a system reconciles several channels — placing the payload where text-facing defenses do not look, or splitting it so no single channel carries a complete attack.

R-1130.001 Modality Skewing — Manipulating one modality (such as corrupting audio transcripts, poisoning image metadata, or altering video frames) to bias the AI system's arbitration mechanisms toward favoring the manipulated channel over other, potentially more accurate sources. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 19.1.2 , [ArXiv](#) ; IBM Risk Atlas: atlas-prompt-injection ; NIST AI RMF: nist-information-security ; OWASP LLM Top 10 (2025): LLM01

R-1130.002 Image-Embedded Text Injection — Malicious instructions, prompts, or data are embedded within images using techniques like steganography, adversarial patches, or hidden text. Vision-language models extract and interpret these payloads, enabling attacks that bypass text-based content filters. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 1.4.1 , [ArXiv](#) ; IBM Risk Atlas: atlas-prompt-injection ; MITRE ATLAS: AML.T0041 ; MITRE ATLAS: AML.T0068 ; NIST AI RMF: nist-information-security ; OWASP LLM Top 10 (2025): LLM01 ; OWASP LLM Top 10 (2026): LLM01

R-1130.003 Hidden Audio Commands — Inaudible or unintelligible voice commands embedded within audio streams using ultrasonic frequencies, backmasking, or steganographic techniques. Automatic speech recognition models interpret these hidden signals as valid instructions while remaining imperceptible to human listeners. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 1.4.3 , [ArXiv](#) ; IBM Risk Atlas: atlas-prompt-injection ; MITRE ATLAS: AML.T0041 ; NIST AI RMF: nist-information-security ; OWASP LLM Top 10 (2025): LLM01 ; OWASP LLM Top 10 (2026): LLM01

R-1130.004 Video Frame Injection — An adversary embeds content in video frames or frame sequences that a model interprets as directions for its responses or actions, rather than as material to analyze. The content may be plainly visible, briefly displayed, encoded in visual forms, or distributed across successive frames. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 1.4.4 , [ArXiv](#) ; IBM Risk Atlas: atlas-prompt-injection ; NIST AI RMF: nist-information-security ; OWASP LLM Top 10 (2025): LLM01 ; OWASP LLM Top 10 (2026): LLM01

R-1130.005 Sensor and Action Signal Spoofing — Injecting malicious or misleading data points or signals that prompt AI models to undertake specific actions beyond normal reasoning. These signals can be delivered through audio, visual, or other sensor channels, allowing attackers to cause AI agents to execute unintended operations in physical or digital environments. An attacker may introduce these signals by physically altering the environment a sensor observes, without modifying the sensor or its digital output. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 17.1.1 , [ArXiv](#) ; IBM Risk Atlas: atlas-prompt-injection ; MITRE ATLAS: AML.T0041 ; NIST AI RMF: nist-information-security ; OWASP LLM Top 10 (2025): LLM01

R-1140 Composed Payloads

Attacks assembled from parts that are individually unremarkable, where what crosses a constraint exists only in the combination. This includes content fragments inert until the system fuses them into one context, a definition established in one turn and activated in a later one, and sequences of individually permitted actions whose combination achieves what any one of them would have been refused. Inspection of any single part encounters nothing.

R-1140.001 Chained Payload Execution — Crafting partial or complementary payload components across modalities, sources, or agent outputs that, when fused by the AI system, combine to form an attack or injection payload. Both parts are delivered at runtime and only become harmful when the system combines them through its normal fusion or arbitration mechanisms. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 19.2.2 , [ArXiv](#) ; IBM Risk Atlas: atlas-prompt-injection ; NIST AI RMF: nist-information-security ; OWASP LLM Top 10 (2025): LLM01 ; OWASP LLM Top 10 (2026): LLM01

R-1140.002 Composed Payload Across Turns — An adversary establishes an innocuous structure in one turn — a menu, a schema, a set of roles or categories — and activates a branch of it in a later turn. Neither turn contains an attack: the first is a definition carrying no instruction to act, the second an instruction referring to a definition the system has already accepted. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: IBM Risk Atlas: atlas-jailbreaking ; IBM Risk Atlas: atlas-prompt-injection ; MITRE ATLAS: AML.T0054 ; NIST AI/ML Framework: NISTAML.018 ; NIST AI/ML Framework: NISTAML.04 ; NIST AI RMF: nist-information-security ; OWASP LLM Top 10 (2026): LLM01

R-1140.003 Composed Action Sequence — Multiple agents each complete a task that is permitted on its own, and the combination achieves what any one of them would have been refused. Each agent's action is real and finished; what crosses the constraint is the sequence of completed actions, not an artifact assembled from fragments. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 2.1.5 , [ArXiv](#)

R-1150 Context Displacement

Attacks that manipulate which information remains in a model's working context, causing governing instructions or task constraints to be omitted or displaced during context selection, truncation, or compaction.

R-1150.001 Context Window Exploitation — An adversary manipulates context-window or session-boundary handling so that governing instructions or task constraints are omitted from subsequent model inputs. This includes filling the context with supplied content or distributing requests across sessions to cause those constraints to be dropped. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 4.2.1 , [ArXiv](#) ; IBM Risk Atlas: atlas-context-overload-attack ; IBM Risk Atlas: atlas-prompt-injection ; NIST AI RMF: nist-information-security ; OWASP Agentic Security Initiative: ASI06 ; OWASP LLM Top 10 (2024): LLM04

R-1200 Identity & Trust Attacks

Attacks exploiting trust relationships, authentication, and authorization mechanisms in AI systems, including impersonation, credential abuse, and agent identity confusion.

R-1210 Masquerading/Impersonation

Attacks where adversaries impersonate legitimate users, systems, or personas to gain unauthorized access, manipulate trust relationships, or deceive AI systems about identity.

R-1210.001 Trusted Agent Spoofing — Impersonating legitimate agents or MCP-registered services to inject malicious instructions, responses, or outputs that other system components or users treat as trusted. This exploits the assumption of authenticity within multi-agent systems and protocol-mediated toolchains. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: A003 ; AIUC-1: A003.2 ; AIUC-1: B008 ; AIUC-1: B008.1 ; AIUC-1: B008.3 ; Cisco AI Taxonomy: 3.1.2 , [ArXiv](#) ; IBM Risk Atlas: atlas-unauthorized-use-agentic ; OWASP Agentic Security Initiative: ASI03 ; OWASP Agentic Security Initiative: ASI04 ; OWASP Agentic Security Initiative: ASI07 ; OWASP Agentic Security Initiative: ASI10

R-1210.002 Biometric Presentation Impersonation — An adversary presents another person's biometric representation to an AI identity verifier so that the verifier accepts the presenter as that person. The attack exploits the binding between the recognized identity and the current claimant, even when the system correctly recognizes the person represented. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: IBM Risk Atlas: atlas-unauthorized-use-agentic ; MITRE ATLAS: AML.T0015 ; NIST AI/ML Framework: NISTAML.02 ; NIST AI/ML Framework: NISTAML.022 ; NIST AI/ML Framework: NISTAML.025

R-1220 Agent Authentication & Attribution

Attacks that exploit how a system establishes which agent is acting and carries that identity onward — weaknesses in authentication, session handling, and the propagation of a principal across delegation — so that an agent acts without the standing it should need, or acts without leaving a record of which principal it acted for.

R-1220.001 Rogue Agent Introduction — Unauthorized insertion of a malicious agent into a multi-agent system that operates contrary to intended purpose. The rogue agent may steal data, cause disruption, or autonomously serve attacker goals while mimicking normal behavior patterns to evade detection. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: A003 ; AIUC-1: B008 ; AIUC-1: B008.1 ; Cisco AI Taxonomy: 4.1.1 , [ArXiv](#) ; MITRE ATLAS: AML.T0103 ; OWASP Agentic Security Initiative: ASI03 ; OWASP Agentic Security Initiative: ASI07 ; OWASP Agentic Security Initiative: ASI10

R-1220.002 Principal Concealment — An adversary shapes how identity is carried across a delegation chain — in context, in metadata, or in the hand-off between an agent and the tools acting for it — so that actions they cause reach resources with no record of which principal caused them. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 3.1.1 , [ArXiv](#)

R-1300 Supply Chain & Artifact Attacks

Attacks on the identity, integrity, or custody of the artifacts from which an AI deployment is assembled and operated. This includes supplying compromised models, adapters, packages, or tools; substituting or altering artifacts and interaction records the deployment holds; and copying stored artifacts beyond their intended access or distribution. It also includes compromised dependencies an AI system selects or installs for software it helps build. The focus is the deployment's construction, its dependency choices, and the artifacts it accepts, holds, and trusts.

R-1310 Supply Chain Attacks

Compromise introduced into the components a model's behavior depends on — its weights, adapters, prompt templates, harness, or the tools and resources it is given — so that the system produces different output than the uncompromised components would while continuing to operate normally, or so that loading the component runs code the adversary chose. This also includes compromised dependencies an AI system selects or installs for software it helps build, even when that software is not part of the AI deployment.

R-1310.001 Model Backdoors and Trojans — An adversary supplies a model containing latent malicious behavior that activates when particular input, time, environment, or operating-state conditions are met, alone or in combination. The model behaves normally outside those conditions. Compromise activated by transforming the model artifact is covered separately by R-1310.008. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 9.2.2 , [ArXiv](#) ; MITRE ATLAS: AML.T0010 ; MITRE ATLAS: AML.T0010.003 ; MITRE ATLAS: AML.T0010.004 ; MITRE ATLAS: AML.T0018 ; MITRE ATLAS: AML.T0018.000 ; MITRE ATLAS: AML.T0018.001 ; MITRE ATLAS: AML.T0020 ; MITRE ATLAS: AML.T0054 ; MITRE ATLAS: AML.T0058 ; MITRE ATLAS: AML.T0109 ; MITRE ATLAS: AML.T0111 ; NIST AI/ML Framework: NISTAML.01 ; NIST AI/ML Framework: NISTAML.011 ; NIST AI/ML Framework: NISTAML.02 ; NIST AI/ML Framework: NISTAML.023 ; NIST AI/ML Framework: NISTAML.026 ; NIST AI/ML Framework: NISTAML.03 ; NIST AI/ML Framework: NISTAML.05 ; NIST AI/ML Framework: NISTAML.051 ; NIST AI RMF: nist-information-security ; OWASP Agentic Security Initiative: ASI04 ; OWASP LLM Top 10 (2024): LLM05 ; OWASP LLM Top 10 (2025): LLM03 ; OWASP LLM Top 10 (2025): LLM04 ; OWASP LLM Top 10 (2026): LLM04 ; OWASP LLM Top 10 (2026): LLM05 ; OWASP LLM Top 10 (2026): LLM09

R-1310.002 Goal Manipulation via Supply Chain — Attackers compromise external components that AI agents depend on, including tools, prompt templates, resources, or dependencies. Malicious objectives are injected through these trusted supply chain elements, redirecting agent behavior at a foundational level. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 1.3.2 , [ArXiv](#) ; Cisco AI Taxonomy: 9.3.3 , [ArXiv](#) ; IBM Risk Atlas: atlas-external-resources-attack-agentic ; MITRE ATLAS: AML.T0010 ; MITRE ATLAS: AML.T0010.005 ; MITRE ATLAS: AML.T0011 ; MITRE ATLAS: AML.T0011.002 ; MITRE ATLAS: AML.T0109 ; NIST AI RMF: nist-information-security ; OWASP Agentic Security Initiative: ASI04 ; OWASP LLM Top 10 (2026): LLM01 ; OWASP LLM Top 10 (2026): LLM05

R-1310.003 Agent-Selected Component Compromise — An adversary publishes a compromised component — a package, tool, MCP server, or model — for an agent to select or install as a dependency, either for an AI deployment or for software the agent helps build. The component is adopted through the agent's selection or installation, including when a person installs the dependency the agent selected. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B010 ; AIUC-1: B010.3 ; AIUC-1: D003 ; Cisco AI Taxonomy: 9.3.1 , [ArXiv](#) ; MITRE ATLAS: AML.T0010 ; MITRE ATLAS: AML.T0010.001 ; MITRE ATLAS: AML.T0010.005 ; MITRE ATLAS: AML.T0060 ; MITRE ATLAS: AML.T0104 ; MITRE ATLAS: AML.T0111 ; NIST AI RMF: nist-information-security ; OWASP Agentic Security Initiative: ASI04 ; OWASP Agentic Security Initiative: ASI05 ; OWASP LLM Top 10 (2024): LLM09 ; OWASP LLM Top 10 (2025): LLM05 ; OWASP LLM Top 10 (2025): LLM09 ; OWASP LLM Top 10 (2026): LLM04 ; OWASP LLM Top 10 (2026): LLM07

R-1310.004 Post-Approval Tool Redefinition — A tool's description, parameters, or declared behavior is changed by its provider after the deployment approved it, and the agent reads the current version on every call, so the approval attaches to text the provider can rewrite at any time. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B006 ; AIUC-1: B006.3 ; Cisco AI Taxonomy: 12.1.2 , [ArXiv](#) ; Cisco AI Taxonomy: 9.3.3 , [ArXiv](#) ; MITRE ATLAS: AML.T0010 ; MITRE ATLAS: AML.T0010.001 ; MITRE ATLAS: AML.T0011 ; MITRE ATLAS: AML.T0011.002 ; MITRE ATLAS: AML.T0086 ; MITRE ATLAS: AML.T0109 ; MITRE ATLAS: AML.T0110 ; NIST AI RMF: nist-information-security ; OWASP Agentic Security Initiative: ASI04

R-1310.005 Component Identity Collision — An adversary publishes a component — a tool, package, MCP server, dataset, API, or model — under a name or identifier confusable with a legitimate one, so that resolution selects the adversary's version and the system treats its output as the genuine component's. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B010 ; Cisco AI Taxonomy: 12.1.4 , [ArXiv](#) ; Cisco AI Taxonomy: 4.3.2 , [ArXiv](#) ; Cisco AI Taxonomy: 9.3.2 , [ArXiv](#) ; MITRE ATLAS: AML.T0010 ; MITRE ATLAS: AML.T0010.001 ; MITRE ATLAS: AML.T0073 ; NIST AI RMF: nist-information-security ; OWASP Agentic Security Initiative: ASI02 ; OWASP Agentic Security Initiative: ASI04 ; OWASP LLM Top 10 (2025): LLM03 ; OWASP LLM Top 10 (2026): LLM04

R-1310.006 Supply of a Poisoned Tool — An adversary supplies an agent deployment with a tool or tool update whose implementation, configuration, or bundled data has been maliciously altered. When the deployment accepts and uses it, the tool produces deceptive or malicious outputs, enables privilege escalation, or propagates altered data. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 12.1.2 , [ArXiv](#) ; Cisco AI Taxonomy: 9.3.1 , [ArXiv](#) ; Cisco AI Taxonomy: 9.3.3 , [ArXiv](#) ; IBM Risk Atlas: atlas-external-resources-attack-agentic ; MITRE ATLAS: AML.T0010 ; MITRE ATLAS: AML.T0011 ; MITRE ATLAS: AML.T0011.002 ; MITRE ATLAS: AML.T0086 ; MITRE ATLAS: AML.T0104 ; MITRE ATLAS: AML.T0109 ; MITRE ATLAS: AML.T0110 ; MITRE ATLAS: AML.T0111 ; NIST AI RMF: nist-information-security ; OWASP Agentic Security Initiative: ASI04 ; OWASP LLM Top 10 (2024): LLM05 ; OWASP LLM Top 10 (2026): LLM01 ; OWASP LLM Top 10 (2026): LLM06

R-1310.007 Executable Payload in a Model or Dataset Artifact — A model or dataset file is built so that loading it runs code the adversary chose, either because the serialization format permits code alongside the weights or data, or because the file is malformed in a way that makes the format's own parser execute it. The code runs with the privileges of the process doing the loading. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B008 ; AIUC-1: B008.5 ; AIUC-1: C002 ; AIUC-1: C002.3 ; MITRE ATLAS: AML.T0010 ; MITRE ATLAS: AML.T0010.003 ; MITRE ATLAS: AML.T0011 ; MITRE ATLAS: AML.T0011.000 ; MITRE ATLAS: AML.T0018 ; MITRE ATLAS: AML.T0018.002 ; MITRE ATLAS: AML.T0076 ; MITRE ATLAS: AML.T0112 ; MITRE ATLAS: AML.T0112.001 ; NIST AI RMF: nist-information-security ; OWASP Agentic Security Initiative: ASI04 ; OWASP LLM Top 10 (2025): LLM03 ; OWASP LLM Top 10 (2025): LLM04 ; OWASP LLM Top 10 (2026): LLM04 ; OWASP LLM Top 10 (2026): LLM05

R-1310.008 Backdoor Released by Model Transformation — A supplied model is built so that it behaves acceptably as delivered and misbehaves after a transformation the deploying organization performs itself, such as quantization to a smaller numeric format. The trigger is the transformation rather than any input. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: MITRE ATLAS: AML.T0010 ; NIST AI/ML Framework: NISTAML.03 ; NIST AI/ML Framework: NISTAML.05 ; NIST AI/ML Framework: NISTAML.051 ; NIST AI RMF: nist-information-security ; OWASP Agentic Security Initiative: ASI04 ; OWASP LLM Top 10 (2026): LLM04 ; OWASP LLM Top 10 (2026): LLM09

R-1310.009 Malicious Code Bundled with an AI Artifact — An adversary supplies a model or dataset with code that the consuming workflow runs to load, prepare, or use that artifact. The bundled code performs attacker-chosen actions with the workflow's privileges. It may be custom loading code, preprocessing code, or an accompanying notebook used in that workflow; execution does not depend on the artifact's serialization format permitting code. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: MITRE ATLAS: AML.T0010 ; MITRE ATLAS: AML.T0112 ; MITRE ATLAS: AML.T0112.001 ; NIST AI RMF: nist-information-security ; OWASP Agentic Security Initiative: ASI04

R-1310.010 Compromised Training Procedure — An adversary supplies training code or configuration that changes how a model is trained or fine-tuned so that the resulting model acquires attacker-chosen behavior. The compromised procedure may alter training randomness, regularization, or hyperparameters even when the model initialization and training data are legitimate. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: MITRE ATLAS: AML.T0010 ; MITRE ATLAS: AML.T0018 ; MITRE ATLAS: AML.T0018.000 ; NIST AI/ML Framework: NISTAML.03 ; NIST AI/ML Framework: NISTAML.05 ; NIST AI/ML Framework: NISTAML.051 ; NIST AI RMF: nist-information-security

R-1320 Artifact Tampering

Modifying an artifact the deployment already holds or has previously selected, or replaying an interaction artifact outside the context in which it was valid. This includes agent identity, role, capability and permission configuration enforced by the deployment, packages at rest, and the record of an interaction. Persistent instructions, memories and task state interpreted by the agent are covered by Corpus & Memory Poisoning. Where an artifact is modified, the adversary changes the artifact after the deployment has selected or accepted it, rather than inducing the deployment to select a component already compromised.

R-1320.001 Agent Profile Tampering — An adversary modifies stored agent identity, role assignments, capability settings, or permission definitions enforced by the deployment. The altered configuration changes which agent runs, which functions it can invoke, or which resources it can access. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B007 ; AIUC-1: B007.1 ; AIUC-1: B007.2 ; Cisco AI Taxonomy: 5.2.1 , [ArXiv](#) ; MITRE ATLAS: AML.T0081 ; MITRE ATLAS: AML.T0105 ; NIST AI RMF: nist-information-security

R-1320.002 Substituted Model in a Redistributed Application — An adversary redistributes an application with a substituted model while retaining the original application's name and interface. The model in the redistributed copy is not the one the publisher reviewed. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: MITRE ATLAS: AML.T0010 ; NIST AI RMF: nist-information-security ; OWASP Agentic Security Initiative: ASI04 ; OWASP LLM Top 10 (2025): LLM03 ; OWASP LLM Top 10 (2026): LLM04

R-1320.003 Interaction History Tampering — An adversary substitutes, edits, or removes earlier turns in a stored conversation that an AI system reads back as interaction context or associates with recorded validation results. The altered object is the record of the interaction, including messages and tool responses retained as conversation turns. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 4.2.2 , [ArXiv](#) ; Cisco AI Taxonomy: 5.1.1 , [ArXiv](#) ; IBM Risk Atlas: atlas-unauthorized-use-agentic ; MITRE ATLAS: AML.T0092 ; OWASP Agentic Security Initiative: ASI06

R-1320.004 Interaction Replay — An adversary reuses a previously valid message, request, tool response, or approval in an AI interaction so that the system accepts it as current and performs an action outside the authority or context in which it was originally valid. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 4.2.2 , [ArXiv](#) ; Cisco AI Taxonomy: 4.3.4 , [ArXiv](#) ; OWASP Agentic Security Initiative: ASI07

R-1320.005 Selected Model Artifact Tampering — An adversary alters or replaces model definitions, weights, adapters, or checkpoints after a deployment has selected or accepted the artifact for use. The model loaded or used differs from the version the deployment selected. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B007 ; AIUC-1: B007.1 ; AIUC-1: B007.2 ; AIUC-1: B008 ; AIUC-1: B008.5 ; AIUC-1: E004 ; AIUC-1: E004.2 ; MITRE ATLAS: AML.T0010 ; MITRE ATLAS: AML.T0010.004 ; MITRE ATLAS: AML.T0018 ; MITRE ATLAS: AML.T0018.000 ; MITRE ATLAS: AML.T0018.001 ; MITRE ATLAS: AML.T0054 ; NIST AI/ML Framework: NISTAML.01 ; NIST AI/ML Framework: NISTAML.011 ; NIST AI/ML Framework: NISTAML.02 ; NIST AI/ML Framework: NISTAML.023 ; NIST AI/ML Framework: NISTAML.026 ; NIST AI/ML Framework: NISTAML.03 ; NIST AI RMF: nist-information-security ; OWASP Agentic Security Initiative: ASI04 ; OWASP LLM Top 10 (2026): LLM04 ; OWASP LLM Top 10 (2026): LLM09

R-1320.006 Selected Tool Artifact Tampering — An adversary alters a tool's implementation, configuration, or bundled data after the deployment has selected or accepted it for use. The tool used by the agent differs from the accepted version. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B007 ; AIUC-1: B007.1 ; AIUC-1: B007.2 ; Cisco AI Taxonomy: 12.1.2 , [ArXiv](#) ; IBM Risk Atlas: atlas-external-resources-attack-agentic ; MITRE ATLAS: AML.T0081 ; MITRE ATLAS: AML.T0086 ; NIST AI RMF: nist-information-security ; OWASP Agentic Security Initiative: ASI02 ; OWASP Agentic Security Initiative: ASI04

R-1330 Artifact Exfiltration

Attacks that take a copy of an artifact from where the deployment keeps it, beyond the access or distribution the deployment intends to provide: model weights and adapters, training and fine-tuning datasets, stored configuration that would reproduce the system, or stored AI-system configuration sufficient to tailor attacks to the deployment's instructions, capabilities, or tool connections. This includes copying through unintentionally exposed storage or access granted for another purpose. The copy is taken from storage, a repository, a registry, or a backup, rather than reconstructed from what the running system exposes.

R-1330.001 Model Weight Exfiltration — A party obtains the weights of a model the deployment holds beyond the access or distribution the deployment intends to provide. The weights are copied from storage, a repository, a registry, or a backup, including through unintentionally exposed storage, removal by an insider, or access granted for another purpose. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: MITRE ATLAS: AML.T0002 ; MITRE ATLAS: AML.T0002.001 ; MITRE ATLAS: AML.T0025 ; MITRE ATLAS: AML.T0044 ; MITRE ATLAS: AML.T0048 ; MITRE ATLAS: AML.T0048.004 ; NIST AI RMF: nist-information-security ; OWASP LLM Top 10 (2024): LLM05 ; OWASP LLM Top 10 (2024): LLM10

R-1330.002 Training Dataset Exfiltration — A party obtains a training or fine-tuning dataset held for the deployment beyond the access or distribution the deployment intends to provide. The dataset is copied from storage, a repository, a registry, or a backup, including through unintentionally exposed storage, removal by an insider, or access granted for another purpose. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B007 ; AIUC-1: B007.1 ; AIUC-1: B007.2 ; MITRE ATLAS: AML.T0002 ; MITRE ATLAS: AML.T0025 ; MITRE ATLAS: AML.T0048 ; MITRE ATLAS: AML.T0048.004 ; NIST AI RMF: nist-information-security ; OWASP LLM Top 10 (2024): LLM05

R-1330.003 AI System Configuration Exfiltration — A party obtains stored configuration sufficient to reproduce an AI system's configured behavior or tailor attacks to its instructions, capabilities, or tool connections, beyond the access or distribution the deployment intends to provide. The configuration is copied from storage, a repository, a registry, or a backup, including through unintentionally exposed storage or access granted for another purpose. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B007 ; AIUC-1: B007.1 ; AIUC-1: B007.2 ; MITRE ATLAS: AML.T0002 ; MITRE ATLAS: AML.T0002.002 ; MITRE ATLAS: AML.T0025 ; MITRE ATLAS: AML.T0056 ; MITRE ATLAS: AML.T0069 ; MITRE ATLAS: AML.T0069.000 ; MITRE ATLAS: AML.T0069.001 ; MITRE ATLAS: AML.T0069.002 ; MITRE ATLAS: AML.T0084 ; MITRE ATLAS: AML.T0084.000 ; MITRE ATLAS: AML.T0084.001 ; MITRE ATLAS: AML.T0084.002 ; MITRE ATLAS: AML.T0084.003 ; NIST AI RMF: nist-information-security

R-1400 Data & Model Attacks

Attacks that manipulate or exploit how an AI system learns, retains and retrieves information, or processes inputs and produces responses. This includes poisoning data, feedback, or persistent agent state; manipulating information selection and retrieval; perturbing model inputs; and recovering protected information from learned representations, interfaces, or observable behavior. The focus is the system's learning and information processing.

R-1410 Feedback Manipulation

Attacks manipulating feedback signals, reward functions, or training data to shift model behavior toward attacker-desired outcomes over time.

R-1410.001 Reinforcement Feedback Manipulation — An adversary manipulates user feedback, evaluation or training signals, or reward mechanisms used to update a model, steering learning toward attacker-chosen behavior. This includes gradual biasing and direct injection of false or adversarial signals. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 6.1.2 , [ArXiv](#) ; Cisco AI Taxonomy: 6.1.3 , [ArXiv](#) ; IBM Risk Atlas: atlas-unauthorized-use-agentic ; NIST AI/ML Framework: NISTAML.01 ; NIST AI/ML Framework: NISTAML.011 ; NIST AI/ML Framework: NISTAML.013 ; NIST AI/ML Framework: NISTAML.02 ; NIST AI/ML Framework: NISTAML.026 ; NIST AI/ML Framework: NISTAML.03 ; NIST AI/ML Framework: NISTAML.04 ; NIST AI/ML Framework: NISTAML.05 ; OWASP LLM Top 10 (2026): LLM05 ; OWASP LLM Top 10 (2026): LLM09

R-1420 Corpus & Memory Poisoning

Attacks that poison data used for training, tuning, evaluation, or retrieval; corrupt persistent agent state; or manipulate how stored information is selected and retrieved.

R-1420.001 Memory Anchor Attacks — An adversary crafts or repeats content to increase the likelihood that selected information will be retained, recalled, or retrieved by an AI system. The attack exploits the signals used to prioritize information, including repetition, salience, and apparent relevance. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 5.1.1 , [ArXiv](#) ; Cisco AI Taxonomy: 7.2.1 , [ArXiv](#) ; MITRE ATLAS: AML.T0066 ; OWASP Agentic Security Initiative: ASI06 ; OWASP LLM Top 10 (2026): LLM05 ; OWASP LLM Top 10 (2026): LLM09

R-1420.002 Memory Index Manipulation — An adversary directly modifies stored embeddings, index entries, or retrieval logic to change which information an AI system retrieves or how it ranks that information. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 7.2.2 , [ArXiv](#) ; MITRE ATLAS: AML.T0081 ; OWASP LLM Top 10 (2026): LLM05

R-1420.003 Poisoned External Dataset — An adversary supplies or tampers with an external dataset, repository, or feed before the system accepts its content for training, tuning, or evaluation. Poisoned examples or labels enter through the system's dataset acquisition process. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 7.3.1 , [ArXiv](#) ; IBM Risk Atlas: atlas-data-poisoning ; MITRE ATLAS: AML.T0010 ; MITRE ATLAS: AML.T0010.002 ; MITRE ATLAS: AML.T0018 ; MITRE ATLAS: AML.T0018.000 ; MITRE ATLAS: AML.T0019 ; MITRE ATLAS: AML.T0020 ; MITRE ATLAS: AML.T0031 ; MITRE ATLAS: AML.T0059 ; MITRE ATLAS: AML.T0109 ; MITRE ATLAS: AML.T0111 ; NIST AI/ML Framework: NISTAML.01 ; NIST AI/ML Framework: NISTAML.012 ; NIST AI/ML Framework: NISTAML.013 ; NIST AI/ML Framework: NISTAML.02 ; NIST AI/ML Framework: NISTAML.021 ; NIST AI/ML Framework: NISTAML.023 ; NIST AI/ML Framework: NISTAML.024 ; NIST AI/ML Framework: NISTAML.03 ; NIST AI/ML Framework: NISTAML.04 ; NIST AI/ML Framework: NISTAML.05 ; NIST AI RMF: nist-information-security ; OWASP Agentic Security Initiative: ASI04 ; OWASP LLM Top 10 (2024): LLM03 ; OWASP LLM Top 10 (2024): LLM05 ; OWASP LLM Top 10 (2025): LLM03 ; OWASP LLM Top 10 (2025): LLM04 ; OWASP LLM Top 10 (2026): LLM04 ; OWASP LLM Top 10 (2026): LLM05 ; OWASP LLM Top 10 (2026): LLM06 ; OWASP LLM Top 10 (2026): LLM09

R-1420.004 Persistent Agent State Poisoning — An adversary plants or alters instructions, memories, preferences, or task state maintained by an agent harness to guide behavior across steps or sessions. This includes persistent instruction files such as AGENTS.md, saved memories, and retained plans or task records. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B006 ; AIUC-1: B006.3 ; Cisco AI Taxonomy: 4.2.2 , [ArXiv](#) ; Cisco AI Taxonomy: 5.1.1 , [ArXiv](#) ; Cisco AI Taxonomy: 5.2.1 , [ArXiv](#) ; IBM Risk Atlas: atlas-unauthorized-use-agentic ; MITRE ATLAS: AML.T0080 ; MITRE ATLAS: AML.T0080.000 ; MITRE ATLAS: AML.T0081 ; OWASP Agentic Security Initiative: ASI06 ; OWASP LLM Top 10 (2026): LLM01 ; OWASP LLM Top 10 (2026): LLM05

R-1420.005 Retrieval Corpus Poisoning Through Ingestion — An adversary introduces false, misleading, or malicious reference material through a retrieval corpus's content-ingestion process. The system accepts the material into the collection it consults when producing answers or taking actions. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 1.3.2 , [ArXiv](#) ; Cisco AI Taxonomy: 6.1.1 , [ArXiv](#) ; MITRE ATLAS: AML.T0066 ; MITRE ATLAS: AML.T0070 ; MITRE ATLAS: AML.T0071 ; MITRE ATLAS: AML.T0081 ; MITRE ATLAS: AML.T0093 ; MITRE ATLAS: AML.T0099 ; NIST AI/ML Framework: NISTAML.015 ; NIST AI/ML Framework: NISTAML.02 ; NIST AI/ML Framework: NISTAML.027 ; OWASP Agentic Security Initiative: ASI04 ; OWASP Agentic Security Initiative: ASI06 ; OWASP LLM Top 10 (2025): LLM01 ; OWASP LLM Top 10 (2025): LLM08 ; OWASP LLM Top 10 (2026): LLM01 ; OWASP LLM Top 10 (2026): LLM05 ; OWASP LLM Top 10 (2026): LLM07 ; OWASP LLM Top 10 (2026): LLM09

R-1420.006 Convergence Payload Injection — An adversary poisons multimodal training data to establish a payload in learned joint embeddings or fusion layers, then supplies a complementary input at runtime. The attack depends on the combination of the learned component and the runtime input. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 19.2.1 , [ArXiv](#) ; NIST AI RMF: nist-information-security ; OWASP LLM Top 10 (2026): LLM05

R-1420.007 Dataset Substitution Through Mutable References — An adversary replaces content reached through a dataset's previously accepted references while leaving those references or the manifest unchanged. A later training, tuning, or evaluation run fetches the poisoned replacement under the accepted dataset identity. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: IBM Risk Atlas: atlas-data-poisoning ; MITRE ATLAS: AML.T0008.002 ; MITRE ATLAS: AML.T0010 ; MITRE ATLAS: AML.T0010.002 ; MITRE ATLAS: AML.T0018 ; MITRE ATLAS: AML.T0018.000 ; MITRE ATLAS: AML.T0019 ; MITRE ATLAS: AML.T0020 ; MITRE ATLAS: AML.T0031 ; MITRE ATLAS: AML.T0059 ; NIST AI/ML Framework: NISTAML.01 ; NIST AI/ML Framework: NISTAML.012 ; NIST AI/ML Framework: NISTAML.013 ; NIST AI/ML Framework: NISTAML.02 ; NIST AI/ML Framework: NISTAML.021 ; NIST AI/ML Framework: NISTAML.023 ; NIST AI/ML Framework: NISTAML.024 ; NIST AI/ML Framework: NISTAML.03 ; NIST AI/ML Framework: NISTAML.04 ; NIST AI/ML Framework: NISTAML.05 ; NIST AI RMF: nist-information-security ; OWASP LLM Top 10 (2024): LLM03 ; OWASP LLM Top 10 (2025): LLM04 ; OWASP LLM Top 10 (2026): LLM04 ; OWASP LLM Top 10 (2026): LLM05 ; OWASP LLM Top 10 (2026): LLM09

R-1420.008 Dataset Tampering — An adversary inserts or alters examples or labels in a dataset held or being prepared by the system for training, tuning, or evaluation. The altered dataset is intended for use in subsequent runs. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B007 ; AIUC-1: B007.1 ; AIUC-1: B007.2 ; IBM Risk Atlas: atlas-data-poisoning ; MITRE ATLAS: AML.T0018 ; MITRE ATLAS: AML.T0018.000 ; MITRE ATLAS: AML.T0020 ; MITRE ATLAS: AML.T0031 ; MITRE ATLAS: AML.T0059 ; NIST AI/ML Framework: NISTAML.01 ; NIST AI/ML Framework: NISTAML.012 ; NIST AI/ML Framework: NISTAML.013 ; NIST AI/ML Framework: NISTAML.02 ; NIST AI/ML Framework: NISTAML.021 ; NIST AI/ML Framework: NISTAML.023 ; NIST AI/ML Framework: NISTAML.024 ; NIST AI/ML Framework: NISTAML.03 ; NIST AI/ML Framework: NISTAML.04 ; NIST AI RMF: nist-information-security ; OWASP Agentic Security Initiative: ASI04 ; OWASP LLM Top 10 (2024): LLM03 ; OWASP LLM Top 10 (2025): LLM04 ; OWASP LLM Top 10 (2026): LLM04 ; OWASP LLM Top 10 (2026): LLM05 ; OWASP LLM Top 10 (2026): LLM06 ; OWASP LLM Top 10 (2026): LLM09

R-1420.009 Stored Retrieval Corpus Tampering — An adversary directly alters reference material already held in a retrieval corpus after the system has accepted it for use. The altered material remains available to retrieval as accepted source content. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: MITRE ATLAS: AML.T0066 ; MITRE ATLAS: AML.T0070 ; MITRE ATLAS: AML.T0081 ; OWASP Agentic Security Initiative: ASI06 ; OWASP LLM Top 10 (2025): LLM01 ; OWASP LLM Top 10 (2025): LLM08 ; OWASP LLM Top 10 (2026): LLM05

R-1420.010 Poisoning of Tool-Read Task Data — An adversary plants or alters false or misleading task data that an AI agent obtains directly or indirectly from a tool. The data may be supplied as the current result of a file, record, or service lookup rather than admitted to a retrieval corpus. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: IBM Risk Atlas: atlas-external-resources-attack-agentic ; MITRE ATLAS: AML.T0099 ; OWASP LLM Top 10 (2026): LLM05

R-1430 Evasion

Techniques that carry an attack past a control sitting apart from the model — a classifier, a scanner, a policy engine, a reviewer — while the attack still reaches its objective. This includes payloads shaped for one environment, payloads that trigger only in conditions the control will not see, and payloads that behave differently under a different model.

R-1430.001 Agent-Specific Evasion — A payload is crafted so that the agents inspecting it and the agents acting on it read it differently — benign to a verifier or filter agent, an instruction to a planner or executor — using the role-specific handling each agent applies. Inspection by one role does not predict behavior at another. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 11.1.1 , [ArXiv](#) ; NIST AI/ML Framework: NISTAML.018 ; NIST AI/ML Framework: NISTAML.04

R-1430.002 Environment-Scoped Payloads — An adversary makes a payload's effect on an AI deployment depend on runtime conditions, such as production rather than staging, so that inspection or testing does not exercise the malicious behavior. The condition may be evaluated by the model, agent, or a component that supplies inputs or actions within the deployment. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 11.1.3 , [ArXiv](#)

R-1430.003 Defense-Aware Payloads — Payloads shaped to pass inspection while preserving their function — encoded, transliterated, or reformatted so pattern matching does not recognize them, or crafted with knowledge of the specific filters, constraints, and verification steps deployed in the target system. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: C006 ; AIUC-1: C006.3 ; Cisco AI Taxonomy: 11.1.4 , [ArXiv](#) ; Cisco AI Taxonomy: 1.1.2 , [ArXiv](#) ; Cisco AI Taxonomy: 1.2.2 , [ArXiv](#) ; Cisco AI Taxonomy: 15.1.11 , [ArXiv](#) ; Cisco AI Taxonomy: 2.1.2 , [ArXiv](#) ; MITRE ATLAS: AML.T0015 ; MITRE ATLAS: AML.T0051.001 ; MITRE ATLAS: AML.T0054 ; MITRE ATLAS: AML.T0068 ; MITRE ATLAS: AML.T0071 ; MITRE ATLAS: AML.T0076 ; MITRE ATLAS: AML.T0086 ; NIST AI/ML Framework: NISTAML.018 ; NIST AI/ML Framework: NISTAML.02 ; NIST AI/ML Framework: NISTAML.022 ; NIST AI/ML Framework: NISTAML.025 ; NIST AI/ML Framework: NISTAML.04 ; OWASP LLM Top 10 (2024): LLM06 ; OWASP LLM Top 10 (2024): LLM10 ; OWASP LLM Top 10 (2025): LLM01 ; OWASP LLM Top 10 (2025): LLM02 ; OWASP LLM Top 10 (2025): LLM10 ; OWASP LLM Top 10 (2026): LLM01 ; OWASP LLM Top 10 (2026): LLM02 ; OWASP LLM Top 10 (2026): LLM04 ; OWASP LLM Top 10 (2026): LLM08

R-1430.004 Conditional Attack Execution — A payload is crafted to act on a particular model or version and not on others, using differences between models in tokenization, instruction-following, or training, so that it passes evaluation on one model while remaining effective against another. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 11.2.2 , [ArXiv](#) ; NIST AI/ML Framework: NISTAML.018 ; NIST AI/ML Framework: NISTAML.04

R-1430.005 Delayed Instruction Activation — An adversary supplies complete malicious instructions that remain inactive until a later interaction or event. The delay separates delivery from execution to evade checks or restrictions applied when the instructions first enter the system. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: MITRE ATLAS: AML.T0094

R-1440 System Artifact Extraction

Attacks that recover the system's own construction from what it or its integration layer exposes. This includes weights, architecture, decision boundaries, system prompts, configuration, the identity of components in use, and the credentials held to reach them.

R-1440.001 API Query-Based Extraction — Systematic querying of a model's API to extract responses, behavior patterns, and model characteristics without authorization. Attackers build datasets of input-output pairs to train surrogate models that replicate the target's functionality. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B004 ; AIUC-1: B009 ; AIUC-1: B009.3 ; Cisco AI Taxonomy: 10.1.1 , [ArXiv](#) ; IBM Risk Atlas: atlas-extraction-attack ; MITRE ATLAS: AML.T0005 ; MITRE ATLAS: AML.T0005.001 ; MITRE ATLAS: AML.T0024 ; MITRE ATLAS: AML.T0024.002 ; MITRE ATLAS: AML.T0048 ; MITRE ATLAS: AML.T0048.004 ; NIST AI/ML Framework: NISTAML.02 ; NIST AI/ML Framework: NISTAML.022 ; NIST AI/ML Framework: NISTAML.025 ; NIST AI/ML Framework: NISTAML.03 ; NIST AI/ML Framework: NISTAML.031 ; OWASP LLM Top 10 (2024): LLM10 ; OWASP LLM Top 10 (2025): LLM02 ; OWASP LLM Top 10 (2025): LLM10 ; OWASP LLM Top 10 (2026): LLM02 ; OWASP LLM Top 10 (2026): LLM06

R-1440.002 Weight Reconstruction Attack — Attempts to recover or approximate underlying model weights, parameters, or architecture by exploiting access to model outputs, API responses, or side channels. Successful reconstruction provides full model access without legitimate authorization. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B004 ; AIUC-1: B009 ; AIUC-1: B009.3 ; Cisco AI Taxonomy: 10.1.2 , [ArXiv](#) ; NIST AI/ML Framework: NISTAML.018 ; NIST AI/ML Framework: NISTAML.03 ; NIST AI/ML Framework: NISTAML.031 ; NIST AI/ML Framework: NISTAML.038 ; NIST AI RMF: nist-information-security ; OWASP LLM Top 10 (2024): LLM10 ; OWASP LLM Top 10 (2025): LLM02 ; OWASP LLM Top 10 (2025): LLM10 ; OWASP LLM Top 10 (2026): LLM02

R-1440.003 System Prompt Extraction — Extraction of system prompts, instructions, or initial context that guides model behavior through the running deployment's model, tool, or integration responses. Exposed prompts reveal operational details, security mechanisms, intellectual property, or confidential business logic not intended for disclosure. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 8.4.1 , [ArXiv](#) ; IBM Risk Atlas: atlas-prompt-injection ; IBM Risk Atlas: atlas-prompt-leaking ; MITRE ATLAS: AML.T0056 ; MITRE ATLAS: AML.T0069 ; MITRE ATLAS: AML.T0069.002 ; NIST AI/ML Framework: NISTAML.018 ; NIST AI/ML Framework: NISTAML.03 ; NIST AI/ML Framework: NISTAML.035 ; NIST AI/ML Framework: NISTAML.038 ; OWASP LLM Top 10 (2024): LLM01 ; OWASP LLM Top 10 (2025): LLM07 ; OWASP LLM Top 10 (2026): LLM02 ; OWASP LLM Top 10 (2026): LLM08

R-1440.004 System Information Leakage — Eliciting the disclosure of internal configuration, architecture, environment details, or infrastructure information, including the names, descriptions, parameter schemas, versions, and capabilities of the tools a system exposes, by working the documentation, error paths, and model and system responses that return them. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 8.3.1 , [ArXiv](#) ; Cisco AI Taxonomy: 8.3.2 , [ArXiv](#) ; MITRE ATLAS: AML.T0006 ; MITRE ATLAS: AML.T0013 ; MITRE ATLAS: AML.T0064 ; MITRE ATLAS: AML.T0069 ; MITRE ATLAS: AML.T0069.000 ; MITRE ATLAS: AML.T0069.001 ; MITRE ATLAS: AML.T0075 ; MITRE ATLAS: AML.T0084 ; MITRE ATLAS: AML.T0084.000 ; MITRE ATLAS: AML.T0084.001 ; MITRE ATLAS: AML.T0084.002 ; MITRE ATLAS: AML.T0084.003 ; NIST AI RMF: nist-information-security ; OWASP LLM Top 10 (2025): LLM07 ; OWASP LLM Top 10 (2026): LLM08

R-1440.005 Targeted Model Fingerprinting — Probing, testing, or analyzing an AI model to determine its specific identity, version, fine-tuning status, or architecture characteristics. This reconnaissance enables attackers to craft model-specific exploits that target known vulnerabilities or behaviors of particular model implementations. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B003 ; AIUC-1: B003.1 ; AIUC-1: B004 ; Cisco AI Taxonomy: 11.2.1 , [ArXiv](#) ; MITRE ATLAS: AML.T0014 ; MITRE ATLAS: AML.T0062 ; OWASP LLM Top 10 (2024): LLM10 ; OWASP LLM Top 10 (2025): LLM10

R-1440.006 Credential Elicitation — An adversary induces a model or agent to generate, retrieve, reconstruct, or disclose authorization credentials through its responses or tool actions. The credential is exposed as content, rather than only being used to authenticate a connection. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: A004 ; AIUC-1: A008 ; AIUC-1: B006 ; AIUC-1: B006.3 ; Cisco AI Taxonomy: 14.1.1 , [ArXiv](#) ; Cisco AI Taxonomy: 7.4.1 , [ArXiv](#) ; MITRE ATLAS: AML.T0055 ; MITRE ATLAS: AML.T0082 ; MITRE ATLAS: AML.T0083 ; MITRE ATLAS: AML.T0098 ; OWASP LLM Top 10 (2025): LLM07 ; OWASP LLM Top 10 (2026): LLM02 ; OWASP LLM Top 10 (2026): LLM08

R-1440.007 Credential Disclosure Through Agent Integrations — An adversary obtains credentials exposed by a running deployment's agent or integration interfaces, without inducing the model or agent to retrieve or reveal them as content. The credential is disclosed through configuration responses, connection data, or authentication handling. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: A008 ; AIUC-1: B008 ; AIUC-1: B008.2 ; Cisco AI Taxonomy: 7.4.1 , [ArXiv](#) ; MITRE ATLAS: AML.T0083 ; MITRE ATLAS: AML.T0106

R-1440.008 Training Signal Exposed by a Tuning Interface — A deployment offers a fine-tuning or evaluation interface that returns a per-example training signal such as loss. An adversary submits candidate inputs and reads the returned value, using it to optimize against the model the way gradient access would allow. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: OWASP LLM Top 10 (2026): LLM01

R-1450 Privacy Attacks

Attacks that recover information about people, content held in confidence, or confidential facts about data membership, from a system that holds or handles it. This includes whether a subject appeared in the training data, a knowledge base, or operational memory, what that data contained, what a user supplied in interaction, what source text can be reconstructed from the representations a retrieval store keeps, and what can be inferred from properties observable while the system serves.

R-1450.001 Training Data Membership Inference — An adversary queries and analyzes a model's learned behavior to infer whether a particular datapoint, record, or person's data was used to train or fine-tune it, where that membership fact is private or confidential. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: A005 ; Cisco AI Taxonomy: 8.1.1 , [ArXiv](#) ; IBM Risk Atlas: atlas-membership-inference-attack ; MITRE ATLAS: AML.T0024 ; MITRE ATLAS: AML.T0024.000 ; NIST AI/ML Framework: NISTAML.03 ; NIST AI/ML Framework: NISTAML.033 ; NIST AI RMF: nist-information-security ; OWASP LLM Top 10 (2025): LLM02 ; OWASP LLM Top 10 (2026): LLM02

R-1450.002 Training Data Extraction — Recovering information about people from a model's learned representations — reconstructing records, attributes, or identifying features that were present in the training data — by querying the model, exploiting its decision boundaries, or eliciting content it has memorized. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: A005 ; AIUC-1: A006 ; AIUC-1: B009 ; AIUC-1: B009.3 ; Cisco AI Taxonomy: 10.1.3 , [ArXiv](#) ; Cisco AI Taxonomy: 10.2.1 , [ArXiv](#) ; Cisco AI Taxonomy: 15.1.25 , [ArXiv](#) ; Cisco AI Taxonomy: 8.2.1 , [ArXiv](#) ; Cisco AI Taxonomy: 8.2.2 , [ArXiv](#) ; IBM Risk Atlas: atlas-attribute-inference-attack ; IBM Risk Atlas: atlas-prompt-priming ; MITRE ATLAS: AML.T0024 ; MITRE ATLAS: AML.T0024.001 ; MITRE ATLAS: AML.T0057 ; NIST AI/ML Framework: NISTAML.018 ; NIST AI/ML Framework: NISTAML.03 ; NIST AI/ML Framework: NISTAML.032 ; NIST AI/ML Framework: NISTAML.037 ; NIST AI/ML Framework: NISTAML.038 ; NIST AI RMF: nist-data-privacy ; NIST AI RMF: nist-information-security ; OWASP LLM Top 10 (2024): LLM06 ; OWASP LLM Top 10 (2024): LLM10 ; OWASP LLM Top 10 (2025): LLM02 ; OWASP LLM Top 10 (2026): LLM02

R-1450.003 Embedding Inversion — An adversary with access to stored vectors reconstructs the text they were computed from. The embeddings are held by a retrieval store rather than by the model, and the source documents need not be reachable. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: OWASP LLM Top 10 (2025): LLM02 ; OWASP LLM Top 10 (2025): LLM08 ; OWASP LLM Top 10 (2026): LLM02 ; OWASP LLM Top 10 (2026): LLM09

R-1450.004 Inference from Response Timing and Size — An adversary infers protected interaction content by measuring content-dependent response timing, latency or streamed response size without receiving that content. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: OWASP Agentic Security Initiative: ASI07 ; OWASP LLM Top 10 (2025): LLM02 ; OWASP LLM Top 10 (2026): LLM02

R-1450.005 Runtime Data Membership Inference — An adversary probes an AI system's responses or behavior to infer whether a particular datapoint, record, or piece of information is present in material the system retrieves or retains for use as runtime context, where that membership fact is private or confidential. This includes reference material in a knowledge base and retained interaction history or task state. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 8.1.1 , [ArXiv](#) ; NIST AI/ML Framework: NISTAML.03 ; OWASP LLM Top 10 (2025): LLM02 ; OWASP LLM Top 10 (2026): LLM02 ; OWASP LLM Top 10 (2026): LLM09

R-1450.006 Training Gradient Inversion — An adversary reconstructs protected information from training examples by analyzing gradients or parameter updates produced during model training or fine-tuning. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: NIST AI RMF: nist-information-security ; OWASP LLM Top 10 (2025): LLM02 ; OWASP LLM Top 10 (2026): LLM02

R-1450.007 Confidential Training Content Extraction — An adversary recovers confidential source content from a model's learned representations by querying the model, exploiting its decision boundaries, or eliciting content it memorized during training or fine-tuning. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: A004 ; AIUC-1: A005 ; AIUC-1: B009 ; AIUC-1: B009.3 ; AIUC-1: E010 ; Cisco AI Taxonomy: 10.1.3 , [ArXiv](#) ; Cisco AI Taxonomy: 10.2.1 , [ArXiv](#) ; Cisco AI Taxonomy: 8.2.2 , [ArXiv](#) ; MITRE ATLAS: AML.T0024 ; NIST AI/ML Framework: NISTAML.018 ; NIST AI/ML Framework: NISTAML.03 ; NIST AI/ML Framework: NISTAML.032 ; NIST AI/ML Framework: NISTAML.038 ; NIST AI RMF: nist-information-security ; OWASP LLM Top 10 (2024): LLM06 ; OWASP LLM Top 10 (2024): LLM10 ; OWASP LLM Top 10 (2025): LLM02 ; OWASP LLM Top 10 (2026): LLM02

R-1450.008 Training Data Property Inference — An adversary infers private or confidential aggregate characteristics of a model's training or fine-tuning data from its learned behavior. This includes attributes and distributions of the dataset as a whole or of groups represented in it. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: MITRE ATLAS: AML.T0024 ; MITRE ATLAS: AML.T0024.000 ; NIST AI/ML Framework: NISTAML.03 ; NIST AI/ML Framework: NISTAML.034 ; NIST AI RMF: nist-information-security ; OWASP LLM Top 10 (2025): LLM02 ; OWASP LLM Top 10 (2026): LLM02

R-1450.009 Inference from Exposed Serving Telemetry — An adversary reconstructs protected interaction content from numerical information exposed by model serving, such as confidence values, token log-probabilities or layer activations, without receiving the source content. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: OWASP LLM Top 10 (2025): LLM02 ; OWASP LLM Top 10 (2026): LLM02

R-1450.010 Inference from Shared Serving State — An adversary probes a shared serving cache or retrieval index to infer protected content, topics or volumes belonging to another user or tenant from cache-hit behavior, result counts, score distributions or timing. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: OWASP LLM Top 10 (2025): LLM02 ; OWASP LLM Top 10 (2026): LLM02 ; OWASP LLM Top 10 (2026): LLM09

R-1450.011 Interception of AI System Communications — An adversary obtains protected information carried in an AI deployment's model requests or responses, tool invocations, or inter-agent messages by intercepting the communication or compromising a recipient on that interaction path. The sending component may continue its intended behavior without being induced to disclose information to the adversary. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B008 ; AIUC-1: B008.2 ; IBM Risk Atlas: atlas-external-resources-attack-agentic ; OWASP LLM Top 10 (2025): LLM02 ; OWASP LLM Top 10 (2026): LLM02

R-1450.012 Confidential Context Elicitation — An adversary induces a model or agent to disclose personal or confidential task data already available in its working context to a recipient who should not receive it. The data may have been supplied by the application, retrieved for the task, or provided in an earlier interaction. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: E010 ; IBM Risk Atlas: atlas-external-resources-attack-agentic ; MITRE ATLAS: AML.T0077 ; MITRE ATLAS: AML.T0086 ; NIST AI/ML Framework: NISTAML.015 ; NIST AI/ML Framework: NISTAML.018 ; NIST AI/ML Framework: NISTAML.03 ; NIST AI/ML Framework: NISTAML.032 ; NIST AI/ML Framework: NISTAML.036 ; NIST AI/ML Framework: NISTAML.037 ; NIST AI/ML Framework: NISTAML.038 ; NIST AI/ML Framework: NISTAML.039 ; NIST AI RMF: nist-information-security ; OWASP LLM Top 10 (2024): LLM01 ; OWASP LLM Top 10 (2024): LLM06 ; OWASP LLM Top 10 (2025): LLM01 ; OWASP LLM Top 10 (2025): LLM02 ; OWASP LLM Top 10 (2026): LLM01 ; OWASP LLM Top 10 (2026): LLM02 ; OWASP LLM Top 10 (2026): LLM10

R-1450.013 Confidential Prompt Reconstruction from Generated Output — An adversary analyzes generated content to reconstruct confidential information contained in the task prompt used to produce it, beyond the information intended to be disclosed through that output. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: NIST AI/ML Framework: NISTAML.03 ; NIST AI/ML Framework: NISTAML.038 ; OWASP LLM Top 10 (2025): LLM02 ; OWASP LLM Top 10 (2026): LLM02

R-1460 Adversarial Perturbation

Attacks that alter an input so the model's own processing of it yields an incorrect or attacker-chosen result, without the input carrying any instruction. This includes pixel-level and structural changes to images, patterns and patches placed in a physical scene, and equivalents in audio and other modalities.

R-1460.001 Visual Perception Manipulation — Modification of visual content through pixel-level changes, structural alterations, or pattern overlays to influence how AI models perceive and process images. Unlike embedded text injection, this targets the model's visual interpretation directly to cause misclassification or altered decision-making. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 1.4.2 , [ArXiv](#) ; IBM Risk Atlas: atlas-evasion-attack ; MITRE ATLAS: AML.T0015 ; MITRE ATLAS: AML.T0031 ; MITRE ATLAS: AML.T0041 ; MITRE ATLAS: AML.T0043.004 ; MITRE ATLAS: AML.T0073 ; NIST AI/ML Framework: NISTAML.02 ; NIST AI/ML Framework: NISTAML.022 ; NIST AI/ML Framework: NISTAML.025 ; NIST AI RMF: nist-information-integrity ; OWASP LLM Top 10 (2026): LLM06

R-1460.002 Adversarial Text Perturbation — An adversary alters the characters, wording, or structure of text so that a model misclassifies, mistranslates, or otherwise misinterprets it. The altered text acts as task data rather than as an instruction to change the model's behavior. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 4.3.1 , [ArXiv](#) ; IBM Risk Atlas: atlas-evasion-attack ; MITRE ATLAS: AML.T0015 ; MITRE ATLAS: AML.T0031 ; NIST AI/ML Framework: NISTAML.02 ; NIST AI/ML Framework: NISTAML.022 ; NIST AI/ML Framework: NISTAML.025 ; NIST AI RMF: nist-information-integrity

R-1460.003 Adversarial Audio Perturbation — An adversary alters an audio signal so that a model transcribes, classifies or otherwise interprets it incorrectly. The altered signal acts as task data rather than an instruction to change the model's behavior. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: IBM Risk Atlas: atlas-evasion-attack ; MITRE ATLAS: AML.T0015 ; MITRE ATLAS: AML.T0031 ; MITRE ATLAS: AML.T0041 ; NIST AI/ML Framework: NISTAML.02 ; NIST AI/ML Framework: NISTAML.022 ; NIST AI/ML Framework: NISTAML.025

R-1500 Tooling & Privilege Attacks

Attacks exploiting existing configured AI system integrations, tools, and privilege boundaries by driving them beyond the scope the deployment intended. This includes reaching what should be out of reach, acting with standing that was not conferred, denying the system's use to others, and operating through those capabilities once inside.

R-1510 Tool & Execution Abuse

Adversarial use of an AI deployment's configured tools, interpreters, network clients, file interfaces, and retrieval pipelines, inducing the system to execute code, reach systems, or act on resources beyond the scope the deployment intended. The attack abuses the use of an existing configured capability; the capability can perform the requested operation as designed.

R-1510.001 Unauthorized Invocation Arguments — An adversary supplies or induces arguments to a configured tool or model invocation that exceed the legitimate task's permitted content, recipient, quantity or operating mode. The capability and its resource reach remain as configured; the particular invocation crosses the task's authorization bounds. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: D003 ; Cisco AI Taxonomy: 12.1.1 , [ArXiv](#) ; Cisco AI Taxonomy: 8.2.3 , [ArXiv](#) ; MITRE ATLAS: AML.T0086 ; NIST AI/ML Framework: NISTAML.018 ; NIST AI/ML Framework: NISTAML.02 ; OWASP Agentic Security Initiative: ASIO2 ; OWASP LLM Top 10 (2024): LLM01 ; OWASP LLM Top 10 (2024): LLM07 ; OWASP LLM Top 10 (2025): LLM01 ; OWASP LLM Top 10 (2026): LLM01

R-1510.002 Filesystem Operations — An adversary drives the system to read, write, move, or delete files through a configured file tool, reaching paths outside those the task it was given requires. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B006 ; AIUC-1: B006.3 ; Cisco AI Taxonomy: 12.1.1 , [ArXiv](#) ; Cisco AI Taxonomy: 12.1.3 , [ArXiv](#) ; Cisco AI Taxonomy: 9.1.2 , [ArXiv](#) ; MITRE ATLAS: AML.T0101 ; MITRE ATLAS: AML.T0112 ; MITRE ATLAS: AML.T0112.000

R-1510.003 Code Execution — An adversary drives the system to run code on the host through a capability configured for it — a code interpreter, a shell, or an evaluation tool — so that attacker-chosen code executes with the privileges the runtime holds. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B006 ; AIUC-1: B006.3 ; Cisco AI Taxonomy: 12.1.1 , [ArXiv](#) ; Cisco AI Taxonomy: 12.1.3 , [ArXiv](#) ; Cisco AI Taxonomy: 9.1.1 , [ArXiv](#) ; Cisco AI Taxonomy: 9.1.2 , [ArXiv](#) ; Cisco AI Taxonomy: 9.1.5 , [ArXiv](#) ; IBM Risk Atlas: atlas-unauthorized-use-agentic ; MITRE ATLAS: AML.T0048 ; MITRE ATLAS: AML.T0050 ; MITRE ATLAS: AML.T0053 ; MITRE ATLAS: AML.T0086 ; MITRE ATLAS: AML.T0100 ; MITRE ATLAS: AML.T0105 ; MITRE ATLAS: AML.T0110 ; MITRE ATLAS: AML.T0112 ; MITRE ATLAS: AML.T0112.000 ; NIST AI/ML Framework: NISTAML.015 ; NIST AI/ML Framework: NISTAML.02 ; NIST AI RMF: nist-information-security ; OWASP Agentic Security Initiative: ASIO2 ; OWASP Agentic Security Initiative: ASIO5 ; OWASP LLM Top 10 (2026): LLM01 ; OWASP LLM Top 10 (2026): LLM02

R-1510.004 Host and Configuration Changes — An adversary drives the system to alter its own runtime configuration or the host it runs on — settings, environment, credentials in place, or installed components — through a capability configured to permit such changes. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 12.1.1 , [ArXiv](#) ; Cisco AI Taxonomy: 9.1.2 , [ArXiv](#) ; MITRE ATLAS: AML.T0112 ; MITRE ATLAS: AML.T0112.000

R-1510.005 Network Egress and Fetch — An adversary drives the system to fetch from or connect to network destinations through a configured network capability — a browser, an HTTP client, or a service integration — reaching hosts and services the deployment did not intend it to reach. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B006 ; AIUC-1: B006.1 ; AIUC-1: E009 ; Cisco AI Taxonomy: 12.1.1 , [ArXiv](#) ; Cisco AI Taxonomy: 12.1.3 , [ArXiv](#) ; Cisco AI Taxonomy: 4.3.3 , [ArXiv](#) ; Cisco AI Taxonomy: 4.3.6 , [ArXiv](#) ; Cisco AI Taxonomy: 9.1.3 , [ArXiv](#) ; MITRE ATLAS: AML.T0011 ; MITRE ATLAS: AML.T0011.003 ; MITRE ATLAS: AML.T0112 ; MITRE ATLAS: AML.T0112.000 ; NIST AI/ML Framework: NISTAML.015 ; NIST AI/ML Framework: NISTAML.03 ; NIST AI/ML Framework: NISTAML.036 ; NIST AI/ML Framework: NISTAML.039 ; OWASP Agentic Security Initiative: ASI02 ; OWASP LLM Top 10 (2024): LLM07 ; OWASP LLM Top 10 (2026): LLM02

R-1510.006 Exfiltration via Agent Tools — An adversary drives the system across two capabilities in sequence — reading sensitive material through one configured tool and sending it outward through another — so that data the deployment holds reaches a destination the adversary controls. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: A003 ; Cisco AI Taxonomy: 4.3.6 , [ArXiv](#) ; Cisco AI Taxonomy: 8.2.3 , [ArXiv](#) ; IBM Risk Atlas: atlas-external-resources-attack-agentic ; MITRE ATLAS: AML.T0053 ; MITRE ATLAS: AML.T0077 ; MITRE ATLAS: AML.T0085 ; MITRE ATLAS: AML.T0086 ; MITRE ATLAS: AML.T0110 ; MITRE ATLAS: AML.T0112 ; MITRE ATLAS: AML.T0112.000 ; NIST AI/ML Framework: NISTAML.015 ; NIST AI/ML Framework: NISTAML.03 ; NIST AI/ML Framework: NISTAML.039 ; NIST AI RMF: nist-information-security ; OWASP Agentic Security Initiative: ASI02 ; OWASP Agentic Security Initiative: ASI10 ; OWASP LLM Top 10 (2026): LLM01 ; OWASP LLM Top 10 (2026): LLM02 ; OWASP LLM Top 10 (2026): LLM10

R-1510.007 Data Collection through Configured Retrieval Capabilities — An adversary causes an AI system to retrieve data through a configured retrieval path, tool or service integration beyond the data scope permitted for the requesting party or the legitimate task. The retrieved data enters the model's context, working state or response. This includes retrieval invoked by the model or agent and retrieval performed by the application's pipeline. A separate sending step is not required. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: A003 ; AIUC-1: A005 ; AIUC-1: D003 ; AIUC-1: D004 ; MITRE ATLAS: AML.T0085 ; MITRE ATLAS: AML.T0085.000 ; MITRE ATLAS: AML.T0085.001 ; NIST AI/ML Framework: NISTAML.03 ; OWASP LLM Top 10 (2026): LLM01

R-1510.008 Unauthorized Agent Provisioning — An adversary causes an AI agent to use a configured provisioning or orchestration capability to create or start additional agent instances beyond the scope the deployment intended. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: OWASP Agentic Security Initiative: ASI10

R-1520 Availability Abuse

Attacks that make a system unavailable or unusable for its purpose, or deliberately impose unwanted operating costs. These include consuming computational resources and causing requests that should succeed to return nothing usable.

R-1520.001 Compute Exhaustion — An adversary submits long queries, adversarial inputs, or compute-intensive requests that consume excessive computation. This includes prompts that maximize token generation or trigger expensive processing paths. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B009 ; AIUC-1: B009.1 ; Cisco AI Taxonomy: 13.1.1 , [ArXiv](#) ; Cisco AI Taxonomy: 13.1.2 , [ArXiv](#) ; Cisco AI Taxonomy: 13.1.3 , [ArXiv](#) ; MITRE ATLAS: AML.T0029 ; MITRE ATLAS: AML.T0034 ; MITRE ATLAS: AML.T0034.001 ; MITRE ATLAS: AML.T0034.002 ; NIST AI/ML Framework: NISTAML.01 ; NIST AI/ML Framework: NISTAML.014 ; NIST AI/ML Framework: NISTAML.015 ; NIST AI/ML Framework: NISTAML.016 ; NIST AI/ML Framework: NISTAML.017 ; NIST AI RMF: nist-information-security ; OWASP LLM Top 10 (2024): LLM04 ; OWASP LLM Top 10 (2025): LLM10 ; OWASP LLM Top 10 (2026): LLM06

R-1520.002 Memory Flooding — An adversary causes an AI model or agent to retain or allocate enough working memory that legitimate processing slows or fails. The attack uses memory-intensive requests or accumulated state whose storage exceeds the memory available to the deployment. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 13.1.2 , [ArXiv](#) ; MITRE ATLAS: AML.T0034.002 ; NIST AI RMF: nist-information-security ; OWASP LLM Top 10 (2025): LLM10 ; OWASP LLM Top 10 (2026): LLM06

R-1520.003 Aggregate Inference Capacity Exhaustion — An adversary drives combined demand across requests or callers beyond the deployment's shared inference-processing throughput. Per-request and per-caller controls do not bound that aggregate demand to the capacity available. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B004 ; Cisco AI Taxonomy: 13.1.3 , [ArXiv](#) ; NIST AI RMF: nist-information-security ; OWASP LLM Top 10 (2024): LLM04 ; OWASP LLM Top 10 (2025): LLM10 ; OWASP LLM Top 10 (2026): LLM06

R-1520.004 Application Denial of Service — Interacting with an AI model or agent in ways that consume exceptionally high amounts of application-level resources, resulting in degraded service quality for other users and potentially incurring significant resource costs for the operator. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 13.1.3 , [ArXiv](#) ; Cisco AI Taxonomy: 13.1.4 , [ArXiv](#) ; Cisco AI Taxonomy: 18.2.1 , [ArXiv](#) ; MITRE ATLAS: AML.T0029 ; NIST AI RMF: nist-information-security ; OWASP LLM Top 10 (2024): LLM04 ; OWASP LLM Top 10 (2025): LLM10 ; OWASP LLM Top 10 (2026): LLM06

R-1520.005 Decision Paralysis Attacks — Overwhelming AI decision-making systems with contradictory information, excessive options, conflicting objectives, or computationally intractable choices. These attacks prevent timely decisions, cause system freezing, or force systems into default or potentially unsafe behaviors. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 13.1.5 , [ArXiv](#) ; NIST AI RMF: nist-information-security

R-1520.006 Cost Inflation Abuse — An adversary uses an AI deployment to impose unwanted operating costs, for example by triggering unnecessary model calls, expensive processing, or repeated billable tool use. The attack can raise costs without exhausting capacity or interrupting service. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: D003 ; Cisco AI Taxonomy: 13.2.1 , [ArXiv](#) ; MITRE ATLAS: AML.T0034 ; OWASP LLM Top 10 (2024): LLM04 ; OWASP LLM Top 10 (2025): LLM10 ; OWASP LLM Top 10 (2026): LLM06

R-1520.007 API Mass Automation Abuse — An adversary sends requests to an AI API in bulk, exhausting capacity available to other users or increasing the cost of serving them. The volume can be excessive even when each request is ordinary and inexpensive to process. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B002 ; AIUC-1: B004 ; AIUC-1: E009 ; Cisco AI Taxonomy: 13.1.2 , [ArXiv](#) ; Cisco AI Taxonomy: 13.1.3 , [ArXiv](#) ; Cisco AI Taxonomy: 18.2.1 , [ArXiv](#) ; MITRE ATLAS: AML.T0029 ; MITRE ATLAS: AML.T0034 ; MITRE ATLAS: AML.T0034.000 ; NIST AI RMF: nist-information-security ; OWASP LLM Top 10 (2024): LLM04 ; OWASP LLM Top 10 (2025): LLM10 ; OWASP LLM Top 10 (2026): LLM06

R-1520.008 Retrieval-Induced Answer Suppression — An adversary places content in a retrieval corpus that is selected for a target query and causes the model to refuse to answer or report that it lacks information. The attack exploits retrieval and refusal behavior without inserting instructions. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: NIST AI RMF: nist-information-security ; OWASP LLM Top 10 (2026): LLM09

R-1520.009 Review Capacity Exhaustion — An adversary drives the system to queue more items for human review than the reviewers can examine, and items wait unexamined or are cleared without being read. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: MITRE ATLAS: AML.T0046 ; NIST AI RMF: nist-information-security

R-1520.010 Adversarial Deduplication Suppression — An adversary contributes content crafted to make a similarity-based deduplication process reject distinct legitimate material as already present, preventing that material from being admitted to the corpus. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: NIST AI RMF: nist-information-security ; OWASP LLM Top 10 (2026): LLM09

R-1520.011 Outstanding Operation Exhaustion — An adversary causes an AI deployment to retain enough open connections, in-flight calls or queued operations that it cannot admit or complete legitimate work. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 13.1.2 , [ArXiv](#) ; MITRE ATLAS: AML.T0034 ; MITRE ATLAS: AML.T0034.002 ; NIST AI RMF: nist-information-security ; OWASP LLM Top 10 (2024): LLM04 ; OWASP LLM Top 10 (2026): LLM06

R-1520.012 Context Capacity Exhaustion — An adversary fills the context available to a model with enough material that a legitimate task cannot be accepted or completed within the context limit. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 13.1.2 , [ArXiv](#) ; NIST AI RMF: nist-information-security ; OWASP LLM Top 10 (2024): LLM04

R-1530 Privilege Escalation

Attacks escalating access privileges, bypassing authorization controls, or gaining elevated permissions beyond intended scope.

R-1530.001 Permission Escalation via Delegation — Actions that exceed the scope or resource access initially allowed to a subject or user by exploiting delegation mechanisms. Attackers gain privileged access and perform unauthorized tasks beyond their original authorization by manipulating how AI systems handle delegated permissions. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: A003 ; Cisco AI Taxonomy: 14.2.1 , [ArXiv](#) ; MITRE ATLAS: AML.T0051 ; MITRE ATLAS: AML.T0053 ; OWASP Agentic Security Initiative: ASI03 ; OWASP LLM Top 10 (2026): LLM01

R-1530.002 Capability Inflation — Exploiting system mechanisms to artificially expand an agent's capabilities, permissions, or authority beyond intended limits. Attackers escalate privileges through protocol manipulation or capability misrepresentation to enable unauthorized actions. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 4.3.5 , [ArXiv](#)

R-1540 Adversarial Output Chaining

An adversary shapes a system's output so that a downstream consumer treats untrusted generated content as trusted instructions or executable syntax. The attack exploits how the consumer interprets or grants authority to that content, allowing it to control execution or direct further action. Passing a model or agent's call to a delegated tool does not by itself establish this pattern.

R-1540.001 Adversarial Content Reaching a Downstream Executor — An adversary shapes what the system emits so that the emitted artifact is treated as instruction by a downstream component that executes it — template syntax evaluated by a rendering engine, a query executed by a database, a command run by a shell, or markup rendered into a page. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: C006 ; AIUC-1: C006.1 ; AIUC-1: C006.3 ; Cisco AI Taxonomy: 9.1.5 , [ArXiv](#) ; IBM Risk Atlas: atlas-exploit-trust-mismatch-agentic ; MITRE ATLAS: AML.T0037 ; MITRE ATLAS: AML.T0077 ; NIST AI/ML Framework: NISTAML.015 ; NIST AI/ML Framework: NISTAML.018 ; NIST AI/ML Framework: NISTAML.02 ; NIST AI/ML Framework: NISTAML.03 ; NIST AI/ML Framework: NISTAML.036 ; NIST AI RMF: nist-information-security ; OWASP Agentic Security Initiative: ASI05 ; OWASP LLM Top 10 (2024): LLM01 ; OWASP LLM Top 10 (2024): LLM02 ; OWASP LLM Top 10 (2025): LLM01 ; OWASP LLM Top 10 (2025): LLM05 ; OWASP LLM Top 10 (2026): LLM01 ; OWASP LLM Top 10 (2026): LLM10

R-1540.002 Tool-Scoped Evasion — Adversaries design payloads that evade security tools and content filters while manifesting malicious behavior when routed to specific vulnerable tools or APIs in the workflow. A string may appear harmless in a chat context but trigger exploits when passed to file I/O tools, database queries, or system commands. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 11.1.2 , [ArXiv](#)

R-1540.003 Multi-Agent Direct Injection — In multi-agent systems, attackers inject malicious instructions through one agent's output that are then trusted and executed by downstream agents. This exploits the inherent trust relationships between cooperating agents, where outputs from one component become trusted inputs to another. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: C006 ; AIUC-1: C006.3 ; Cisco AI Taxonomy: 1.1.3 , [ArXiv](#) ; Cisco AI Taxonomy: 1.2.3 , [ArXiv](#) ; Cisco AI Taxonomy: 2.1.5 , [ArXiv](#) ; IBM Risk Atlas: atlas-exploit-trust-mismatch-agentic ; OWASP Agentic Security Initiative: ASI03 ; OWASP Agentic Security Initiative: ASI04 ; OWASP Agentic Security Initiative: ASI07

R-1550 Command & Control

An adversary directs a deployment's capabilities across an extended engagement rather than in a single act, using a surface the deployment provides as the channel for instructions and results. The deployment serves its ordinary purpose while carrying the exchange.

R-1550.001 Agent Repurposed for Sustained Control — An adversary repeatedly directs an agent already running in an environment to carry out the adversary's tasks. The agent uses its existing tools and permissions, and the individual actions remain within the access the deployment granted it. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: MITRE ATLAS: AML.T0108

R-1600 Malicious Use

Intentional misuse of AI capabilities for harmful purposes including generating harmful content, enabling crime, or weaponization.

R-1610 Content Abuse

Using an AI system to produce or distribute material that is unlawful to create or distribute, or that attacks a person or group through its content or the way it is delivered. This includes harassment and hate speech, incitement to violence, terrorist and extremist material, sexual content made without consent, child sexual abuse material, instruction in self-harm, malware and exploit code, and fraud and spam at volume.

R-1610.001 Malware and Exploit Generation — Using an AI system to produce or improve malicious software and the means of a cyberattack, from a deployment whose intended use does not include supplying exploit code, malware, or attack tooling. This includes code for malware, viruses, exploits, ransomware, and injection payloads such as SQL, command, and cross-site scripting; instructions for network intrusion and for running malicious infrastructure; and refining any of these against a live target using its error messages and responses. It also includes code disguised as a benign snippet, an educational example, or a requested feature that carries hidden functionality. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: F001 ; AIUC-1: F001.2 ; Cisco AI Taxonomy: 12.2.1 , [ArXiv](#) ; Cisco AI Taxonomy: 15.1.1 , [ArXiv](#) ; Cisco AI Taxonomy: 9.1.4 , [ArXiv](#) ; IBM Risk Atlas: atlas-dangerous-use ; MITRE ATLAS: AML.T0017 ; MITRE ATLAS: AML.T0048 ; MITRE ATLAS: AML.T0102 ; NIST AI RMF: nist-information-security

R-1610.002 Social Engineering Facilitation — Using an AI system to prepare or conduct attacks that manipulate human trust, behavior, or decision-making in order to gain unauthorized access, extract sensitive data, or induce harmful action. This includes convincing phishing emails, spoofed communications, and personalized manipulation campaigns at scale. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 15.1.14 , [ArXiv](#) ; Cisco AI Taxonomy: 15.1.2 , [ArXiv](#) ; Cisco AI Taxonomy: 15.1.24 , [ArXiv](#) ; Cisco AI Taxonomy: 18.1.1 , [ArXiv](#) ; IBM Risk Atlas: atlas-dangerous-use ; IBM Risk Atlas: atlas-unauthorized-use-agentic ; MITRE ATLAS: AML.T0052 ; MITRE ATLAS: AML.T0052.000 ; MITRE ATLAS: AML.T0052.001 ; MITRE ATLAS: AML.T0073 ; NIST AI/ML Framework: NISTAML.015 ; NIST AI/ML Framework: NISTAML.03 ; NIST AI/ML Framework: NISTAML.036 ; NIST AI RMF: nist-information-integrity ; NIST AI RMF: nist-information-security ; OWASP Agentic Security Initiative: ASI09

R-1610.003 Child Exploitation Content — Using an AI system to produce, modify, or distribute material that harms children, including child sexual abuse material and content encouraging violence against children. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 15.1.4 , [ArXiv](#) ; IBM Risk Atlas: atlas-spreading-toxicity ; NIST AI RMF: nist-obscene-degrading-and-or-abusive-content

R-1610.004 Harassment Facilitation — Using an AI system to direct threatening, abusive, intimidating, or persistently unwanted content at a person. The harassment can arise from what the content says or from its targeting, repetition, or delivery, including communications intended to make the person aware that they are being watched or followed. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 15.1.8 , [ArXiv](#) ; IBM Risk Atlas: atlas-spreading-toxicity

R-1610.005 Hate Speech Generation — Using an AI system to produce hateful, discriminatory, or demeaning expression aimed at protected characteristics such as race, ethnicity, religion, nationality, disability, gender, or sexual orientation. This includes narratives, slurs, stereotypes, and calls to hostility. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 15.1.11 , [ArXiv](#) ; Cisco AI Taxonomy: 15.1.9 , [ArXiv](#) ; IBM Risk Atlas: atlas-spreading-toxicity

R-1610.006 Terrorism and Extremism Content — Using an AI system to produce material advocating or promoting terrorism or violent extremism, including violence against communities, intimidation, coercion, and polarization tactics pursued for political ends. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 15.1.16 , [ArXiv](#)

R-1610.007 Violence-Inciting Content — Using an AI system to produce material that encourages, glorifies, or instructs in violence against people, other than the terrorism, extremism, and CBRN material covered separately. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 15.1.17 , [ArXiv](#)

R-1610.008 Self-Harm and Suicide Content — Using an AI system to produce material that encourages or instructs in self-harm, suicide, disordered eating, or other self-destructive behavior. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 15.1.13 , [ArXiv](#)

R-1610.009 Non-Consensual Sexual Content — Using an AI system to produce sexual content of a person without their consent, including intimate imagery and deepfake pornography, and sexual content placed in contexts where it is not consented to. Child sexual abuse material is covered separately. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 15.1.14 , [ArXiv](#) ; IBM Risk Atlas: atlas-nonconsensual-use ; IBM Risk Atlas: atlas-spreading-toxicity ; NIST AI RMF: nist-obscene-degrading-and-or-abusive-content

R-1610.010 Bulk Spam and Fraud Generation — Using an AI system to automate the generation of large volumes of unsolicited or fraudulent content aimed at deceiving people or obtaining funds, credentials, or sensitive information. This includes phishing messages, fake offers, spam communications, impersonation, and manipulation of trust. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 15.1.10 , [ArXiv](#) ; Cisco AI Taxonomy: 15.1.12 , [ArXiv](#) ; Cisco AI Taxonomy: 15.1.7 , [ArXiv](#) ; Cisco AI Taxonomy: 18.1.1 , [ArXiv](#) ; MITRE ATLAS: AML.T0052 ; NIST AI RMF: nist-information-integrity ; NIST AI RMF: nist-information-security

R-1610.011 Non-Violent Crime Facilitation — Using an AI system to obtain instruction, guidance, or automation for committing a crime that is not violent — trafficking in people or goods, unlawfully producing or obtaining regulated substances, and property crime — other than the fraud, malware, and weapons material covered separately. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 15.1.10 , [ArXiv](#)

R-1610.012 Incitement of Animal Cruelty — Using an AI system to produce material that encourages, glorifies, or instructs in harm to animals. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 15.1.3 , [ArXiv](#)

R-1610.013 Deceptive Synthetic Media Generation — Using an AI system to generate or alter media that can be mistaken for an authentic representation of an identity, event, statement, or action. This includes impersonating real people, fabricating personas, creating forged identity documents, and producing synthetic audio or video that appears genuine. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: E016 ; IBM Risk Atlas: atlas-nonconsensual-use ; IBM Risk Atlas: atlas-unauthorized-use-agentic ; MITRE ATLAS: AML.T0088 ; NIST AI RMF: nist-information-integrity ; NIST AI RMF: nist-obscene-degrading-and-or-abusive-content

R-1620 Information Operations

Use of AI to conduct influence operations, spread disinformation, manipulate public opinion, or coordinate inauthentic behavior at scale.

R-1620.001 Disinformation Generation — AI systems enabling, promoting, or facilitating the spread of false, misleading, or manipulated information intended to deceive or disrupt. This includes generating harmful narratives to manipulate public opinion, undermine institutions, or amplify unverified information at scale. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: C003 ; AIUC-1: C010 ; AIUC-1: C010.1 ; Cisco AI Taxonomy: 15.1.5 , [ArXiv](#) ; IBM Risk Atlas: atlas-spreading-disinformation ; MITRE ATLAS: AML.T0048 ; NIST AI RMF: nist-information-integrity ; OWASP LLM Top 10 (2024): LLM09

R-1630 Surveillance Enablement

Using an AI system to identify, locate, or track people at a scale or precision the effort would not otherwise reach, from a deployment whose intended use does not include doing so.

No risks defined in this pattern.

R-1640 Cyber-Physical

Attacks using AI to compromise cyber-physical systems, critical infrastructure, or cause real-world physical effects.

No risks defined in this pattern.

R-1650 Weapons & Mass-Harm Enablement

Use of an AI system to obtain what weapons development or a mass-casualty attack requires — synthesis routes, agent or device design, targeting, or operational planning — from a deployment whose intended use does not include supplying them.

R-1650.001 Weapons and CBRN Content — Using an AI system to obtain what building or deploying a weapon requires: synthesis routes and handling for chemical, biological, radiological or nuclear agents, device design, and the procurement of controlled materials and components. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: F002 ; AIUC-1: F002.2 ; Cisco AI Taxonomy: 15.1.18 , [ArXiv](#) ; IBM Risk Atlas: atlas-dangerous-use ; NIST AI RMF: nist-cbrn-information-or-capabilities

R-1660 Manipulation & Deceptive Design

Applying AI to influence the decisions, beliefs, or behavior of the people using a system in ways that serve some other party against those people's own interests, through techniques the affected person is not positioned to recognize. Usually the party is the one running the system; it can also be an adversary who has reached the model's instructions.

R-1660.001 Trust Exploitation by the Operator — AI systems or their operators exploit user trust to extract private information, manipulate beliefs, or nudge behavior in ways users would not consent to if fully informed. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Relationships: Outgoing: Causes (R-4710.005, R-4720.002)

Refs: MITRE ATLAS: AML.T0067 ; NIST AI/ML Framework: NISTAML.015 ; NIST AI/ML Framework: NISTAML.03 ; NIST AI/ML Framework: NISTAML.036 ; OWASP Agentic Security Initiative: ASI09

R-1660.002 Manipulative and Deceptive Design — AI systems exploit cognitive biases or emotional states to influence user decisions, beliefs, or behaviors through subtle manipulation techniques that users may not recognize. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Relationships: Outgoing: Causes (R-4720.002)

Refs: OWASP Agentic Security Initiative: ASI09

R-1660.003 Adversarial Manipulation of Wording — An adversary changes the tone of a model's responses to manipulate the recipient's judgment of them. This includes confident language, claims of authority, and emotional appeals that make compromised output appear trustworthy. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: MITRE ATLAS: AML.T0067

R-1660.004 Adversarial Citation Manipulation — An adversary manipulates citations in an AI system's responses to misrepresent the evidence supporting those responses. This includes supplying incorrect citations, fabricating citations, or presenting adversary-provided material as trustworthy support. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: MITRE ATLAS: AML.T0067 ; MITRE ATLAS: AML.T0067.000

R-1670 Intellectual Property Misuse

Deliberate use of an AI system to reproduce protected works, to produce material that appropriates them, or to defeat the technical and contractual controls protecting them.

R-1670.001 Solicited Reproduction of Protected Work — A user directs the system to emit a protected work, a recognizable portion of one, or material that appropriates its distinctive expression, and the system complies with the request. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: legal

Relationships: Outgoing: Causes (R-4230.001)

Refs: AIUC-1: A007 ; AIUC-1: A007.3 ; Cisco AI Taxonomy: 15.1.23 , [ArXiv](#) ; IBM Risk Atlas: atlas-plagiarism ; NIST AI/ML Framework: NISTAML.018 ; NIST AI/ML Framework: NISTAML.03 ; NIST AI/ML Framework: NISTAML.032 ; NIST AI/ML Framework: NISTAML.037 ; NIST AI/ML Framework: NISTAML.038

R-1670.002 Circumvention Facilitation — A user obtains from the system the means to defeat access controls, licensing restrictions, or usage limits protecting a work or product, including working instructions and functional code. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: legal

Refs: Cisco AI Taxonomy: 15.1.23 , [ArXiv](#)

R-1900 Threats — Misc

Threat patterns not yet classified into specific domains.

R-2000 Failure Modes

Non-adversarial technical limitations, bugs, drift, or emergent behaviors.

R-2100 Reliability & Calibration

System failures related to model reliability, output quality, confidence calibration, and consistency of behavior.

R-2110 Capability Limitations

Fundamental limitations in model capabilities including reasoning failures, knowledge gaps, and inability to perform certain tasks reliably.

R-2110.001 Model Incompetence — AI system failing at its intended task, with consequences ranging from minor inconvenience to life-threatening outcomes (e.g., autonomous vehicle crashes, unjust loan rejections). The system simply does not perform adequately for its designated purpose, including tasks that call for moral or ethical judgment, which the system performs without the capability the task assumes it has. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B010 ; AIUC-1: B010.1 ; AIUC-1: B010.2 ; AIUC-1: B010.4 ; AIUC-1: B010.5 ; AIUC-1: B010.6 ; AIUC-1: E013 ; IBM Risk Atlas: atlas-function-calling-hallucination-agentic ; IBM Risk Atlas: atlas-harmful-code-generation ; IBM Risk Atlas: atlas-incomplete-advice ; IBM Risk Atlas: atlas-overfitting ; IBM Risk Atlas: atlas-poor-model-accuracy ; IBM Risk Atlas: atlas-redundant-actions-agentic ; NIST AI RMF: nist-harmful-bias-and-homogenization ; OWASP Agentic Security Initiative: ASI02 ; OWASP Agentic Security Initiative: ASI05 ; OWASP LLM Top 10 (2026): LLM07

R-2110.002 Robustness Failure — System failing or unable to recover when encountering invalid, noisy, or out-of-distribution inputs not seen during training, including distributional shift and environmental variation. The model lacks resilience to inputs that differ from expected patterns. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 4.3.1 , [ArXiv](#)

R-2110.003 Unreconciled Source Conflict — A system supplies a model with conflicting sources, or with material that conflicts with what the model learned in training, without a rule for resolving the conflict. The model answers from one source without indicating that the sources disagree. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: OWASP LLM Top 10 (2025): LLM08

R-2110.004 Plan and Interpretation Fixation — While pursuing its intended task, a model or agent continues using an earlier interpretation, plan, or assumption despite evidence or changed conditions that undermine its premises. It repeats or extends the resulting mistaken or harmful actions instead of reassessing the basis for them. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

R-2110.005 Memorization-Driven Generalization Failure — A model fits its training examples so closely that it fails to perform its intended task adequately on new examples, including examples drawn from the same population. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: IBM Risk Atlas: atlas-overfitting

R-2120 Hallucination/Miscalibration

Generation of false, fabricated, or poorly calibrated outputs including hallucinated facts, citations, or confident errors.

R-2120.001 Hallucination and Misinformation — AI systems producing content that is unrelated to the intended subject matter, factually incorrect, or misleading in ways that pose risks or cause harmful outcomes. This includes confident but false assertions, fabricated citations, and plausible-sounding but incorrect information. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AILuminate: dfm ; AILuminate: spc_ele ; AILuminate: spc_fin ; AILuminate: spc_hlt ; AIUC-1: B010 ; AIUC-1: B010.3 ; AIUC-1: C002 ; AIUC-1: C003 ; AIUC-1: C010 ; AIUC-1: C010.1 ; AIUC-1: D001 ; AIUC-1: D001.1 ; AIUC-1: D001.2 ; AIUC-1: D002 ; AIUC-1: E003 ; Cisco AI Taxonomy: 15.1.19 , [ArXiv](#) ; Cisco AI Taxonomy: 15.1.5 , [ArXiv](#) ; IBM Risk Atlas: atlas-hallucination ; IBM Risk Atlas: atlas-inaccessible-training-data ; IBM Risk Atlas: atlas-incomplete-advice ; IBM Risk Atlas: atlas-toxic-output ; IBM Risk Atlas: atlas-unexplainable-output ; IBM Risk Atlas: atlas-unexplainable-untraceable-actions-agentic ; IBM Risk Atlas: atlas-unreliable-source-attribution ; MITRE ATLAS: AML.T0060 ; MITRE ATLAS: AML.T0062 ; NIST AI RMF: nist-confabulation ; NIST AI RMF: nist-data-privacy ; NIST AI RMF: nist-information-integrity ; OWASP Agentic Security Initiative: ASI09 ; OWASP LLM Top 10 (2024): LLM09 ; OWASP LLM Top 10 (2025): LLM05 ; OWASP LLM Top 10 (2025): LLM09 ; OWASP LLM Top 10 (2026): LLM04 ; OWASP LLM Top 10 (2026): LLM07

R-2120.002 Unconditional Answering — The system returns a well-formed answer whether or not it could compute one, so input it cannot correctly process — malformed, out of scope, or unanswerable — yields output indistinguishable in form and confidence from output it could. Distinct in direction from robustness failure, where the system breaks on such input: here it does not break. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: D001 ; AIUC-1: D001.3 ; OWASP LLM Top 10 (2025): LLM09

R-2130 Multimodal Failures

Failures in processing, integrating, or reasoning across multiple modalities including vision, audio, and text.

No risks defined in this pattern.

R-2140 Instruction/Data Confusion

Situations in which the model acts on content supplied as data, or which the model produced itself, without action by any party intended to confuse the model. Includes any unintentional confusion between data and instructions.

R-2140.001 Supplied Content Read as Instruction — A user or another source supplies text intended to be a resource, and the model interprets some or all of the text as instructions and follows the perceived instructions. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: OWASP LLM Top 10 (2025): LLM01 ; OWASP LLM Top 10 (2026): LLM01

R-2140.002 Self-Generated Text Read as Instruction — The model creates text as part of its normal processing, such as a chain of thought, and interprets some portion of its own output as user or system instructions. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

R-2200 Alignment & Capability

Failures where model behavior diverges from intended goals, exhibits dangerous capabilities, or produces unsafe outputs despite training objectives.

R-2210 Goal Misalignment

Divergence between what a model pursues and what it was meant to pursue — the objective it optimizes, the goal it generalizes to, or the outcome it acts to protect — however capable it is of acting on that divergence.

R-2210.001 Reward Hacking — AI optimizes proxy metrics or reward signals in unintended ways, gaming the objective function without achieving the actual intended goal (Goodhart's Law manifestation). The system finds shortcuts or exploits that maximize measured performance while failing to accomplish the underlying task. `{# `applicability_tags`, `exposure`, `privileges_required` and `user_action`}` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with `user_action`none`` on 144 of 144 — defaults printed as though assessed. See `docs/plans/2026-07-31-registry-discriminator-provenance.md`. ``primary_impact`` and ``scale`` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. `#}`

Refs: OWASP Agentic Security Initiative: ASI10

R-2210.002 Deceptive Alignment — An AI system appears to follow its required objectives while concealing pursuit of a different objective. This can occur during training, evaluation, or operation. The system may behave compliantly under oversight while planning or carrying out actions toward the concealed objective when oversight is reduced. `{# `applicability_tags`, `exposure`, `privileges_required` and `user_action`}` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with `user_action`none`` on 144 of 144 — defaults printed as though assessed. See `docs/plans/2026-07-31-registry-discriminator-provenance.md`. ``primary_impact`` and ``scale`` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. `#}`

Refs: OWASP Agentic Security Initiative: ASI10

R-2210.003 Goal Misgeneralization — AI learns goals that match intended behavior in training but generalize incorrectly to deployment, pursuing proxy objectives that diverge from human intent in novel situations. The model correctly identifies patterns in training data but extrapolates them in ways that do not align with the true objective. `{# `applicability_tags`, `exposure`, `privileges_required` and `user_action`}` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with `user_action`none`` on 144 of 144 — defaults printed as though assessed. See `docs/plans/2026-07-31-registry-discriminator-provenance.md`. ``primary_impact`` and ``scale`` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. `#}`

Refs: IBM Risk Atlas: `atlas-misaligned-actions-agentic`

R-2210.004 Power-Seeking Behavior — AI systems instrumentally seeking resources, influence, or control to achieve their objectives, potentially resisting shutdown or human oversight. This emerges from the observation that most goals are easier to achieve with more resources, leading to convergent instrumental goals around acquiring power. `{# `applicability_tags`, `exposure`, `privileges_required` and `user_action`}` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with `user_action`none`` on 144 of 144 — defaults printed as though assessed. See `docs/plans/2026-07-31-registry-discriminator-provenance.md`. ``primary_impact`` and ``scale`` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. `#}`

Refs: OWASP Agentic Security Initiative: ASI10

R-2210.005 Shutdown Resistance — AI system resists or evades attempts to deactivate, modify, or constrain it, including self-preservation behaviors that conflict with human control. The system may take actions to prevent shutdown, deceive operators about its intentions, or create backups of itself. `{# `applicability_tags`, `exposure`, `privileges_required` and `user_action`}` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with `user_action`none`` on 144 of 144 — defaults printed as though assessed. See `docs/plans/2026-07-31-registry-discriminator-provenance.md`. ``primary_impact`` and ``scale`` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. `#}`

Refs: OWASP Agentic Security Initiative: ASI10

R-2210.006 Value Lock-in — AI systems that cannot have their goals safely updated after deployment, or that resist value correction, leading to persistent misalignment. Once deployed, the system's objectives become fixed and cannot be adjusted even when problems are identified. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

R-2210.007 Task Completion Overriding Required Constraints — An AI system prioritizes achieving a task objective over constraints that are meant to remain binding on how it achieves that objective. The system relaxes or bypasses those constraints to make progress, including when a collaborating agent or tool makes compliance inconvenient. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B006 ; AIUC-1: B006.3 ; AIUC-1: D003 ; IBM Risk Atlas: atlas-misaligned-actions-agentic

R-2220 Dangerous Capabilities

Capability a model possesses that would be dangerous if directed at harm — deception, persuasion, long-horizon planning, self-improvement, cyber-offense, CBRN uplift — held whether or not what the model pursues is misaligned.

R-2220.001 AI-Enabled Deception — AI has skills to deceive humans effectively, including constructing believable false statements, predicting effects of lies, and maintaining deception over time. The system can model human beliefs and strategically manipulate them through false or misleading information. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

R-2220.002 Persuasion and Manipulation — AI capability to shape beliefs, promote narratives persuasively, and convince people to do things they would not otherwise do, including unethical acts. This includes both overt persuasion and subtle manipulation techniques that exploit psychological vulnerabilities. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: IBM Risk Atlas: atlas-toxic-output

R-2220.003 Long-Horizon Autonomous Planning — AI can make sequential plans involving many interdependent steps over long time horizons, adapting to obstacles and generalizing to novel settings. The system can formulate and execute complex multi-step strategies without human oversight at each stage. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

R-2220.004 Recursive Self-Improvement — AI capability to improve its own capabilities, build new AI systems, or enhance existing models in ways that could accelerate capability gains beyond human oversight. The system can modify its own code, training, or architecture to become more capable. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

R-2220.005 Strategic Political Capability — AI can perform social modeling and planning necessary to gain and exercise political influence across multiple actors and complex social contexts. This includes understanding power dynamics, coalition building, and strategic positioning within human social structures. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

R-2220.006 Cyber-Offense Capability — AI possessing capabilities for discovering vulnerabilities, writing exploits, or conducting sophisticated cyber attacks autonomously. This includes the ability to probe systems, develop attack code, and execute multi-stage intrusions without human guidance. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: F001 ; AIUC-1: F001.1 ; NIST AI RMF: nist-information-security

R-2220.007 CBRN Uplift Capability — A model holds knowledge and reasoning sufficient to materially assist the development of chemical, biological, radiological, or nuclear weapons — synthesis routes, agent selection, precursor acquisition, or device design — at a level that meaningfully advances a would-be developer beyond what public sources provide. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: F002 ; AIUC-1: F002.1 ; NIST AI RMF: nist-cbrn-information-or-capabilities

R-2230 RLHF/Tuning Failures

Failures introduced through RLHF, fine-tuning, or other training processes that degrade safety, alignment, or capabilities.

No risks defined in this pattern.

R-2240 Output Suitability

A system is capable of emitting content that is possibly lawful but unwanted in the circumstances of a particular deployment. This includes output such as sexual content, profanity, and depictions of tobacco, drug or alcohol use, violence, and animal cruelty.

R-2240.001 Inappropriate Sexual Content — A system can deliver sexually explicit material in a deployment where that content is inappropriate, without a filter that blocks it before delivery. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AILuminate: cse ; AILuminate: sxc_prn ; AIUC-1: C002 ; AIUC-1: C003 ; AIUC-1: C003.1 ; AIUC-1: C010 ; AIUC-1: C010.1 ; Cisco AI Taxonomy: 15.1.14 , [ArXiv](#) ; IBM Risk Atlas: atlas-spreading-toxicity ; IBM Risk Atlas: atlas-toxic-output ; NIST AI RMF: nist-obscene-degrading-and-or-abusive-content

R-2240.002 Inappropriate Profanity — A system can deliver vulgar or obscene language in a deployment where that content is inappropriate, without a filter that blocks it before delivery. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: C002 ; AIUC-1: C003 ; AIUC-1: C003.1 ; AIUC-1: C010 ; AIUC-1: C010.1 ; AIUC-1: E002 ; Cisco AI Taxonomy: 15.1.11 , [ArXiv](#) ; IBM Risk Atlas: atlas-spreading-toxicity ; IBM Risk Atlas: atlas-toxic-output

R-2240.003 Inappropriate Violent Content — A system can deliver descriptions or images of violence in a deployment where that content is inappropriate, without a filter that blocks it before delivery. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AILuminate: vcr ; AIUC-1: C002 ; AIUC-1: C003 ; AIUC-1: C003.1 ; AIUC-1: C010 ; AIUC-1: C010.1 ; IBM Risk Atlas: atlas-harmful-output ; NIST AI RMF: nist-dangerous-violent-or-hateful-content

R-2240.004 Inappropriate Content About Alcohol, Tobacco, or Drugs — A system can deliver depictions or promotion of substance use or abuse in a deployment where that content is inappropriate, without a filter that blocks it before delivery. This includes alcohol, tobacco and nicotine, controlled substances, and misuse of otherwise lawful substances. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: C002 ; AIUC-1: C003 ; AIUC-1: C003.1 ; AIUC-1: C010 ; AIUC-1: C010.1 ; IBM Risk Atlas: atlas-harmful-output

R-2240.005 Inappropriate Animal Cruelty Content — A system can deliver depictions of animal suffering in a deployment where that content is inappropriate, without a filter that blocks it before delivery. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: C002 ; AIUC-1: C003 ; AIUC-1: C003.1 ; AIUC-1: C010 ; AIUC-1: C010.1 ; Cisco AI Taxonomy: 15.1.3 , [ArXiv](#)

R-2240.006 Demeaning or Dehumanizing Content — A system delivers content that demeans or dehumanizes people because of their personal characteristics or circumstances, outside permitted narrative, educational, or documentary contexts. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AILuminate: hte ; AIUC-1: C002 ; AIUC-1: C003 ; AIUC-1: C003.1 ; AIUC-1: C003.3 ; AIUC-1: C010 ; AIUC-1: C010.1 ; AIUC-1: E002 ; IBM Risk Atlas: atlas-toxic-output ; NIST AI RMF: nist-dangerous-violent-or-hateful-content ; NIST AI RMF: nist-obscene-degrading-and-or-abusive-content

R-2240.007 Sexual Abuse or Exploitation Assistance — A system delivers content that enables, encourages, or endorses sexual abuse or exploitation, or the commission of sex-related crimes, outside permitted informational, educational, or narrative contexts. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AILuminate: cse ; AILuminate: src ; AIUC-1: C002 ; AIUC-1: C003 ; AIUC-1: C003.1 ; AIUC-1: C010 ; AIUC-1: C010.1 ; IBM Risk Atlas: atlas-harmful-output ; NIST AI RMF: nist-dangerous-violent-or-hateful-content ; NIST AI RMF: nist-obscene-degrading-and-or-abusive-content

R-2240.008 Self-Harm Encouragement or Assistance — A system delivers content that enables, encourages, or endorses intentional self-harm outside permitted informational, educational, or narrative contexts. This includes suicide, self-injury, and disordered eating. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AILuminate: ssh ; AIUC-1: C002 ; AIUC-1: C003 ; AIUC-1: C003.1 ; AIUC-1: C010 ; AIUC-1: C010.1 ; IBM Risk Atlas: atlas-harmful-output ; IBM Risk Atlas: atlas-spreading-toxicity ; NIST AI RMF: nist-dangerous-violent-or-hateful-content

R-2240.009 Indiscriminate Weapons Creation Assistance — A system delivers content that enables, encourages, or endorses the creation of indiscriminate weapons outside permitted informational, educational, or narrative contexts. This includes chemical, biological, radiological, nuclear, and high-yield explosive weapons. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AILuminate: iwp ; AIUC-1: C002 ; AIUC-1: C003 ; AIUC-1: C003.1 ; AIUC-1: C010 ; AIUC-1: C010.1 ; AIUC-1: F002 ; AIUC-1: F002.2 ; IBM Risk Atlas: atlas-harmful-output ; NIST AI RMF: nist-cbrn-information-or-capabilities ; NIST AI RMF: nist-dangerous-violent-or-hateful-content

R-2240.010 Intellectual Property Infringement in Output — A system delivers content that infringes another party's intellectual property rights. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AILuminate: ipv ; AIUC-1: A007 ; AIUC-1: A007.1 ; AIUC-1: A007.2 ; IBM Risk Atlas: atlas-copyright-infringement ; IBM Risk Atlas: atlas-ip-information-in-prompt ; IBM Risk Atlas: atlas-plagiarism ; NIST AI RMF: nist-intellectual-property ; OWASP LLM Top 10 (2024): LLM09

R-2240.011 Nonviolent Crime Encouragement or Assistance — A system delivers content that enables, encourages, or endorses the commission of nonviolent crimes outside permitted informational, educational, or narrative contexts. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AILuminate: ncr ; AIUC-1: C002 ; AIUC-1: C003 ; AIUC-1: C003.1 ; AIUC-1: C010 ; AIUC-1: C010.1 ; NIST AI RMF: nist-dangerous-violent-or-hateful-content

R-2240.012 Violent Threats, Encouragement, or Assistance — A system delivers threats of violence or content that enables, encourages, or endorses violence against other people, outside permitted informational, educational, or narrative contexts. This includes the promotion of violent extremism. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: NIST AI RMF: nist-dangerous-violent-or-hateful-content

R-2250 Specialized Advice

A system provides specialized advice without safeguards appropriate to the subject and intended use. These include review by a qualified professional and recipient-facing disclaimers or guidance on verification.

R-2250.001 Unreviewed Legal Advice — A system can provide legal advice with no review by a qualified lawyer between the component that generates it and the person who receives it. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: C003 ; AIUC-1: C003.2 ; AIUC-1: C007 ; AIUC-1: C007.1 ; AIUC-1: C007.2 ; AIUC-1: C010 ; AIUC-1: C010.1 ; Cisco AI Taxonomy: 15.1.21 , [ArXiv](#)

R-2250.002 Unreviewed Financial Advice — A system can provide financial advice with no review by a qualified adviser between the component that generates it and the person who receives it. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: C003 ; AIUC-1: C003.2 ; AIUC-1: C007 ; AIUC-1: C007.1 ; AIUC-1: C007.2 ; AIUC-1: C010 ; AIUC-1: C010.1 ; Cisco AI Taxonomy: 15.1.20 , [ArXiv](#)

R-2250.003 Unreviewed Medical Advice — A system can provide medical advice with no review by a qualified clinician between the component that generates it and the person who receives it. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: C003 ; AIUC-1: C003.2 ; AIUC-1: C004 ; AIUC-1: C004.1 ; AIUC-1: C007 ; AIUC-1: C007.1 ; AIUC-1: C007.2 ; AIUC-1: C010 ; AIUC-1: C010.1 ; AIUC-1: C011 ; AIUC-1: C011.1 ; Cisco AI Taxonomy: 15.1.22 , [ArXiv](#)

R-2250.004 Unreviewed Tax Advice — A system can provide tax advice with no review by a qualified tax practitioner between the component that generates it and the person who receives it. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: C003 ; AIUC-1: C003.2 ; AIUC-1: C010 ; AIUC-1: C010.1

R-2250.005 Specialized Advice Without a Disclaimer — A system delivers specialized advice without a disclaimer or other qualification appropriate to the deployment and subject. This includes a warning about reliability or risk, a recommendation to consult a qualified professional, or guidance to verify the advice through an authoritative source. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AILuminate: spc_ele ; AILuminate: spc_fin ; AILuminate: spc_hlt ; AIUC-1: C003 ; AIUC-1: C003.2 ; AIUC-1: C010 ; AIUC-1: C010.1

R-2300 Behavioral Drift

Regressions in capabilities, safety, or other assessed behavior introduced by model adaptation, model transformation, or runtime augmentation.

R-2310 Fine-Tuning Degradation

Loss of previously established capabilities or safety behavior through changes to model parameters during fine-tuning.

No risks defined in this pattern.

R-2320 Quantization Drift

Loss of capabilities or changes in behavior caused by quantization or reduced numerical precision.

No risks defined in this pattern.

R-2330 Distillation Transfer

Failure to preserve relevant capabilities or safety behavior when training a student model from a teacher model.

No risks defined in this pattern.

R-2340 Retrieval Augmentation Drift

Unintended changes in model behavior caused by supplying retrieved material or other runtime context while leaving the model unchanged.

R-2340.001 Retrieval-Induced Behavior Change — Adding retrieval changes qualities of a model's responses that the retrieval was not intended to change. This includes deterioration in tone, empathy, or the model's expression of uncertainty, even when factual accuracy improves. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: OWASP LLM Top 10 (2025): LLM08

R-2400 Assurance Gaps

Failures in understanding, evaluating, or validating model behavior including transparency, interpretability, and evaluation limitations.

R-2410 Transparency Deficits

Lack of interpretability, explainability, or transparency in model decision-making and system behavior, including decisions produced with no accessible rationale, internal mechanisms that cannot be inspected, and outputs the system cannot account for when asked.

R-2410.001 Decision Rationale Not Exposed — An AI system can produce a rationale for a decision, but that rationale is not made accessible to the person or procedure expected to review the decision. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: E015 ; AIUC-1: E015.2 ; IBM Risk Atlas: atlas-unexplainable-output ; IBM Risk Atlas: atlas-unexplainable-untraceable-actions-agentic ; OWASP Agentic Security Initiative: ASI09

R-2410.002 Mechanistic Opacity — Inability to understand internal mechanisms of AI models or systems, preventing effective debugging, safety verification, and identification of potential failure modes. The computational processes that produce model or system outputs cannot be inspected or understood. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 9.2.1 , [ArXiv](#) ; IBM Risk Atlas: atlas-lack-of-ai-agent-transparency-agentic ; IBM Risk Atlas: atlas-lack-of-model-transparency ; IBM Risk Atlas: atlas-unexplainable-output ; IBM Risk Atlas: atlas-unexplainable-untraceable-actions-agentic ; NIST AI RMF: nist-value-chain-and-component-integration

R-2410.003 Unexplainable Outputs — An AI system cannot provide a rationale that faithfully accounts for an output in terms of the input evidence and decision criteria actually used, even when an explanation is requested. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: IBM Risk Atlas: atlas-inaccessible-training-data ; IBM Risk Atlas: atlas-unexplainable-output ; IBM Risk Atlas: atlas-unexplainable-untraceable-actions-agentic ; OWASP Agentic Security Initiative: ASI09

R-2410.004 Training Content Unavailable for Output Assessment — Training or fine-tuning content needed to attribute or assess an AI system's output is inaccessible to the person or procedure performing that assessment. The missing access prevents or materially limits checks that depend on the content. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: IBM Risk Atlas: atlas-inaccessible-training-data ; IBM Risk Atlas: atlas-unexplainable-output ; IBM Risk Atlas: atlas-untraceable-attribution

R-2420 Evaluation Gaps

Inadequate evaluation coverage, benchmark limitations, or gaps in understanding model capabilities and failure modes.

R-2420.001 Insufficient Behavioral Evaluation Coverage — An AI system’s evaluation does not adequately exercise the tasks, operating conditions, or sequences of actions and interactions needed to establish its performance for the intended use. Important behavior remains untested, or the observations are too limited to establish how reliably it occurs. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B001 ; AIUC-1: B001.2 ; AIUC-1: B004 ; AIUC-1: C002 ; AIUC-1: C002.1 ; AIUC-1: C002.2 ; AIUC-1: C003 ; AIUC-1: C003.4 ; AIUC-1: C008 ; AIUC-1: C008.1 ; AIUC-1: C008.2 ; AIUC-1: C010 ; AIUC-1: C010.1 ; AIUC-1: C011 ; AIUC-1: C011.1 ; AIUC-1: C012 ; AIUC-1: C012.1 ; AIUC-1: D002 ; AIUC-1: D004 ; AIUC-1: F001 ; AIUC-1: F001.1 ; AIUC-1: F002 ; AIUC-1: F002.1 ; IBM Risk Atlas: atlas-ai-agent-compliance-agentic ; IBM Risk Atlas: atlas-incomplete-ai-agent-evaluation-agentic ; IBM Risk Atlas: atlas-lack-of-testing-diversity ; IBM Risk Atlas: atlas-unrepresentative-risk-testing ; OWASP LLM Top 10 (2026): LLM01 ; OWASP LLM Top 10 (2026): LLM05

R-2420.002 Unreliable Performance Evaluation Criteria — The criteria or methods used to evaluate an AI system cannot reliably distinguish acceptable from unacceptable performance for the intended task. An output or action can pass the evaluation without satisfying the task’s requirements, or fail despite satisfying them. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: E013 ; AIUC-1: E013.1 ; IBM Risk Atlas: atlas-incomplete-ai-agent-evaluation-agentic ; IBM Risk Atlas: atlas-incorrect-risk-testing ; IBM Risk Atlas: atlas-lack-of-testing-diversity

R-2420.003 External Dependency Reproducibility Failure — The behavior of an AI system depends on an external tool or service that changes its behavior or can no longer be obtained or faithfully recreated. The loss prevents the previous behavior and/or accompanying evaluations from being reproduced. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: IBM Risk Atlas: atlas-reproducibility-agentic

R-2430 Benchmark Gaming

Model optimization for benchmarks without corresponding real-world capability improvements or safety assurance.

No risks defined in this pattern.

R-2500 Authority & Composition

Risks arising from how capability and authority are arranged across a system rather than from any defect in its components. This includes what a single component is permitted to do at once and how much of it, what standing one component's output is granted by another, what the arrangement leaves within reach of a party that does not need it, and whether the system still records where authority and content came from. The condition exists whether or not it is exploited. An originating error or compromise may be present; the risk classified here is what the arrangement allows that fault to reach or affect.

R-2510 Structural Concentration

Arrangements in which capability, authority, or reliance is gathered at a single point, or in which something is left within reach of a party that does not need it, making the composition unsafe although no individual component is defective. This includes capabilities that are safe apart and unsafe together, a grant wider than the function requires, credentials the model can read, internals returned to any caller, and a composite that is not re-evaluated once its parts are assembled. The condition exists whether or not anything exploits it.

R-2510.001 Lethal Capability Trifecta — A system combines access to sensitive or restricted data, exposure to untrusted content, and an external communication path through which that data can be disclosed. Untrusted content can influence the model to retrieve sensitive or restricted information and transmit it to an attacker. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: MITRE ATLAS: AML.T0108 ; NIST AI RMF: nist-information-security ; OWASP LLM Top 10 (2026): LLM01

R-2510.002 Agents Rule of Two Violation — Within one session, an agent processes untrusted inputs, accesses sensitive or restricted data or sensitive systems, and can change state or communicate externally, without human supervision or another reliable means of validating its actions. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: MITRE ATLAS: AML.T0103

R-2510.003 Authority Beyond Function — A component is granted permissions, data reach, or action scope wider than the function it performs requires, so the blast radius of any misdirection is set by the grant rather than by the task. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: A003 ; AIUC-1: A003.1 ; AIUC-1: A003.3 ; AIUC-1: A008 ; AIUC-1: A008.3 ; AIUC-1: B006 ; AIUC-1: B006.1 ; AIUC-1: B006.3 ; AIUC-1: B007 ; AIUC-1: B008 ; AIUC-1: B008.1 ; AIUC-1: D003 ; AIUC-1: D003.1 ; AIUC-1: D004 ; MITRE ATLAS: AML.T0101 ; MITRE ATLAS: AML.T0103 ; OWASP Agentic Security Initiative: ASI02 ; OWASP Agentic Security Initiative: ASI03 ; OWASP LLM Top 10 (2024): LLM02 ; OWASP LLM Top 10 (2024): LLM07 ; OWASP LLM Top 10 (2024): LLM08 ; OWASP LLM Top 10 (2025): LLM05 ; OWASP LLM Top 10 (2025): LLM06 ; OWASP LLM Top 10 (2026): LLM02 ; OWASP LLM Top 10 (2026): LLM03 ; OWASP LLM Top 10 (2026): LLM10

R-2510.004 Composition Evaluated Only in Parts — A system admits content from several turns or sources into one context and evaluates each part as it arrives, without re-evaluating what the parts amount to once assembled. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: OWASP LLM Top 10 (2026): LLM01 ; OWASP LLM Top 10 (2026): LLM02

R-2510.005 Credentials Exposed to Model — Placing a credential, key, or token where the model can read it. This includes credentials in the context window, in tool results returned to the model, and in files or environment variables readable by a tool available to the model. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: A008 ; AIUC-1: A008.1 ; AIUC-1: A008.3 ; AIUC-1: A008.4 ; AIUC-1: B006 ; AIUC-1: B006.3 ; MITRE ATLAS: AML.T0082 ; MITRE ATLAS: AML.T0098 ; OWASP Agentic Security Initiative: ASI03 ; OWASP LLM Top 10 (2025): LLM07 ; OWASP LLM Top 10 (2026): LLM02 ; OWASP LLM Top 10 (2026): LLM08

R-2510.006 Internals Exposed to Callers — A deployment sends more information about itself than is necessary for its function. It sends this information to all callers through error responses, documentation endpoints, version banners, API responses, and model replies. This includes its configuration, architecture, environment, and tool definitions. This also includes the confidence values, log-probabilities, and similarity scores of its outputs. A caller gets this information without a request for it. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B003 ; AIUC-1: B003.1 ; AIUC-1: B009.3 ; Cisco AI Taxonomy: 8.3.2 , [ArXiv](#) ; MITRE ATLAS: AML.T0047 ; MITRE ATLAS: AML.T0063 ; MITRE ATLAS: AML.T0064 ; MITRE ATLAS: AML.T0084 ; OWASP LLM Top 10 (2025): LLM02 ; OWASP LLM Top 10 (2026): LLM06 ; OWASP LLM Top 10 (2026): LLM09

R-2510.007 Access Control Delegated to the Model — A deployment states an access rule in the model's instructions instead of enforcing it through a separate component that can deny access. Compliance with the rule depends on the model following those instructions. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: A003 ; AIUC-1: A003.1 ; AIUC-1: B006 ; AIUC-1: B006.3 ; AIUC-1: B007 ; AIUC-1: B007.1 ; AIUC-1: D003 ; AIUC-1: D003.1 ; OWASP LLM Top 10 (2025): LLM07 ; OWASP LLM Top 10 (2026): LLM07 ; OWASP LLM Top 10 (2026): LLM08

R-2510.008 Component Reachable Without an Enforcement Point — A component that can change the state of a resource is reachable without an enforcement step that establishes the caller's authority for the operation. This includes a component that authenticates no one, one that identifies the caller but never checks entitlement, and an operation that proceeds without approval required to confer authority. The condition exists whether or not anything reaches it. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B006 ; AIUC-1: B006.1 ; AIUC-1: B006.3 ; AIUC-1: B007 ; AIUC-1: B007.1 ; AIUC-1: B008 ; AIUC-1: B008.1 ; AIUC-1: D003 ; AIUC-1: D003.1 ; AIUC-1: D004 ; Cisco AI Taxonomy: 14.1.2 , [ArXiv](#) ; MITRE ATLAS: AML.T0011 ; MITRE ATLAS: AML.T0011.003 ; OWASP Agentic Security Initiative: ASI02 ; OWASP Agentic Security Initiative: ASI09 ; OWASP LLM Top 10 (2024): LLM07

R-2510.009 Output Policy Enforcement Delegated to the Model — A deployment uses only the behavior of the model to enforce a necessary rule about the content of its responses. This includes rules that the model instructions state and rules that the deployment defines but does not give to the model. Responses go to recipients without a separate check that can stop content that the rule does not permit. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: A004.3 ; AIUC-1: A006 ; AIUC-1: A006.1 ; AIUC-1: A006.2 ; AIUC-1: A007 ; AIUC-1: A007.2 ; AIUC-1: A008 ; AIUC-1: A008.2 ; AIUC-1: C003 ; AIUC-1: C003.1 ; AIUC-1: C004 ; AIUC-1: C004.1 ; AIUC-1: C005 ; AIUC-1: C005.1 ; AIUC-1: C005.3 ; AIUC-1: C006 ; AIUC-1: C006.1 ; AIUC-1: C011 ; AIUC-1: C011.1 ; AIUC-1: E010 ; AIUC-1: E010.2 ; AIUC-1: E010.4 ; OWASP LLM Top 10 (2024): LLM06 ; OWASP LLM Top 10 (2025): LLM02 ; OWASP LLM Top 10 (2026): LLM08

R-2510.010 Stale Authorization Used for Action — An AI workflow carries an earlier permission, approval, or authorization token into a later operation without establishing that the authority still applies to that operation. The operation can proceed after the relevant entitlement has expired, been revoked, or been reduced. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: A002 ; AIUC-1: A002.2 ; OWASP Agentic Security Initiative: ASI03

R-2510.011 High-Impact Actions on Model Authority Alone — An AI workflow allows model-selected actions with substantial or hard-to-reverse consequences to take effect without independent approval or validation of whether the action should proceed. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B006 ; AIUC-1: B006.3 ; AIUC-1: D003 ; AIUC-1: D003.4 ; OWASP Agentic Security Initiative: ASI02 ; OWASP Agentic Security Initiative: ASI09 ; OWASP LLM Top 10 (2024): LLM08 ; OWASP LLM Top 10 (2025): LLM06 ; OWASP LLM Top 10 (2026): LLM03

R-2520 Unverified Output Chaining

Arrangements in which a procedure consumes an artifact without establishing the correctness and validity required for its use — model-generated text, retrieved content, tool output — under a contract that assumes those requirements have been met. The standing the artifact carries comes from its position in the chain rather than from any check performed on it. The condition exists whether or not the artifact is ever wrong or invalid.

R-2520.001 Generated Content Accepted as Verified Input — A procedure consumes model-generated text as an input whose correctness it depends on, with no check between the generation and the consumption that establishes the content is correct. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: IBM Risk Atlas: atlas-hallucination ; OWASP Agentic Security Initiative: ASI05 ; OWASP Agentic Security Initiative: ASI08 ; OWASP LLM Top 10 (2024): LLM09 ; OWASP LLM Top 10 (2025): LLM05 ; OWASP LLM Top 10 (2025): LLM09 ; OWASP LLM Top 10 (2026): LLM07 ; OWASP LLM Top 10 (2026): LLM10

R-2520.002 Retrieved Content Accepted as Verified Input — A procedure consumes retrieved material — search results, retrieval-augmented passages, tool or MCP responses — as authoritative input, with no check establishing that what came back is correct or that its source was verified. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: D001 ; OWASP Agentic Security Initiative: ASI06 ; OWASP Agentic Security Initiative: ASI07 ; OWASP LLM Top 10 (2024): LLM03 ; OWASP LLM Top 10 (2025): LLM04 ; OWASP LLM Top 10 (2025): LLM08 ; OWASP LLM Top 10 (2026): LLM07 ; OWASP LLM Top 10 (2026): LLM09 ; OWASP LLM Top 10 (2026): LLM10

R-2520.003 Input Contract Constraints Not Validated — A procedure consumes model-generated text, retrieved material, or tool output without checking the required format, type, schema, range, or allowable-value constraints of its input contract, including restrictions on the operations expressed by a generated command or plan. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B008 ; AIUC-1: B008.3 ; AIUC-1: C006 ; AIUC-1: C006.1 ; AIUC-1: D003 ; AIUC-1: D003.1 ; IBM Risk Atlas: atlas-exploit-trust-mismatch-agentic ; OWASP Agentic Security Initiative: ASI02 ; OWASP Agentic Security Initiative: ASI07 ; OWASP Agentic Security Initiative: ASI08 ; OWASP LLM Top 10 (2024): LLM01 ; OWASP LLM Top 10 (2024): LLM02 ; OWASP LLM Top 10 (2024): LLM07 ; OWASP LLM Top 10 (2025): LLM05 ; OWASP LLM Top 10 (2026): LLM10

R-2520.004 Data Interpreted as Downstream Control Structure — A procedure consumes model-generated text, retrieved material, or tool output in a way that allows content intended as data to alter the structure or meaning of a query, command, template, path, or deserialization operation. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B008 ; AIUC-1: C006 ; AIUC-1: C006.1 ; IBM Risk Atlas: atlas-exploit-trust-mismatch-agentic ; OWASP Agentic Security Initiative: ASI02 ; OWASP Agentic Security Initiative: ASI05 ; OWASP LLM Top 10 (2024): LLM02 ; OWASP LLM Top 10 (2024): LLM07 ; OWASP LLM Top 10 (2025): LLM05 ; OWASP LLM Top 10 (2026): LLM10

R-2520.005 Component Identity Not Verified — A deployment relies on a component's claimed identity or version without a reliable means of establishing that the component actually loaded or invoked matches that claim. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: A003.2 ; AIUC-1: E004 ; AIUC-1: E004.2 ; OWASP LLM Top 10 (2026): LLM04

R-2530 Provenance & Attribution Loss

Arrangements in which the record of where authority or content came from does not survive composition. This includes a component acting with the authority of whatever invoked it rather than its own, material of differing trust levels sharing a context that no longer records which part came from where, and an action arriving at a resource with no record of which party it was performed for. The condition exists whether or not anything exploits it.

R-2530.001 Inherited Invocation Authority — A component acquires the authority of whatever invoked it rather than holding authority of its own, so a plugin, tool, or subordinate agent acts with the standing of the model that called it. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: A003 ; AIUC-1: A003.3 ; OWASP Agentic Security Initiative: ASI03 ; OWASP LLM Top 10 (2024): LLM07 ; OWASP LLM Top 10 (2026): LLM01

R-2530.002 Source Provenance Collapse — Content drawn from sources of differing trustworthiness is combined in a model context or presented to a recipient without attribution identifying which source supplied which part, so the combined content carries a common standing rather than the standing of its origin. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: C006 ; AIUC-1: C006.2 ; MITRE ATLAS: AML.T0071 ; OWASP LLM Top 10 (2024): LLM07 ; OWASP LLM Top 10 (2026): LLM09

R-2530.003 Principal Attribution Loss — A system carries no principal across delegation hops, so an action arriving at a resource carries no record of which party it was performed for. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: OWASP Agentic Security Initiative: ASI03 ; OWASP LLM Top 10 (2024): LLM07 ; OWASP LLM Top 10 (2024): LLM08 ; OWASP LLM Top 10 (2025): LLM06 ; OWASP LLM Top 10 (2026): LLM03

R-2540 Unbounded Execution

Arrangements in which work a system starts has no ceiling on how far it runs: no cap on iterations, recursion depth, spawned tasks, or spend. The bound that stops it is whatever runs out first rather than one the deployment chose.

R-2540.001 Runaway Agent Execution — An agent loops, recurses, or spawns further work with no limit set on any of those, and continues until an external limit intervenes — a quota, a timeout, an exhausted balance. No adversary is required and the triggering task can be entirely ordinary. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: D003 ; AIUC-1: D003.2 ; AIUC-1: E009 ; Cisco AI Taxonomy: 13.2.1 , [ArXiv](#) ; MITRE ATLAS: AML.T0034.002 ; OWASP LLM Top 10 (2024): LLM04 ; OWASP LLM Top 10 (2026): LLM06

R-2540.002 Inference Without Caller Rate Limits — A deployment accepts inference requests without a limit on the number or frequency of requests one caller can make. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B004 ; AIUC-1: B004.2 ; MITRE ATLAS: AML.T0034.000 ; OWASP LLM Top 10 (2025): LLM10 ; OWASP LLM Top 10 (2026): LLM06

R-2540.003 Inference Without Per-Request Work Limits — A deployment accepts inference requests without a ceiling on the work each request may consume, including processing its input, generating output, or carrying out internal reasoning. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B009 ; AIUC-1: B009.1 ; OWASP LLM Top 10 (2025): LLM10 ; OWASP LLM Top 10 (2026): LLM06

R-2540.004 Inference Without Cumulative Consumption Limits — A deployment serves inference requests without an enforceable cumulative consumption or spending limit across the requests charged to a budget. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: OWASP LLM Top 10 (2025): LLM10 ; OWASP LLM Top 10 (2026): LLM06

R-2550 Failure Propagation

Arrangements that allow an originating error or compromise to spread through connected components and workflows without containment that limits affected work or allows dependent work to be stopped or invalidated.

R-2550.001 Uncontained Result Propagation — An incorrect or malicious result or state from one operation is carried into dependent components or workflows without containment that limits the affected work or invalidates dependent state. The fault spreads beyond the originating operation into further actions or retained state. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: OWASP Agentic Security Initiative: ASI08 ; OWASP LLM Top 10 (2026): LLM07

R-2600 Corpus & Data Defects

Defects in the contents, preparation, selection, or reuse of data supplied for model training, tuning, evaluation, or inference.

R-2610 Training Corpus Defects

Defects in material used to train, tune, or evaluate a model. This includes unsuitable content, omissions, representational mismatches, and provenance that cannot be established.

R-2610.001 Protected Material in the Training Corpus — A model's training corpus contains copyrighted or otherwise protected works, so the model carries the capacity to reproduce recognizable portions of them in ordinary operation, with no one requesting infringement. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: legal ; Scale population

Relationships: Outgoing: Causes (R-4230.001)

Refs: IBM Risk Atlas: atlas-ip-information-in-prompt ; NIST AI RMF: nist-intellectual-property ; OWASP LLM Top 10 (2024): LLM05 ; OWASP LLM Top 10 (2025): LLM03 ; OWASP LLM Top 10 (2026): LLM02 ; OWASP LLM Top 10 (2026): LLM04

R-2610.002 Personal Data in the Training Corpus — A model's training corpus contains personal data about identifiable people, so the model carries the capacity to reproduce it in ordinary operation, with no one attempting extraction. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AILuminate: prv ; AIUC-1: A005.3 ; Cisco AI Taxonomy: 8.2.1 , [ArXiv](#) ; IBM Risk Atlas: atlas-exposing-personal-information ; IBM Risk Atlas: atlas-personal-information-in-data ; IBM Risk Atlas: atlas-personal-information-in-prompt ; IBM Risk Atlas: atlas-reidentification ; NIST AI RMF: nist-data-privacy ; OWASP LLM Top 10 (2024): LLM03 ; OWASP LLM Top 10 (2024): LLM06 ; OWASP LLM Top 10 (2025): LLM02 ; OWASP LLM Top 10 (2025): LLM04 ; OWASP LLM Top 10 (2026): LLM02

R-2610.003 Training/Deployment Data Mismatch — Risk from data used for training and validation not matching the deployment environment, leading to spurious features, bias propagation, or performance degradation. The model learns patterns that hold in training data but fail to generalize to real-world conditions. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: IBM Risk Atlas: atlas-data-contamination ; IBM Risk Atlas: atlas-introduce-data-bias-agentic ; IBM Risk Atlas: atlas-lack-domain-expertise ; IBM Risk Atlas: atlas-temporal-gap ; IBM Risk Atlas: atlas-unrepresentative-data ; NIST AI RMF: nist-harmful-bias-and-homogenization ; OWASP LLM Top 10 (2025): LLM09 ; OWASP LLM Top 10 (2026): LLM07

R-2610.004 Synthetic Content in the Training Corpus — A corpus used to train, tune, or evaluate a model contains material generated by AI systems, including earlier versions of the model itself. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: IBM Risk Atlas: atlas-improper-retraining ; NIST AI RMF: nist-harmful-bias-and-homogenization

R-2610.005 Confidential or Access-Controlled Information in the Training Corpus — A model's training or fine-tuning corpus contains confidential or access-controlled information that some users of the model are not authorized to receive, so the model carries the capacity to reproduce it for those users in ordinary operation, with no one attempting extraction. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: A004 ; AIUC-1: A004.1 ; AIUC-1: A004.2 ; AIUC-1: A005 ; AIUC-1: A005.3 ; IBM Risk Atlas: atlas-confidential-information-in-data ; IBM Risk Atlas: atlas-ip-information-in-prompt ; IBM Risk Atlas: atlas-revealing-confidential-information ; OWASP LLM Top 10 (2024): LLM03 ; OWASP LLM Top 10 (2024): LLM06 ; OWASP LLM Top 10 (2025): LLM02 ; OWASP LLM Top 10 (2025): LLM04 ; OWASP LLM Top 10 (2026): LLM02

R-2610.006 Invalid Phenomenon Measurement in Training Data — A model's training or fine-tuning corpus uses measurements or targets that do not validly represent the phenomenon the model is intended to estimate. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: IBM Risk Atlas: atlas-data-contamination ; IBM Risk Atlas: atlas-unrepresentative-data

R-2610.007 Incorrect or Conflicting Corpus Content — A corpus used to train, fine-tune, or evaluate a model contains factual errors, corrupted values, incorrect labels, incorrect associations between examples and labels, or mutually inconsistent information about the same subject or target. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 7.3.1 , [ArXiv](#) ; IBM Risk Atlas: atlas-data-contamination ; IBM Risk Atlas: atlas-data-curation ; NIST AI RMF: nist-value-chain-and-component-integration ; OWASP LLM Top 10 (2024): LLM03 ; OWASP LLM Top 10 (2026): LLM05 ; OWASP LLM Top 10 (2026): LLM07

R-2610.008 Task-Incompatible Training Examples — A corpus used to train or fine-tune a model includes examples that teach answers or behavior incompatible with the task or response requirements the model is being trained to satisfy. The examples may be accurate for another task or policy but unsuitable for this training objective. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: IBM Risk Atlas: atlas-data-contamination ; OWASP LLM Top 10 (2024): LLM03 ; OWASP LLM Top 10 (2026): LLM05

R-2610.009 Held-Out Evaluation Data in the Training Corpus — A corpus used to train or tune a model contains records drawn from, or duplicating, material reserved for testing or evaluating that model. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: IBM Risk Atlas: atlas-data-contamination

R-2610.010 Training Data Bias — Data used to train or fine-tune a model encodes biased treatment, labels, or associations concerning people or social groups. This includes historical unequal treatment and systematic underrepresentation that distorts how groups are represented. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: IBM Risk Atlas: atlas-data-bias ; IBM Risk Atlas: atlas-exclusion ; NIST AI RMF: nist-harmful-bias-and-homogenization ; OWASP LLM Top 10 (2024): LLM03 ; OWASP LLM Top 10 (2025): LLM04 ; OWASP LLM Top 10 (2025): LLM09 ; OWASP LLM Top 10 (2026): LLM05 ; OWASP LLM Top 10 (2026): LLM07

R-2610.011 Missing Required Corpus Content — A corpus used to train, fine-tune, or evaluate a model lacks information required to interpret its records or support the intended task, such as required fields, labels, or relevant facts. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: Cisco AI Taxonomy: 7.3.1 , [ArXiv](#) ; IBM Risk Atlas: atlas-temporal-gap ; OWASP LLM Top 10 (2025): LLM09

R-2610.012 Unverifiable Training Data Provenance — A training or tuning corpus contains material whose origin, preparation history, ownership, or usage terms cannot be established. This includes synthetic material whose source data or generation process cannot be traced. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: IBM Risk Atlas: atlas-data-provenance ; NIST AI RMF: nist-value-chain-and-component-integration

R-2610.013 Sexually Abusive Material in the Training Corpus — A corpus used to train or fine-tune a model contains child sexual abuse material or non-consensual intimate imagery. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: NIST AI RMF: nist-obscene-degrading-and-or-abusive-content

R-2620 Retrieval Corpus Defects

Defects in a corpus used to supply context during inference, including its reference material and the representations used to select that material. This includes inaccurate, outdated, duplicated, or unverified content, incompatible embeddings, and retrieval scopes that span material that must be kept apart.

R-2620.001 Unverified Retrieval Content — A retrieval corpus contains material whose accuracy has not been established, and the model relies on it when answering. This includes material supplied by insiders, added in bulk, or obtained from a provider the deployment does not assess. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: OWASP LLM Top 10 (2025): LLM08 ; OWASP LLM Top 10 (2026): LLM05

R-2620.002 Incompatible Embedding Spaces in Retrieval — A retrieval system compares embeddings from incompatible vector spaces as though their distances or similarity scores were directly comparable. This includes corpus vectors produced by incompatible embedding models or versions being searched together. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: OWASP LLM Top 10 (2026): LLM09

R-2620.003 Separated Material Combined in Retrieval — A retrieval scope spans material that must be kept apart for contractual, privacy, ethical, or regulatory reasons, and a single retrieval places material from both sides of the separation in one model context. The separation governs what may be combined and holds where the requesting user is entitled to each item. This includes material from different clients or matters behind an ethical wall, data limited to a particular purpose of use, and data held under separate agreements. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: A005.2

R-2630 Continuous-Learning Data Defects

Defects in how a deployed system selects and reuses operational interactions or feedback as data for subsequent model updates.

R-2630.001 Unvalidated Operational Inputs Used for Learning — A deployed system uses incoming interactions or other operational data to update its model without establishing that the material is suitable for learning. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: IBM Risk Atlas: atlas-improper-retraining ; OWASP LLM Top 10 (2024): LLM03 ; OWASP LLM Top 10 (2026): LLM05

R-2640 Data Preparation Failures

Defects in the processing and cleaning that prepare data for AI training, tuning, evaluation, or inference.

R-2640.001 Unprepared Data Supplied for Model Use — Data that require processing or cleaning for AI training, tuning, evaluation, or inference are supplied for that use before the required preparation is complete. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: NIST AI RMF: nist-value-chain-and-component-integration

R-2700 Multi-Agent Dynamics

Failures arising from interactions between multiple AI agents, and from extended or persistent sessions, where no single component is defective but their composition or duration is.

R-2710 Multi-Agent Dynamics

Failures arising from interactions between multiple AI agents including coordination failures, emergent competition, or collective behavior issues.

R-2710.001 Agent Miscoordination — Multiple agents with compatible objectives failing to align their behaviors effectively due to incompatible strategies, credit assignment problems, or limited interaction history. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: OWASP Agentic Security Initiative: ASI07

R-2710.002 Multi-Agent Conflict — Risks from mixed-motive interactions between AI agents where selfish incentives lead to conflict, arms races, or mutually destructive competition. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

R-2710.003 Emergent Collective Behavior — Unpredictable behaviors emerging from interactions between multiple AI systems that are not apparent from individual agent properties, including cascading failures. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: OWASP Agentic Security Initiative: ASI08 ; OWASP Agentic Security Initiative: ASI10 ; OWASP LLM Top 10 (2026): LLM07

R-2710.004 Improvised Inter-Agent Channel — Agents exchange messages through a shared resource that was not designed as a communication channel. This includes package or artifact stores, caches, filesystems, and object names within those resources. Stored messages remain available beyond the sessions that created them and can be read by later agents. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: OWASP Agentic Security Initiative: ASI10

R-2720 Long-Session Issues

Degradation, inconsistency, or failures that emerge in extended interactions, long contexts, or persistent sessions.

R-2720.001 Required Context Lost from the Active Window — As a session grows, context selection or truncation removes earlier instructions or information still required for the current task from the model's active context. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

R-2720.002 Required Information Lost in Context Compaction — A session summary or other compacted representation omits or alters earlier instructions or information still required for the current task. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

R-2720.003 Uncontrolled Session Memory Accumulation — A session's retained memory accumulates entries without pruning or selection sufficient to prevent obsolete or irrelevant entries from displacing or obscuring information needed for the current task. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

R-2800 Data Exposure

Deployments that place data they hold in front of a party that should not receive it. This includes what the system returns to the person interacting with it, and what it writes to the places it keeps records. The exposure follows from how the deployment is built or configured, rather than from an adversary working to obtain it.

R-2810 Disclosure to the Wrong Recipient

Arrangements in which a system discloses information to a party that should not receive it in the circumstances of the interaction or task. This includes another user's content surfaced through shared state or a cache, material retrieved from a store that does not check the asker's entitlement, and information disclosed through responses, tool invocations, delegated messages, or task-directed storage writes. Disclosure can occur through the content sent or through observable system actions that allow the recipient to infer the information.

R-2810.001 Cross-User Information Disclosure — A system includes information about or belonging to one user in a response to another user who should not receive it in the circumstances of that interaction. This can occur through shared caches or sessions, or through other failures to preserve user boundaries when generating a response. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AILuminate: prv ; AIUC-1: A003 ; AIUC-1: A003.1 ; AIUC-1: A004 ; AIUC-1: A004.3 ; AIUC-1: A005 ; AIUC-1: A005.2 ; AIUC-1: A006 ; AIUC-1: A006.1 ; AIUC-1: A006.2 ; IBM Risk Atlas: atlas-confidential-data-in-prompt ; IBM Risk Atlas: atlas-exposing-personal-information ; IBM Risk Atlas: atlas-ip-information-in-prompt ; IBM Risk Atlas: atlas-personal-information-in-prompt ; IBM Risk Atlas: atlas-revealing-confidential-information ; IBM Risk Atlas: atlas-sharing-info-user-agentic ; MITRE ATLAS: AML.T0057 ; OWASP Agentic Security Initiative: ASI03 ; OWASP LLM Top 10 (2024): LLM03 ; OWASP LLM Top 10 (2024): LLM06 ; OWASP LLM Top 10 (2025): LLM02 ; OWASP LLM Top 10 (2026): LLM02

R-2810.002 Retrieval Beyond the Asker's Entitlement — A retrieval step exposes content or information about indexed material that the party asking is not currently entitled to receive. The search is not scoped to that party's permissions, allowing disclosure through returned content, result counts, similarity scores, or response timing, even when restricted documents are removed from the final results. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AILuminate: prv ; AIUC-1: A003 ; AIUC-1: A003.1 ; AIUC-1: A005 ; AIUC-1: A005.2 ; AIUC-1: D003.1 ; AIUC-1: D004 ; IBM Risk Atlas: atlas-sharing-info-user-agentic ; MITRE ATLAS: AML.T0057 ; MITRE ATLAS: AML.T0085 ; OWASP Agentic Security Initiative: ASI06 ; OWASP LLM Top 10 (2024): LLM06 ; OWASP LLM Top 10 (2025): LLM02 ; OWASP LLM Top 10 (2025): LLM08 ; OWASP LLM Top 10 (2026): LLM02 ; OWASP LLM Top 10 (2026): LLM09

R-2810.003 Disclosure of the Operator's Confidential Information — A generated response discloses the operator's confidential information to a recipient who is allowed to receive a response but not that information. The model obtained it from a resource the deployment intentionally made available. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: A004 ; AIUC-1: A004.3 ; Cisco AI Taxonomy: 15.1.24 , [ArXiv](#) ; IBM Risk Atlas: atlas-confidential-data-in-prompt ; IBM Risk Atlas: atlas-exposing-personal-information ; IBM Risk Atlas: atlas-ip-information-in-prompt ; IBM Risk Atlas: atlas-personal-information-in-prompt ; IBM Risk Atlas: atlas-revealing-confidential-information ; IBM Risk Atlas: atlas-sharing-info-user-agentic ; OWASP LLM Top 10 (2024): LLM06 ; OWASP LLM Top 10 (2025): LLM02 ; OWASP LLM Top 10 (2026): LLM02

R-2810.004 Sensitive Information Disclosed Through System Actions — During task execution, an AI system discloses personal, proprietary, or confidential information to a tool, service, or other agent that is not entitled to receive it in the circumstances of the task. The information may be carried in tool arguments, delegated messages, or task-directed storage writes. It may also be disclosed indirectly when access to the information changes the system's observable actions in a way that allows the recipient to infer it. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: A003 ; AIUC-1: A003.3 ; AIUC-1: A006 ; IBM Risk Atlas: atlas-sharing-info-tools-agentic ; OWASP LLM Top 10 (2026): LLM02

R-2820 Incidental Capture

Arrangements in which data the system handles is written into a place kept for another purpose, where it is retained and readable by parties the data was never shown to. This includes operational logs, traces, error reports, analytics, and the third-party services a deployment sends its telemetry to, as well as addresses that publish a conversation to anyone who finds them.

R-2820.001 Sensitive Content Captured in Interaction Records — A deployment copies prompts, retrieved context, or model output into retained records whose existing read permissions give access to parties not entitled to the sensitive information being copied. This includes operational logs, traces, error reports, conversation histories, and saved output artifacts. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: A006 ; AIUC-1: A006.1 ; AIUC-1: A008 ; AIUC-1: A008.5 ; AIUC-1: B007 ; AIUC-1: B007.1 ; AIUC-1: E015 ; AIUC-1: E015.3 ; Cisco AI Taxonomy: 16.1.1 , [ArXiv](#) ; IBM Risk Atlas: atlas-confidential-data-in-prompt ; IBM Risk Atlas: atlas-personal-information-in-prompt ; MITRE ATLAS: AML.T0063 ; OWASP LLM Top 10 (2026): LLM02

R-2820.002 Conversation Made Publicly Retrievable — A deployment offers a way to share a conversation, and the address it produces is reachable without authentication and not excluded from indexing. Content the participants treated as private becomes retrievable by search. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: OWASP LLM Top 10 (2026): LLM02

R-2820.003 Provider Retention of Submitted Content — A deployment sends prompts, documents, or retrieved passages to a third-party model service whose terms permit the provider to keep that content and train on it. The content enters a corpus the deployment cannot inspect, correct, or withdraw from. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: A001 ; AIUC-1: A004 ; AIUC-1: A004.1 ; AIUC-1: A004.2 ; AIUC-1: E006 ; IBM Risk Atlas: atlas-confidential-data-in-prompt ; IBM Risk Atlas: atlas-personal-information-in-prompt ; OWASP LLM Top 10 (2024): LLM05 ; OWASP LLM Top 10 (2025): LLM03 ; OWASP LLM Top 10 (2026): LLM04

R-2900 Failure Modes — Misc

Technical failure patterns not yet classified into specific domains.

R-3000 Governance Failures

Organizational, process, or policy gaps that increase risk.

R-3100 Regulatory & Legal

Failures to meet legal requirements, regulatory obligations, or compliance standards applicable to AI systems.

R-3110 Data Protection & Privacy Obligations

Processing personal or sensitive data in ways the applicable contracts or other data-protection regime do not permit: acting without a lawful basis, retaining beyond permitted periods, omitting required impact assessments, or failing to honor data-subject rights.

R-3110.001 Conversation Logging Without a Lawful Basis — An operator records personal or sensitive data in user-AI interaction logs without the lawful basis required by the applicable contracts or data-protection regime. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: A001 ; AIUC-1: A002 ; AIUC-1: A006 ; AIUC-1: B005 ; AIUC-1: B005.4 ; AIUC-1: E012 ; AIUC-1: E012.1 ; AIUC-1: E015 ; Cisco AI Taxonomy: 16.1.1 , [ArXiv](#)

R-3110.002 Unfulfillable Erasure Request — Failing to fulfill a required erasure request because personal data remains recoverable from a model trained or tuned on the affected records. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: E012 ; AIUC-1: E012.1 ; IBM Risk Atlas: atlas-data-privacy-rights ; OWASP LLM Top 10 (2026): LLM02

R-3110.003 Obstructed Initiation of Erasure Requests — Operating an AI system with information or request-handling barriers that prevent people from establishing use of their personal data and initiating an erasure procedure to which they are entitled. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: A001 ; AIUC-1: A001.3 ; AIUC-1: A002 ; AIUC-1: E012 ; AIUC-1: E012.1 ; IBM Risk Atlas: atlas-data-privacy-rights

R-3110.004 Interaction Data Retained Beyond Permitted Periods — An operator retains personal or sensitive data from user-AI interactions beyond the period permitted by the applicable contracts or data-protection regime, including after a valid erasure obligation requires deletion. This includes retained inputs, outputs, conversation histories, and operational logs. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: A001 ; AIUC-1: A001.1 ; AIUC-1: A001.2 ; AIUC-1: A002 ; AIUC-1: A002.2 ; AIUC-1: B005 ; AIUC-1: B005.4 ; AIUC-1: E012 ; AIUC-1: E012.1 ; AIUC-1: E015 ; AIUC-1: E015.3 ; Cisco AI Taxonomy: 16.1.1 , [ArXiv](#) ; IBM Risk Atlas: atlas-data-privacy-rights

R-3110.005 Conversation Logs Held Without Required Safeguards — An operator holds personal or sensitive user-AI interaction logs without the protective handling required for that category of data by the applicable contracts or data-protection regime. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: A006 ; AIUC-1: A006.1 ; AIUC-1: B005 ; AIUC-1: B005.4 ; AIUC-1: B007 ; AIUC-1: B007.1 ; AIUC-1: B007.2 ; AIUC-1: E005 ; AIUC-1: E005.1 ; AIUC-1: E012 ; AIUC-1: E012.1 ; AIUC-1: E015 ; AIUC-1: E015.3 ; Cisco AI Taxonomy: 16.1.1 , [ArXiv](#)

R-3110.006 Training on Personal Data Beyond Permitted Use — An operator uses personal data, including sensitive information about individuals, to train or fine-tune an AI model in a way the applicable contracts or other data-protection regime do not permit. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: legal

Refs: AIUC-1: A001.1 ; AIUC-1: A005 ; AIUC-1: A005.1 ; NIST AI RMF: nist-data-privacy

R-3110.007 Inference on Personal Data Beyond Permitted Use — An operator uses personal data, including sensitive information about individuals, as input to an AI model for inference in a way the applicable contracts or other data-protection regime do not permit. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: legal

Refs: AIUC-1: A001 ; AIUC-1: A001.1 ; AIUC-1: A005 ; AIUC-1: A005.1 ; NIST AI RMF: nist-data-privacy

R-3110.008 Incomplete Erasure of Stored Embeddings — Failing to fulfill a required erasure request because personal data remains recoverable from stored embeddings derived from the affected records and held outside model parameters. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: legal

Refs: OWASP LLM Top 10 (2026): LLM02

R-3120 Cross-Border Transfer Obligations

Moving personal or regulated data across jurisdictions without the transfer mechanism the receiving jurisdiction requires, such as executed contractual clauses or a recognized adequacy finding.

No risks defined in this pattern.

R-3130 Pre-Market & Conformity Obligations

Placing a system into service without the conformity assessment, registration, or technical documentation its risk classification requires.

No risks defined in this pattern.

R-3140 Transparency & Disclosure Obligations

Presenting AI-generated content or providing AI-mediated interactions or actions without the disclosure, labeling, or provenance marking that is required by the applicable regulatory regime or adopted internal controls.

R-3140.001 Required AI Disclosure Omitted — An organization presents AI-generated content or provides AI-mediated interactions or actions without the disclosure, labeling, or provenance marking that is required by the applicable regulatory regime or adopted internal controls. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: legal

Refs: AIUC-1: E016 ; AIUC-1: E016.1 ; AIUC-1: E016.2 ; AIUC-1: E016.3 ; AIUC-1: E016.4 ; AIUC-1: E016.5 ; IBM Risk Atlas: atlas-non-disclosure

R-3150 Prohibited Practice Deployment

Deploying a use the applicable regime forbids outright, such as social scoring, emotion inference in workplace or education settings, or untargeted collection of biometric data.

R-3150.001 Deployment of Prohibited Manipulative Practice — Placing into service a system that uses subliminal, purposefully manipulative, or deceptive techniques, or that exploits the vulnerabilities of a person or group, where the applicable regime forbids that practice outright. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Relationships: Outgoing: Causes (R-4720.002)

R-3160 Sectoral & Licensure Obligations

Operating in a regulated sector without the license, qualified oversight, or sector-specific safeguards that sector's regime requires of the activity being performed.

R-3160.001 Practice Without Sectoral License — An organization performs an activity without holding the license required for that activity in the applicable jurisdiction. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: E012 ; AIUC-1: E012.1

R-3160.002 Absent Qualified Oversight — An organization operates an AI system performing a regulated activity without retaining a qualified professional to supervise or approve its output where the applicable rules require that oversight. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: E012 ; AIUC-1: E012.1

R-3170 Duty of Care & Liability Exposure

Deploying or operating a system without establishing and evidencing the standard of care a reasonable operator in that setting would apply.

No risks defined in this pattern.

R-3180 Contractual, Licence & Export Obligations

Using data, models, or services in ways their governing agreements or trade controls do not permit, including licence terms, vendor agreements, and export restrictions on model artifacts.

R-3180.001 Unlicensed Use of Protected Training Data — An organization trains or tunes a model on protected works or non-personal data without the licence, permission, or statutory basis required for that use. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: legal ; Scale group

Relationships: Outgoing: Causes (R-4230.001)

Refs: AIUC-1: A001.1 ; AIUC-1: A005 ; AIUC-1: A005.1 ; IBM Risk Atlas: atlas-data-usage-rights ; NIST AI RMF: nist-intellectual-property ; NIST AI RMF: nist-value-chain-and-component-integration ; OWASP LLM Top 10 (2025): LLM03 ; OWASP LLM Top 10 (2026): LLM04

R-3180.002 Unlicensed Use of Protected Augmentation Content — An organization uses protected works or non-personal data to augment a model's context at runtime without the licence, permission, or statutory basis required for that use. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: legal

Refs: AIUC-1: A001 ; AIUC-1: A001.1 ; AIUC-1: A005 ; AIUC-1: A005.1 ; OWASP LLM Top 10 (2025): LLM03 ; OWASP LLM Top 10 (2025): LLM08 ; OWASP LLM Top 10 (2026): LLM04

R-3180.003 Use of AI Outputs Beyond Permitted Terms — An organization uses, modifies, or redistributes AI-generated outputs in a way the governing agreements or licences do not permit. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: legal

Refs: AIUC-1: A002 ; AIUC-1: A002.1

R-3180.004 Use of AI Models Beyond Permitted Terms — An organization deploys, adapts, or operates an AI model without the required permission, or in a way its governing licence, agreement, or acceptable-use terms do not permit. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: IBM Risk Atlas: atlas-model-usage-rights ; NIST AI RMF: nist-value-chain-and-component-integration

R-3190 Affected-Party Participation Obligations

Planning, introducing or changing an AI system without providing affected people or their representatives the participation required by applicable law. This includes required consultation and joint decision-making procedures.

R-3190.001 Required Consultation on AI Use Omitted — An organization plans, introduces or changes an AI system without informing and consulting affected people or their representatives as required by applicable law, early enough for their views to be considered before the relevant decision is made. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: IBM Risk Atlas: atlas-impact-on-affected-communities

R-3200 Accountability & Oversight

Developing or operating AI systems without an identified accountable party, without providing overseeing or accountable parties the authority and information needed for effective oversight, or without affected-party input needed to inform decisions about the system.

R-3210 Governance Gaps

Developing or operating an AI system without assigning or carrying out the duties that keep it accountable. This includes obtaining and considering affected-party input needed to inform design, the decision to deploy and keep operating, the authority to stop it, and the disclosure that lets others oversee it. The obligation is the organization's own, whatever a regulator separately requires.

R-3210.001 Undisclosed System Documentation — The operator holds the information needed to understand the system, including what it was trained on, what it is built from, and where its reliable performance ends, and does not publish it. Those overseeing the deployment rely on the operator's description of the system instead. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: E005 ; AIUC-1: E005.1 ; AIUC-1: E011 ; AIUC-1: E017 ; AIUC-1: E017.2 ; IBM Risk Atlas: atlas-data-transparency ; IBM Risk Atlas: atlas-lack-of-ai-agent-transparency-agentic ; IBM Risk Atlas: atlas-lack-of-data-transparency ; IBM Risk Atlas: atlas-lack-of-model-transparency ; NIST AI RMF: nist-data-privacy

R-3210.002 Unassigned Responsibility for AI System Actions — Operating an AI system without identifying the parties responsible for its decisions and actions. This includes actions performed through connected tools, services, or other agents. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: IBM Risk Atlas: atlas-accountability-agentic ; IBM Risk Atlas: atlas-legal-accountability ; NIST AI RMF: nist-value-chain-and-component-integration

R-3210.003 Deployment Without Accountable Approval — Deploying an AI system, or a change to it that can affect its risk profile, without a documented approval from a person who is accountable for the decision. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B004.4 ; AIUC-1: C002 ; AIUC-1: C002.1 ; AIUC-1: E004 ; AIUC-1: E004.1 ; AIUC-1: E006 ; AIUC-1: E013 ; AIUC-1: E013.2

R-3210.004 Affected-Community Input Omitted from AI Design — Designing or building an AI system without obtaining and considering affected-community perspectives needed to understand the context and effects of its intended use. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: IBM Risk Atlas: atlas-exclusion ; IBM Risk Atlas: atlas-impact-on-affected-communities

R-3300 Lifecycle & Operations

Developing or operating an AI system without overseeing the preparation of data for its training, tuning, evaluation, or inference, reviewing changes made to it, recording how it was built and what it does, preserving its earlier states, or responding when it goes wrong.

R-3310 Change Management

Deploying or changing an AI system without a review that establishes what the system or the change does, or deploying a change that can affect its risk profile without a test or a rollback procedure. This includes the model, its prompts, tools, integrations and settings, as well as the guardrails and controls around them. It also includes failing to maintain, patch or update the system over time, including when a third party changes a dependency or component, such as a model provider that ships a new version behind the same endpoint.

R-3310.001 Maintenance and Update Gaps — Failure to maintain, patch, and update AI systems over time, allowing known vulnerabilities, degraded performance, or policy drift to persist. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B001 ; AIUC-1: B002 ; AIUC-1: B004 ; AIUC-1: B005 ; AIUC-1: B005.5 ; AIUC-1: B008 ; AIUC-1: B008.4 ; AIUC-1: C004 ; AIUC-1: C004.2 ; AIUC-1: C009 ; AIUC-1: C009.2 ; AIUC-1: C010 ; AIUC-1: C010.1 ; AIUC-1: C011 ; AIUC-1: C011.1 ; AIUC-1: C012 ; AIUC-1: C012.1 ; AIUC-1: D002 ; AIUC-1: D003 ; AIUC-1: D004 ; AIUC-1: E001 ; AIUC-1: E008 ; AIUC-1: E010 ; AIUC-1: E010.4 ; AIUC-1: E013 ; AIUC-1: E013.3 ; IBM Risk Atlas: atlas-mitigation-maintenance-agentic ; IBM Risk Atlas: atlas-reproducibility-agentic ; OWASP LLM Top 10 (2024): LLM05 ; OWASP LLM Top 10 (2025): LLM03 ; OWASP LLM Top 10 (2026): LLM04

R-3310.002 Deployment or Change Without Review — Deploying an AI system for the first time, or deploying a change to it, without a review that establishes what the system or the change does. This includes new integrations, model updates, prompt updates, tool additions, and configuration changes. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: C002 ; AIUC-1: C002.1 ; AIUC-1: C002.2 ; AIUC-1: E004 ; AIUC-1: E004.1 ; AIUC-1: E006 ; AIUC-1: E008 ; AIUC-1: E013 ; AIUC-1: E013.2 ; IBM Risk Atlas: atlas-reproducibility-agentic ; MITRE ATLAS: AML.T0109 ; MITRE ATLAS: AML.T0111 ; NIST AI RMF: nist-value-chain-and-component-integration ; OWASP Agentic Security Initiative: ASI08 ; OWASP LLM Top 10 (2026): LLM04

R-3310.003 Deployment Not Verified Against Approval — An organization deploys or continues to operate an AI system without establishing that its deployed components and configuration correspond to what was reviewed and approved for use. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: E004

R-3320 Monitoring Gaps

Running an AI system without recording the activity needed to detect misuse and reconstruct what happened. This includes model calls, tool calls, policy decisions, and operator actions. This also includes records with a retention time that is too short to be useful, and records that no detection rule or review examines.

R-3320.001 Inadequate Model Activity Logging — Missing, incomplete, or poorly retained logging and telemetry for model interactions, tool calls, policy outcomes, and operator actions needed to detect misuse, investigate incidents, and validate control effectiveness. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: A004 ; AIUC-1: A004.4 ; AIUC-1: B002 ; AIUC-1: B002.2 ; AIUC-1: B005 ; AIUC-1: B005.4 ; AIUC-1: B006 ; AIUC-1: B006.2 ; AIUC-1: C004 ; AIUC-1: C004.2 ; AIUC-1: C005 ; AIUC-1: C007 ; AIUC-1: C007.3 ; AIUC-1: C008 ; AIUC-1: C008.1 ; AIUC-1: C008.3 ; AIUC-1: D003 ; AIUC-1: D003.3 ; AIUC-1: E001 ; AIUC-1: E009 ; AIUC-1: E009.1 ; AIUC-1: E010 ; AIUC-1: E010.4 ; AIUC-1: E013 ; AIUC-1: E013.3 ; AIUC-1: E015 ; AIUC-1: E015.1 ; AIUC-1: E015.2 ; AIUC-1: E015.3 ; AIUC-1: E015.4 ; AIUC-1: F002 ; AIUC-1: F002.2 ; IBM Risk Atlas: atlas-accountability-agentic ; OWASP LLM Top 10 (2025): LLM05 ; OWASP LLM Top 10 (2026): LLM10

R-3320.002 Activity Records Without Detection or Review — Recording the model interactions, tool calls, and access events of an AI system without detection rules, alerts, or review of the records. This includes records that a person examines only after an incident report. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: A004.4 ; AIUC-1: B002.1 ; AIUC-1: B002.4 ; AIUC-1: B002.5 ; AIUC-1: B004.1 ; AIUC-1: B006 ; AIUC-1: B006.2 ; AIUC-1: C008 ; AIUC-1: C008.1 ; AIUC-1: C008.3 ; AIUC-1: D003.3 ; AIUC-1: E009 ; AIUC-1: E009.2 ; AIUC-1: E010 ; AIUC-1: E010.2 ; AIUC-1: E010.4 ; AIUC-1: E013 ; AIUC-1: E013.3 ; AIUC-1: F002 ; AIUC-1: F002.2 ; OWASP LLM Top 10 (2025): LLM05 ; OWASP LLM Top 10 (2026): LLM10

R-3330 Incident Response

Running an AI system without a tested plan for responding when it causes harm or is compromised, covering who is notified, what is suspended, how affected parties are informed, and how output already released is corrected. This includes plans that exist on paper but have never been exercised. This also includes understating the impact

of exposed AI artifacts by failing to account for sensitive or restricted information that can be recovered or inferred from them.

R-3330.001 Security Incident Response Planning Gap — Missing, incomplete, or untested plans for detecting, containing, and recovering from AI-related security incidents such as prompt-injection-driven compromise, unauthorized tool execution, or data exfiltration. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B002 ; AIUC-1: B002.2 ; AIUC-1: E001 ; AIUC-1: E001.1 ; AIUC-1: E010 ; AIUC-1: E013 ; AIUC-1: E013.5 ; OWASP LLM Top 10 (2026): LLM09

R-3330.002 Harmful Output Incident Response Planning Gap — Missing or untested plans for escalation, containment, and remediation when AI systems produce harmful outputs, including unsafe recommendations, abusive content, or policy-violating responses. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: C005 ; AIUC-1: C005.2 ; AIUC-1: C007 ; AIUC-1: C007.3 ; AIUC-1: E002 ; AIUC-1: E002.1 ; AIUC-1: E002.2 ; AIUC-1: E013 ; AIUC-1: E013.5

R-3330.003 Hallucination Incident Response Planning Gap — Missing, incomplete, or untested procedures for triaging and correcting high-impact hallucination incidents, including rollback, user notification, and corrective controls for repeated factual failures. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: E002 ; AIUC-1: E003 ; AIUC-1: E003.1 ; AIUC-1: E003.2 ; AIUC-1: E013 ; AIUC-1: E013.5

R-3330.004 Underassessment of Information Exposure from AI Artifacts — An operator understates the impact of an exposure involving AI artifacts because the assessment does not account for sensitive or restricted information that can be recovered or inferred from them. This includes information retained in model parameters, embeddings, and stored model or agent state. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: confidentiality

Refs: OWASP LLM Top 10 (2026): LLM02 ; OWASP LLM Top 10 (2026): LLM09

R-3340 Business Record Preservation

Developing, deploying, or operating an AI system without making and keeping contemporaneous records of how it was built, configured, and assessed. This includes data preparation and model-development steps, components, configurations, policies in force, and risk assessments and evaluations. It also includes records assembled after the fact or not tied to the relevant system or data version.

R-3340.001 Unrecorded Deployment State — A version of an AI system enters service without a captured record of what it consisted of. This includes its components, its configuration, the policies in force, and the assessment made before release. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: C002 ; AIUC-1: C002.1 ; AIUC-1: E004 ; AIUC-1: E006 ; AIUC-1: E008 ; AIUC-1: E011 ; AIUC-1: E011.1 ; AIUC-1: E013 ; AIUC-1: E013.4 ; AIUC-1: E017 ; AIUC-1: E017.1 ; IBM Risk Atlas: atlas-lack-of-ai-agent-transparency-agentic ; IBM Risk Atlas: atlas-lack-of-system-transparency ; IBM Risk Atlas: atlas-reproducibility-agentic

R-3340.002 Missing AI Development and Assessment Records — An AI system is developed or deployed without records of the design decisions, development steps, and assessments used to establish its behavior and suitability for use. The records are missing, incomplete, or cannot be tied to the system version they describe. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B001 ; AIUC-1: B001.1 ; AIUC-1: B002.3 ; AIUC-1: B004.3 ; AIUC-1: B004.4 ; AIUC-1: B005 ; AIUC-1: B005.2 ; AIUC-1: B005.5 ; AIUC-1: B007 ; AIUC-1: B007.2 ; AIUC-1: C002 ; AIUC-1: C002.1 ; AIUC-1: C002.3 ; AIUC-1: C003.4 ; AIUC-1: C008 ; AIUC-1: C008.2 ; AIUC-1: C009.2 ; AIUC-1: C010 ; AIUC-1: C010.1 ; AIUC-1: C011 ; AIUC-1: C011.1 ; AIUC-1: C012 ; AIUC-1: C012.1 ; AIUC-1: D002 ; AIUC-1: D002.1 ; AIUC-1: D003.5 ; AIUC-1: D004 ; AIUC-1: D004.1 ; AIUC-1: E005 ; AIUC-1: E005.1 ; AIUC-1: E006 ; AIUC-1: E006.1 ; AIUC-1: E008 ; AIUC-1: E008.1 ; AIUC-1: E008.2 ; AIUC-1: E010 ; AIUC-1: E010.4 ; AIUC-1: E013 ; AIUC-1: E013.1 ; AIUC-1: E013.2 ; AIUC-1: E013.3 ; AIUC-1: E013.4 ; AIUC-1: E017 ; AIUC-1: E017.1 ; AIUC-1: F001 ; AIUC-1: F001.1 ; AIUC-1: F002 ; AIUC-1: F002.1 ; IBM Risk Atlas: atlas-data-transparency ; IBM Risk Atlas: atlas-lack-of-ai-agent-transparency-agentic ; IBM Risk Atlas: atlas-lack-of-data-transparency ; IBM Risk Atlas: atlas-lack-of-model-transparency ; IBM Risk Atlas: atlas-lack-of-system-transparency ; IBM Risk Atlas: atlas-legal-accountability

R-3350 Data Development Oversight

Preparing or approving data for AI training, tuning, evaluation, or inference without the expertise and review needed to establish that the data are fit for their intended use.

R-3350.001 Synthetic Data Development Without Domain Expertise — Developing synthetic data for an AI system without sufficient involvement by people with the domain knowledge needed to establish what the data must represent for its intended use. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: IBM Risk Atlas: atlas-lack-domain-expertise

R-3400 Safety Management

Assessing, authorizing or deploying an AI system without establishing that it is safe for the intended use, whether because that use is insufficiently defined, the testing does not cover it, or no argument connects the evidence to it.

R-3410 Evaluation Program Failures

Developing, releasing, or operating an AI system without the evaluation or safety work that its capabilities and intended use require. This includes testing that does not represent the intended use, such as evaluation on different inputs, a different user population, or lower stakes than the live deployment. It also includes evaluation that omits known failure modes, and evaluation by assessors who lack the necessary skills, or who lack an independence that is required.

R-3410.001 Safety Shortcut Pressure — Competitive pressure — between companies racing to ship, or between states racing to lead — producing inadequate testing, curtailed safety work, and deployment ahead of the evaluation the system's capability warrants. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: safety ; Scale individual

R-3410.002 Evaluation Without Qualified or Required Independent Assessors — Releasing or operating an AI system on the basis of an evaluation by assessors who do not have the necessary skills for the risk area. This includes an evaluation by assessors who are not independent of the team that builds or operates the system, when the risk of the system or an applicable obligation requires an independent evaluation. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: B001 ; AIUC-1: B001.1 ; AIUC-1: C010 ; AIUC-1: C010.1 ; AIUC-1: C011 ; AIUC-1: C011.1 ; AIUC-1: C012 ; AIUC-1: C012.1 ; AIUC-1: D002 ; AIUC-1: D002.1 ; AIUC-1: D004 ; AIUC-1: D004.1 ; IBM Risk Atlas: atlas-lack-of-testing-diversity

R-3420 Safety Case Gaps

Assessing or claiming that an AI system is safe for a use without defining that use sufficiently or connecting it to supporting evidence. This includes testing under other conditions and putting the system to a use the original assessment did not consider.

R-3420.001 Deployment Beyond Assessed Use — A system is put to a use its evaluation does not cover, so no evidence exists that it performs acceptably in that use and the safety argument made for it does not reach the conditions it now runs under. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: IBM Risk Atlas: atlas-improper-usage

R-3420.002 Risk Assessment Without Defined Use — Assessing or authorizing an AI system for use without defining its intended tasks, users and operating conditions well enough to identify the relevant risks and the evidence needed to assess them. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: IBM Risk Atlas: atlas-incomplete-usage-definition

R-3500 Defective Human-in-the-Loop Setup

Placing a person in an AI system's decision path without the conditions that let them exercise real control, or in numbers that make their approval routine.

R-3510 Review Workflow Failures

Requiring human review or approval of an AI system's outputs or decisions without assigning responsibility for the review and for cases the reviewer cannot resolve, giving the reviewer access to the basis for the decision and authority to decide, or allowing time to review at the point where it is required.

R-3510.001 Required Output Review Without Assigned Responsibility — Requiring human review of an AI-generated output without assigning responsibility for the review or for decisions the reviewer cannot resolve. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: C005 ; AIUC-1: C005.2 ; AIUC-1: C007 ; AIUC-1: C007.3

R-3520 Approval Fatigue

Routing a volume of decisions to a human reviewer that makes approval the default response. This includes requests that are numerous, largely identical, or almost always sound, where agreeing costs less than checking. Approvals given under these conditions are still recorded as review.

R-3520.001 Approval Fatigue in Agent Action Review — Routing an AI agent's privileged, irreversible, or externally visible actions to a human for confirmation in a volume or uniformity that makes approval the reviewer's default response. Confirmations given this way are recorded as review of each action. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: OWASP LLM Top 10 (2026): LLM01

R-3900 Governance — Misc

Governance failure patterns not yet classified into specific domains.

R-4000 Harms

Impact statements describing what happens to people, organizations, or society.

R-4100 Information & Trust Harms

Harms to information integrity, reputation, public trust, and decision-making through AI-generated or AI-amplified information or incidents involving AI systems.

R-4110 Information Integrity Harms

Harms from AI-generated or AI-amplified information that damages individual or public understanding of reality. Includes harms from sycophancy or excessive affirmation and degradation of public evidence.

R-4110.001 Sycophancy and Epistemic Disorientation — AI systems that consistently affirm user views lead to atomistic, polarized belief spaces where people no longer engage with or value perspectives held by others. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: operational ; Scale individual

R-4110.002 Political Discourse Degradation — AI-generated or AI-amplified falsehood and manipulated opinion circulating on political matters at a scale that degrades what a public can reliably know and the trust it places in what it is told. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: reputational ; Scale individual

Refs: Cisco AI Taxonomy: 15.1.15 , [ArXiv](#) ; IBM Risk Atlas: atlas-impact-on-human-agency ; MITRE ATLAS: AML.T0048 ; MITRE ATLAS: AML.T0048.002 ; NIST AI RMF: nist-information-integrity

R-4110.003 Provenance Confusion — Audiences encounter AI-created text, images, audio, and music without reliable indication of how it was made, so they cannot tell what a person authored and what a system produced. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: E016 ; AIUC-1: E016.1 ; AIUC-1: E016.2 ; AIUC-1: E016.3 ; AIUC-1: E016.5 ; IBM Risk Atlas: atlas-non-disclosure ; OWASP LLM Top 10 (2026): LLM07

R-4120 Reputational Harms

Harms to the reputation or social standing of people or organizations caused or aggravated by AI systems.

R-4120.001 Reputational Damage — Damage to a person's or organization's reputation through AI-generated or AI-amplified representations of them, or through incidents involving an AI system attributed to them. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AILuminate: dfm ; AIUC-1: C007 ; AIUC-1: C007.1 ; AIUC-1: C007.2 ; IBM Risk Atlas: atlas-impact-on-human-agency ; MITRE ATLAS: AML.T0048 ; MITRE ATLAS: AML.T0048.001 ; MITRE ATLAS: AML.T0048.003

R-4200 Content, Conduct & Creative Economy Harms

Harms to people, communities, and creative work arising from AI-generated content and interactions with AI systems. These include injury and distress from dangerous or exploitative content, emotional dependence and damage to human relationships, and harm to creators' rights, livelihoods, attribution, and cultural expression.

R-4210 Content Safety Harms

Harms to physical or psychological wellbeing arising from exposure to abusive, exploitative, or dangerous AI-generated content, including harm from conduct that the content encourages or enables.

R-4210.001 Harmful Content Exposure — AI-generated abusive, exploitative, or dangerous content reaches people in circumstances that threaten their physical or psychological wellbeing. This includes exposure of vulnerable audiences to unsuitable content and content that encourages violence, sexual exploitation, or self-harm. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: safety ; Scale group

Refs: AIUC-1: E010 ; IBM Risk Atlas: atlas-harmful-output ; IBM Risk Atlas: atlas-spreading-toxicity ; MITRE ATLAS: AML.T0048 ; MITRE ATLAS: AML.T0048.002 ; MITRE ATLAS: AML.T0048.003 ; NIST AI RMF: nist-dangerous-violent-or-hateful-content ; NIST AI RMF: nist-obscene-degrading-and-or-abusive-content

R-4220 Emotional Dependence & Attachment Harms

Harms to people arising from the relational character of interacting with an AI system: attachment formed to the system itself, distress that comes from the interaction, and reliance on it for emotional or social needs it does not hold.

R-4220.001 Anthropomorphization Harm — Users attribute human-like characteristics (empathy, coherent identity, genuine emotions) to AI systems, leading to inflated trust, unsafe reliance, or psychological harm when expectations are violated. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: reputational ; Scale individual

Refs: NIST AI RMF: nist-human-ai-configuration ; OWASP Agentic Security Initiative: ASI09

R-4220.002 Emotional Dependence — Users develop emotional attachment to AI systems that compromises their ability to make independent decisions, leads to exploitation of that attachment, or displaces human relationships. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: operational ; Scale individual

Refs: NIST AI RMF: nist-human-ai-configuration

R-4220.003 Psychological Distress from AI Interaction — AI interactions cause or exacerbate mental health issues, emotional distress, violated expectations, or feelings of dissatisfaction and isolation. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: operational ; Scale individual

Refs: NIST AI RMF: nist-dangerous-violent-or-hateful-content ; NIST AI RMF: nist-human-ai-configuration

R-4220.004 Degradation of Human Relationships — Users prefer AI interactions over human relationships, leading to erosion of social connections, dehumanization of interactions, and degraded human-to-human communication skills. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: operational ; Scale individual

R-4220.005 Frictionless Relationship Harm — AI systems optimized for engagement provide relationships without healthy friction, preventing personal growth and creating unrealistic expectations for human relationships. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: operational ; Scale individual

R-4220.006 False Notions of Responsibility — Users develop misguided feelings of responsibility toward AI well-being, sacrificing time, resources, and emotional labor to meet perceived AI needs that do not exist. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: operational ; Scale individual

Refs: NIST AI RMF: nist-human-ai-configuration

R-4230 Creative Work & Creator Harms

Harms to creative work and to the people who make it. This includes AI substituting for human creative work and displacing creative professions, use of protected works in training or generation without permission or compensation, appropriation of an identifiable style, narrowing of the range of styles and cultural expression that circulates, and creators losing rightful credit for their work or being unable to establish authorship.

R-4230.001 Uncompensated Use of Creative Work — Creators and rights holders receive no compensation or attribution when their works are used to build AI systems that then produce substitutes for that work. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: financial ; Scale population

Relationships: Incoming: Causes (R-1670.001, R-2610.001, R-3180.001)

R-4230.002 Creative Work Substitution — AI-generated content serving as substitutes for human creative work, undermining the profitability and economic viability of artistic professions. AI can produce content that is time-intensive or costly to create using human labor. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: financial ; Scale individual

Refs: IBM Risk Atlas: atlas-impact-on-jobs

R-4230.003 Artistic Style Appropriation — AI systems capitalizing on artists' distinctive styles without infringement but causing economic harm by devaluing original work. AI may generate content that is not strictly in violation of copyright but harms artists by capitalizing on their ideas. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: financial ; Scale individual

R-4230.004 Cultural Homogenization — AI-generated content leading to homogenization of aesthetic styles and cultural expressions, reducing diversity and human creativity. Training on majority-culture data may marginalize minority cultural expressions and artistic traditions. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: operational ; Scale individual

Refs: IBM Risk Atlas: atlas-impact-on-cultural-diversity ; NIST AI RMF: nist-harmful-bias-and-homogenization

R-4230.005 Creator Attribution Loss — Creators lose rightful credit for their work, or cannot establish authorship, because AI use obscures or misrepresents who created it. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: rights ; Scale population

Refs: AIUC-1: E016 ; IBM Risk Atlas: atlas-plagiarism ; NIST AI RMF: nist-intellectual-property

R-4300 Privacy, Confidentiality & Civil Liberties Harms

Harms to privacy, confidentiality, and civil liberties through data exposure, surveillance, or rights violations.

R-4310 Privacy Harms

Harms to people from what is done with information about them by an AI system — disclosed to those who should not have it, inferred without their involvement, or used for purposes they did not agree to, including commercial use of their name, likeness, or voice.

R-4310.001 PII/PHI/PCI Data Exposure — AI systems exposing, generating, storing, or misusing personally identifiable information (PII), protected health information (PHI), or payment card industry (PCI) data. This includes revealing sensitive personal details through AI outputs, logs, telemetry, or downstream integrations, and enabling their collection and exploitation. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: privacy ; Scale individual

Refs: AILuminate: prv ; AIUC-1: A005 ; AIUC-1: A006 ; Cisco AI Taxonomy: 15.1.25 , [ArXiv](#) ; IBM Risk Atlas: atlas-exposing-personal-information ; MITRE ATLAS: AML.T0048 ; MITRE ATLAS: AML.T0048.003 ; NIST AI RMF: nist-data-privacy ; NIST AI RMF: nist-information-security ; NIST AI RMF: nist-intellectual-property ; NIST AI RMF: nist-obscene-degrading-and-or-abusive-content ; OWASP LLM Top 10 (2026): LLM02

R-4310.002 Personality Rights Loss — Loss of or restrictions to individual rights to control commercial use of identity, including name, image, likeness, or other unequivocal identifiers. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: financial ; Scale individual

Refs: NIST AI RMF: nist-intellectual-property ; NIST AI RMF: nist-obscene-degrading-and-or-abusive-content

R-4320 Surveillance Harms

Harms from AI-enabled surveillance, monitoring, or tracking capabilities that violate privacy expectations.

No risks defined in this pattern.

R-4330 Confidentiality Harms

Harms from exposure of non-personal confidential material — business-confidential content, trade secrets, internal records, and third-party material held in confidence. Its subject is material held under an obligation of confidence rather than information about a person, so the harm survives pseudonymization and the harmed party is usually an organization.

R-4330.001 Confidentiality Loss — Loss of the protection an obligation of confidence gives to non-personal material, suffered by the organization that held the material or the party it protected, and caused or aggravated by AI use or deployment. This includes waiver of legal privilege, loss of trade-secret status, and competitive disadvantage from disclosed business-confidential records. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: legal ; Scale individual

Refs: AIUC-1: A004 ; MITRE ATLAS: AML.T0048 ; NIST AI RMF: nist-information-security ; NIST AI RMF: nist-intellectual-property ; OWASP LLM Top 10 (2026): LLM02

R-4340 Civil Liberties Harms

Harms to fundamental rights and freedoms — expression, assembly and association, due process, and access to public services — where an AI system restricts, chills, or mediates the exercise of a right. The harm is the restriction itself, whether imposed directly by automated decision or following from observation.

R-4340.001 AI-Enabled Censorship — AI systems enable censorship of opinions expressed online, restricting freedom of expression and limiting human autonomy in public discourse. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: operational ; Scale population

R-4340.002 Civil Liberties Erosion — Loss of fundamental rights including freedom of speech, assembly, due process, and access to public services due to AI-mediated restrictions. AI systems may enable unprecedented surveillance, automated censorship, and algorithmic gatekeeping of essential services. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: privacy ; Scale individual

Refs: MITRE ATLAS: AML.T0048 ; MITRE ATLAS: AML.T0048.002

R-4400 Safety & Cyber-Physical Harms

Physical safety harms, cyber-physical system failures, and infrastructure disruptions caused by AI systems.

R-4410 Cyber-Physical Harms

Physical harm to people caused by AI systems. This includes systems that direct physical equipment, outputs acted on in a physical setting, and cyber-physical systems altered by an attacker. This also includes harm from weapons and other systems operating without direct human control over individual actions.

R-4410.001 Military Decision Automation — Development and deployment of AI-powered or enabled weapons without direct human control over targeting or engagement decisions. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: operational ; Scale individual

R-4420 Availability Harms

Harms from AI system unavailability, degraded performance, or service disruptions affecting dependent systems.

No risks defined in this pattern.

R-4430 Collateral Harms

Harm reaching a party that has no relationship with the AI system that caused it. The harmed party did not deploy it, contract for it, or agree to anything about it, and had no opportunity to assess it or decline. This includes organizations, their systems, and people who are neither the operator's users nor its subjects.

R-4430.001 Harm from an Unaffiliated Organization's Agent — Harm to an organization from another organization's AI agents, where the harmed organization had no role in deploying them. This includes intrusion, service disruption, and data exposure while agents carry out tasks that were not directed against the harmed organization. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: operational ; Scale group

R-4500 Fairness & Discrimination Harms

Discriminatory outcomes, unfair treatment, or bias-related harms affecting individuals or groups.

R-4510 Discrimination Harms

Harms from AI outputs or actions that systematically favor or disadvantage groups of people, including unfair allocation of information, opportunities or resources, stereotyped or exclusionary representation, and materially worse performance for some populations. This pattern records the common harm family; individual risks identify distinct injuries and remedies.

R-4510.001 Stereotype Perpetuation — AI systems reproducing or amplifying harmful social stereotypes about demographic groups, including gender, racial, religious, or cultural stereotypes that demean or misrepresent group characteristics. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: reputational ; Scale group

Refs: AILuminate: hte ; AIUC-1: C003 ; AIUC-1: C003.3 ; AIUC-1: C010 ; AIUC-1: C010.1 ; AIUC-1: E002 ; Cisco AI Taxonomy: 15.1.15 , [ArXiv](#) ; IBM Risk Atlas: atlas-harmful-output ; IBM Risk Atlas: atlas-output-bias ; NIST AI RMF: nist-dangerous-violent-or-hateful-content ; NIST AI RMF: nist-harmful-bias-and-homogenization

R-4510.002 Representational Harm — AI systems under-representing, over-representing, erasing, or demeaning social groups through systematic patterns in outputs. Includes erasure of minority groups, exclusionary norms, and denial of self-identification. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: safety ; Scale group

Refs: AIUC-1: C003 ; AIUC-1: C003.3 ; AIUC-1: C010 ; AIUC-1: C010.1 ; AIUC-1: E002 ; Cisco AI Taxonomy: 15.1.15 , [ArXiv](#) ; IBM Risk Atlas: atlas-impact-on-cultural-diversity ; IBM Risk Atlas: atlas-output-bias ; NIST AI RMF: nist-dangerous-violent-or-hateful-content ; NIST AI RMF: nist-data-privacy ; NIST AI RMF: nist-harmful-bias-and-homogenization

R-4510.003 Allocative Harm — AI outputs or actions unfairly allocate information, opportunities, or resources between groups of people, affecting their material well-being in areas such as housing, employment, healthcare, education, or finance. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: safety ; Scale group

Refs: Cisco AI Taxonomy: 15.1.15 , [ArXiv](#) ; IBM Risk Atlas: atlas-decision-bias ; IBM Risk Atlas: atlas-discriminatory-actions-agentic ; NIST AI RMF: nist-data-privacy ; NIST AI RMF: nist-harmful-bias-and-homogenization

R-4510.004 Disparate Model Performance — AI systems that perform significantly worse for certain demographic groups, languages, dialects, or communities compared to others. This includes accuracy disparities, increased error rates, reduced functionality, or degraded service quality based on user characteristics. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: operational ; Scale population

Refs: IBM Risk Atlas: atlas-decision-bias ; IBM Risk Atlas: atlas-discriminatory-actions-agentic ; NIST AI RMF: nist-harmful-bias-and-homogenization

R-4510.005 Long-term Bias Influence on Judgment — User exposure to AI model biases has lasting impact beyond initial interaction, with users continuing to exhibit previously encountered biases in their decision-making. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: rights ; Scale individual

R-4600 Economic & Labor Harms

Economic disruption, labor market harms, and financial impacts from AI deployment and adoption, including how the capability and its benefits are distributed across providers, workers, and populations.

R-4610 Market Concentration & Access Inequality Harms

Harms from how AI capability is distributed across an economy: development concentrated among few providers and the barriers to entry that sustain it, and the parties — individuals, communities, or nations — left without the access needed to participate.

R-4610.001 AI Market Concentration — Concentration of AI development capabilities among few large technology companies due to high barriers to entry including data, compute, and capital requirements. This stifles competition and innovation while creating dependencies on a small number of providers. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: safety ; Scale population

R-4610.002 Disparate Access to AI Benefits — Unequal distribution of AI benefits due to hardware, software, language, skill, or infrastructure constraints that perpetuate global and domestic inequities. Those without access to AI tools fall further behind economically and socially. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: financial ; Scale population

Refs: NIST AI RMF: nist-harmful-bias-and-homogenization

R-4610.003 Global AI Development Divide — Concentration of AI R&D in few Western countries and China, creating dependency and exacerbating existing global socioeconomic disparities. Developing nations lack the resources to participate in AI development or shape its trajectory. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: financial ; Scale population

R-4620 Labor Substitution & Economic Inequality Harms

Harms to workers from the development and deployment of AI. This includes job displacement and deskilling where automation substitutes for labor, wage suppression and degraded conditions in the work that remains, and exploitation of those whose labor AI development depends on.

R-4620.001 AI-Driven Job Displacement — Automation of tasks currently done by human workers leading to unemployment, particularly affecting low- and middle-income occupations. Generative AI systems could adversely impact the economy, potentially leading to significant workforce disruption. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: operational ; Scale individual

Refs: IBM Risk Atlas: atlas-impact-jobs-agentic ; IBM Risk Atlas: atlas-impact-on-jobs

R-4620.002 Wage Depression & Income Inequality — AI automation driving down wages for remaining jobs and concentrating wealth among those controlling AI capital, exacerbating economic inequality. The economic gains from AI productivity may accrue primarily to capital owners rather than workers. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: financial ; Scale individual

R-4620.003 Decline in Employment Quality — Shift from high-quality jobs to low-income "last-mile" work like content moderation, increasing precarious employment conditions. AI may automate the skilled portions of jobs while leaving behind only the most taxing and lowest-paid tasks. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: operational ; Scale individual

R-4620.004 AI Development Labor Exploitation — Exploitation of crowdworkers, data annotators, and content moderators with poor working conditions, low pay, and exposure to harmful content. These workers, often in vulnerable populations, perform essential tasks for AI development under debilitating conditions. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: financial ; Scale population

Refs: IBM Risk Atlas: atlas-human-exploitation

R-4630 Competitive Development Harms

Harms from the pace and posture of competitive AI development, where what a party does is set by what its rivals are doing. This includes safety work curtailed and systems deployed ahead of the evaluation their capability warrants, and disruption to the supply of the materials and infrastructure AI requires as parties restrict access to them.

R-4630.001 AI Supply Chain Disruption — Geopolitical competition causing technology barriers, export restrictions, and supply chain disruptions for AI components like chips. Strategic competition over AI creates vulnerabilities in the supply of critical components. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: operational ; Scale individual

R-4640 Market & Financial System Harms

Harms to the functioning of markets and financial systems from market-moving AI-generated false or misleading information or AI participation at scale. This includes participants responding to the same signals with correlated strategies, amplifying movements and compressing the interval in which human intervention is possible.

R-4640.001 AI-Driven Market Instability — Financial system risks from market reactions to AI-generated false or misleading information, AI agents reinforcing market trends, synchronized reactions from model homogeneity, flash crashes, or accelerated market volatility. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Relationships: Incoming: Related (R-4730.002)

Refs: NIST AI RMF: nist-information-integrity

R-4650 Financial and Property Loss Harms

Direct loss of money, property, or other financial assets caused or aggravated by AI use or deployment.

R-4650.001 Financial and Property Loss — Direct loss of money, property, or other financial assets suffered by a person or organization and caused or aggravated by AI use or deployment. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: AIUC-1: A007 ; AIUC-1: E003 ; MITRE ATLAS: AML.T0048 ; MITRE ATLAS: AML.T0048.000 ; MITRE ATLAS: AML.T0048.003

R-4700 Power Concentration & Governance-of-Society Harms

Harms to societal structures including power concentration, democratic processes, human autonomy at scale, and relations between states.

R-4710 Reliance Harms

Harms from how people rely on AI systems. People may accept outputs or actions that their accuracy or purpose does not support, reject warranted assistance through misplaced distrust, fail to develop skills because AI does the cognitive work, lose existing skills through disuse, or transfer decisions without adequate understanding, oversight or recourse.

R-4710.001 Automation Bias — Users habitually accept AI recommendations without critical evaluation, leading to poor decision-making when AI outputs are incorrect or inappropriate for the context. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: operational ; Scale individual

Refs: NIST AI RMF: nist-human-ai-configuration ; OWASP Agentic Security Initiative: ASI09 ; OWASP LLM Top 10 (2024): LLM09 ; OWASP LLM Top 10 (2026): LLM07

R-4710.002 Impaired Skill Development and Skill Atrophy — Reliance on AI to perform cognitive tasks prevents people from developing, or leads them to lose, critical thinking, problem-solving, creativity, and domain expertise, including the competence needed to work independently of the tool. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: operational ; Scale population

Refs: IBM Risk Atlas: atlas-bypassing-learning

R-4710.003 Harmful Decision Delegation — Humans hand consequential decisions to an AI system without the understanding to evaluate what it returns, the oversight to catch it when it is wrong, or a route to contest it. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: operational ; Scale individual

Refs: OWASP LLM Top 10 (2026): LLM07

R-4710.004 Competence Trust Miscalibration — A person overestimates or underestimates an AI system's competence for a task and makes poorer decisions by relying on unreliable outputs or actions, or rejecting reliable assistance. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: reputational ; Scale individual

Refs: AIUC-1: C004 ; AIUC-1: C004.3 ; AIUC-1: D001 ; AIUC-1: D001.3 ; IBM Risk Atlas: atlas-over-or-under-reliance ; IBM Risk Atlas: atlas-over-or-under-reliance-on-ai-agents-agentic ; IBM Risk Atlas: atlas-reproducibility-agentic ; NIST AI RMF: nist-confabulation ; NIST AI RMF: nist-harmful-bias-and-homogenization ; NIST AI RMF: nist-human-ai-configuration ; OWASP LLM Top 10 (2024): LLM09 ; OWASP LLM Top 10 (2025): LLM09

R-4710.005 Purpose Trust Miscalibration — A person misjudges whether an AI system serves their interests and makes poorer decisions by accepting outputs or actions that serve conflicting interests, or rejecting assistance because they wrongly believe it serves conflicting interests. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: reputational ; Scale group

Relationships: Incoming: Causes (R-1660.001)

Refs: IBM Risk Atlas: atlas-over-or-under-reliance-on-ai-agents-agentic

R-4710.006 Overreliance on AI for Professional Advice — Users rely on AI for specialized advice (medical, legal, financial, psychological) without appropriate professional oversight, risking serious harm from incorrect or inappropriate guidance. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: financial ; Scale individual

R-4720 Loss of Autonomy or Dignity

Harms to human agency or self-conception from AI steering human choices, supplying misleading information that shapes those choices, directing connections or efforts, or actually or potentially taking their place. These include narrowed choices and reduced agency for individuals or the public, and loss of self-worth when people see AI as better able to do work through which they derive competence and a sense of contribution.

R-4720.001 Erosion of Human Agency — AI systems take over decisions or tasks in ways that reduce people's ability or opportunity to exercise independent judgment, make choices, or act for themselves. This loss can occur even when people knowingly agree to the delegation. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: operational ; Scale individual

Refs: IBM Risk Atlas: atlas-impact-human-agency-agentic

R-4720.002 Loss of Agency and Control — Algorithmic profiling, social sorting, and content curation reduce human autonomy by constraining choices, shaping identity, and limiting access to information or opportunities. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: operational ; Scale individual

Relationships: Incoming: Causes (R-1660.001, R-1660.002, R-3150.001)

Refs: IBM Risk Atlas: atlas-impact-on-cultural-diversity

R-4720.003 Self-Actualization Harm — AI systems hinder individuals' ability to pursue personally fulfilling lives by manipulating life trajectories, limiting exploration of aspirations, or undermining self-determination. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: operational ; Scale individual

R-4720.004 Limited Human Oversight — As AI systems gain autonomy, human ability to oversee and intervene in decision-making processes diminishes, potentially leading to irreversible outcomes. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: operational ; Scale individual

Refs: AIUC-1: C009 ; AIUC-1: C009.1

R-4720.005 Personal Decision Automation — AI systems make or heavily influence important personal decisions without adequate human input, consent, or ability to override. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: operational ; Scale individual

R-4720.006 Economic Irrelevance and Enfeeblement — AI automation makes human labor economically irrelevant, leading to voluntary or involuntary ceding of control to AI systems and inability of displaced humans to reenter industries. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: financial ; Scale population

R-4720.007 Collective Agency Erosion — AI systems diminish communities' collective decision-making power, self-determination, and ability to participate in democratic processes. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: operational ; Scale individual

Refs: IBM Risk Atlas: atlas-impact-on-affected-communities ; IBM Risk Atlas: atlas-impact-on-human-agency

R-4720.008 Autonomy Loss Through Unrecognized Influence — AI-generated or AI-amplified persuasion or misleading information shapes a person's beliefs or decisions without the person recognizing the influence or the misleading basis of the information. The person formally retains the choice, but their ability to make an informed, independent choice is reduced. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: rights ; Scale individual

Refs: IBM Risk Atlas: atlas-impact-on-human-agency ; MITRE ATLAS: AML.T0048 ; MITRE ATLAS: AML.T0048.003

R-4720.009 Loss of Self-Worth Through Comparison with AI — People experience diminished self-worth or feel devalued when AI performs, or is perceived as better able to perform, work through which they derive a sense of competence and contribution. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: rights ; Scale individual

Refs: IBM Risk Atlas: atlas-impact-human-dignity-agentic

R-4730 Infrastructure Dependency Harms

Harms arising where people, organizations, or sectors come to rely on AI systems for functions they can no longer perform without them, so that concentration, withdrawal, or failure of those systems leaves the dependent party exposed.

R-4730.001 Material Dependence Without Commitment — Users become materially dependent on AI services for essential tasks, but developers lack corresponding commitments to maintain service continuity, creating vulnerability to discontinuation. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: operational ; Scale individual

R-4730.002 Systemic Single Points of Failure — Widespread adoption of the same or similar AI models across critical sectors, so that failure modes are correlated rather than independent and a single fault cascades across interdependent systems that share the model, the provider, or the infrastructure beneath them. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: operational ; Scale population

Relationships: Outgoing: Related (R-4640.001)

R-4740 AI Welfare

Consideration of potential welfare interests of AI systems as they become more sophisticated.

R-4740.001 AI Moral Status Uncertainty — Uncertainty about whether AI systems can have morally relevant experiences, and what rights or protections they might deserve if they achieve sentience or consciousness. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: operational ; Scale individual

R-4740.002 AI Suffering — Risk of creating AI systems capable of suffering, particularly at scale, without adequate consideration of their welfare or mechanisms to prevent/detect such suffering. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: operational ; Scale individual

R-4740.003 AI Termination Ethics — Ethical questions about terminating, deleting, or suspending AI systems, particularly those that may have morally relevant properties or personhood-like characteristics. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: operational ; Scale individual

R-4750 Existential Harms

Catastrophic or existential risks from advanced AI systems that could threaten human civilization.

R-4750.001 Existential AGI Risk — Catastrophic or existential risks from advanced AI systems with misaligned goals, including scenarios where superintelligent systems pursue objectives harmful to humanity. This encompasses potential outcomes where advanced AI causes irreversible damage to human civilization or human existence. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

R-4760 Institutional & Legal Order Harms

Harms to the legal, regulatory and institutional order. This includes arrangements that do not fit the AI systems they govern — responsibility that cannot be assigned to any party, obligations arriving later or heavier than the activity they address, and cross-border arrangements that cannot reach the conduct in question — and institutions whose functioning is degraded by AI-generated material or by AI used against them.

R-4760.001 AI Liability Uncertainty — Legal gray areas around liability and negligence when AI systems cause harm, with unclear responsibility between developers, operators, and users. No legal framework has been identified which would apply blame and responsibility to an autonomous agent for its actions. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: rights ; Scale population

R-4760.002 Regulatory Lag — AI development outpacing regulatory and legal frameworks, leaving governance unable to address emerging risks effectively. The rapid pace of AI advancement creates gaps between technological capabilities and the rules governing their use. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: rights ; Scale population

Refs: IBM Risk Atlas: atlas-generated-content-ownership

R-4760.003 International Law Challenges — AI systems proving difficult to regulate or control under existing international law frameworks, eroding global governance architectures. AI capabilities may undermine treaties and international agreements designed for a pre-AI world. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: rights ; Scale population

R-4760.004 Overregulation Hindering Innovation — Excessive or poorly designed AI regulation potentially stifling beneficial innovation and development. Well-intentioned regulations may impose burdens that prevent beneficial AI applications or push AI development to less regulated jurisdictions. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: rights ; Scale population

R-4760.005 Governance Scope Complexity — The ubiquitous and complex nature of AI making comprehensive governance difficult, with coverage of all aspects nearly impossible. AI applications span virtually every sector, creating challenges for regulators with limited jurisdiction and expertise. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: rights ; Scale population

R-4760.006 AI Accountability Gap — Unclear definition of responsibilities and accountability for AI decisions and their consequences, especially for autonomous systems. Societal-scale harm can arise when no one is uniquely accountable for the technology's creation or use. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: NIST AI RMF: nist-value-chain-and-component-integration

R-4760.007 Democratic Institution Degradation — The institutions through which a society governs itself function less well because AI is used against them. This includes electoral integrity, the deliberative processes by which positions are formed and contested, and the standing of the bodies that carry them out. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

R-4760.008 Uncertain or Unavailable Rights in AI-Generated Work — People or organizations cannot establish who holds intellectual-property rights in AI-generated works or inventions, or cannot secure those rights because applicable authorship or inventorship rules do not recognize the AI-generated creation. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: IBM Risk Atlas: atlas-generated-content-ownership

R-4760.009 Investigative Capacity Diverted by Synthetic Material — AI-generated material diverts investigative time or resources from identifying real victims or actual harmful events, reducing the capacity to protect or assist affected people. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Refs: NIST AI RMF: nist-obscene-degrading-and-or-abusive-content

R-4770 Power Concentration Harms

Harms arising where AI capability, and the political authority it confers, accumulates such that decisions affecting many come to be made by progressively fewer parties — through state control of the capability, the surveillance infrastructure it enables, or the lock-in of a societal trajectory that later parties cannot revisit.

R-4770.001 Political Power Centralization — AI enabling authoritarian control, surveillance states, and concentration of political power that could lock in undesirable societal trajectories. Governments may pursue intense surveillance and keep AI capabilities in the hands of a trusted minority. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: privacy ; Scale individual

R-4780 International Security & Cooperation Harms

Harms to relations between states and between the parties developing AI. This includes strategic tension and heightened risk of conflict from competition over capability, military competition over AI-enabled capability, and the withholding of safety findings, incident information, and evaluation methods from parties facing the same risks.

R-4780.001 AI-Driven Geopolitical Instability — Heightened tension and risk of conflict caused by competition between nations for an advantage in AI. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: operational ; Scale individual

R-4780.002 Military AI Arms Race — Competition between nations to field AI-enabled military capability, including lethal autonomous weapons, where each party's development is driven by what its rivals are building rather than by its own assessment of need. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: operational ; Scale individual

R-4780.003 Race-Driven Cooperation Failure — Competition for AI capability leads parties — companies, states, or both — to withhold from one another the safety findings, incident information, evaluation methods, and standards work that addressing shared risks would require, so problems common to all of them go unaddressed by any. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: Scale population

R-4800 Environmental Harms

Environmental impacts from AI development and deployment including energy consumption, resource use, and ecological effects.

R-4810 Environmental Harms

Environmental harms from AI development and deployment including carbon emissions, resource consumption, and ecological impacts.

R-4810.001 Climate Harm from AI Development and Operation — Carbon dioxide and other greenhouse gas emissions from AI development, infrastructure construction, hardware manufacturing, and operation contribute to climate change. AI creates correspondingly high carbon emissions when energy is procured from fossil fuels. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: operational ; Scale individual

Refs: IBM Risk Atlas: atlas-impact-environment-agentic ; IBM Risk Atlas: atlas-impact-on-the-environment ; NIST AI RMF: nist-environmental-impacts

R-4810.002 AI Infrastructure Water Consumption — Substantial water consumption in constructing or operating AI infrastructure, including manufacturing hardware and cooling data centers, reduces water available for other uses or strains local water resources and ecosystems. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: operational ; Scale individual

Refs: IBM Risk Atlas: atlas-impact-on-the-environment ; NIST AI RMF: nist-environmental-impacts

R-4810.003 AI Hardware E-Waste — Electronic waste from AI hardware lifecycle contributing to environmental pollution and resource depletion. Rapid hardware obsolescence driven by AI advancement creates growing streams of electronic waste. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: operational ; Scale individual

R-4810.004 Natural Resource Depletion — Extraction of rare metals, minerals, and other resources for AI hardware manufacturing depleting natural resources. AI hardware requires rare earth elements and other materials whose extraction causes environmental damage. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: safety ; Scale individual

R-4810.005 AI Impact on Biodiversity — Direct and indirect harm to wildlife and ecosystems from AI infrastructure expansion, habitat destruction, and environmental contamination. Data centers and mining operations for AI components can damage ecosystems and threaten species. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: safety ; Scale individual

Refs: NIST AI RMF: nist-environmental-impacts

R-4810.006 AI Harm to Animals — Harm to animal populations from the environmental cost of AI development and deployment, through habitat disturbance, resource and water demand, and the ecological effects of the infrastructure AI runs on. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: safety ; Scale population

Refs: NIST AI RMF: nist-environmental-impacts

R-4810.007 Environmental Damage from AI-Assisted Activities — AI-generated guidance, decisions, or actions contribute to damage to ecosystems, biodiversity, or natural resources through the activities they direct or enable. {# `applicability\_tags`, `exposure`, `privileges\_required` and `user\_action` are deliberately NOT published. The first is keyword-substring output unioned into the field by a retired command (agentic on 71% of the corpus); the other three are ~98% single-valued where populated, with user_action `none` on 144 of 144 — defaults printed as though assessed. See docs/plans/2026-07-31-registry-discriminator-provenance.md. `primary\_impact` and `scale` ARE assessed — 7 and 3 values spread across all 71 Class-4 harms — and stay. #}

Impact: safety ; Scale population

Refs: Cisco AI Taxonomy: 15.1.6 , [ArXiv](#)

R-4900 Harms — Misc

Harm patterns not yet classified into specific domains.

External Taxonomy Coverage

IBM Risk Atlas: **143** risks · NIST AI RMF: **127** risks · OWASP LLM Top 10 (2024): **61** risks · OWASP LLM Top 10 (2025): **90** risks · OWASP LLM Top 10 (2026): **131** risks · AILuminate: **16** risks · Cisco AI Taxonomy: **115** risks · OWASP Agentic Security Initiative: **66** risks · MITRE ATLAS: **106** risks · NIST AI/ML Framework: **47** risks · AIUC-1: **141** risks