

AI Risk vs Security Risk Matrix

DocID: ODA3-2026-07-CHT-SEC-004
Classification: Public

CHT-SEC-004: AI Risk vs Security Risk Matrix

ODA3 Institute | Practitioner Cheat Sheet

Field	Value
Classification	Public — Practitioner Reference
Intended Audience	CISOs, Security Architects, AI Governance Leads, Compliance Officers
Evidence Confidence	Moderate (T2–T3 weighted; see tier tags)
Review Cycle	6 months
Framework Cross-References	Maps to UAIF™ v1.0 (classification), GAISSE™ v1.0 (assurance)
Related Publications	CHT-SEC-001, CHT-SEC-002, CHT-SEC-003

GAISSE™, UAIF™, and AI-IRF™ are frameworks of ODA3 Institute, referenced under the GAISSE Ecosystem License (GEL) v1.0.

Scope note: This cheat sheet centers on the AI Risk ↔ AI Security Risk distinction and the incident-routing decisions that follow from it. Section 1 situates both inside the wider enterprise risk landscape (privacy, operational, regulatory, business, third-party, supply chain, reputational risk), but those adjacent categories are treated at reference depth only, not full depth.

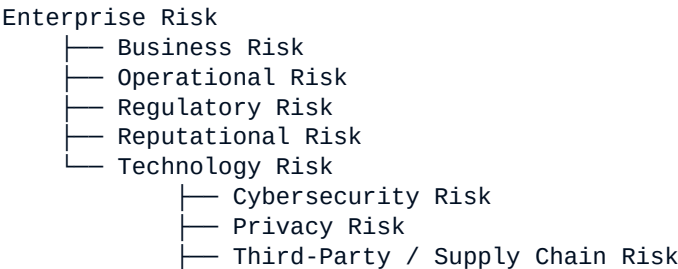
Why This Matters

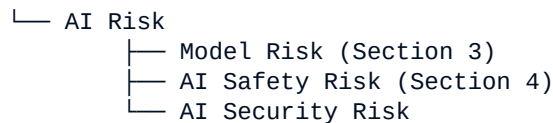
Security teams and AI governance teams often use "risk" to mean different things — and route incidents to different owners as a result. This causes real operational drag: incidents get triaged twice, escalation paths stall, budgets get misallocated, and nobody owns the overlap zone. Practitioners who cannot cleanly separate AI risk (what the system does, or fails to do) from AI security risk (what is done to the system) tend to either bolt bias and drift problems onto a firewall-shaped solution, or dismiss data poisoning as a "model quality" issue that never reaches the SOC.

This cheat sheet gives practitioners a fast, evidence-tiered reference for telling the two risk types apart, placing them in enterprise context, mapping them to owners and controls, and recognizing where they collapse into one.

1. Where AI Risk and AI Security Risk Sit in the Enterprise

AI Risk is not a standalone category — it is a subset of Technology Risk, which is itself one branch of Enterprise Risk. AI Security Risk, in turn, is a subset of AI Risk. Collapsing these layers is a common source of the routing disputes this cheat sheet addresses.





Two adjacent categories deserve a one-line placement, without expanding scope beyond this cheat sheet's focus:

- **Privacy Risk** — inappropriate collection, exposure, or retention of personal data through an AI system. Frequently intersects with AI Security Risk (e.g., prompt leakage exposing PII) but is governed separately by privacy legislation.
- **Third-Party / Supply Chain Risk** — exposure introduced by foundation model vendors, hosted inference APIs, or upstream datasets. Frequently intersects with AI Security Risk (a poisoned upstream model is both a supply-chain and a security event) but carries its own vendor-governance track.

2. Two Risk Lineages, Not One

AI Risk (the umbrella category): The complete set of risks arising from the development, deployment, governance, and use of AI systems — including bias, drift, hallucination, misuse, opacity, safety failures, and business/strategic exposure — independent of whether an adversary is involved. This lineage draws primarily from ML research and governance frameworks (NIST AI RMF, ISO/IEC 42001, EU AI Act) rather than CVE-style taxonomy.

AI Security Risk (a specialized subset): Risk arising specifically from adversarial exploitation, compromise, abuse, or unauthorized manipulation of AI systems — prompt injection, data poisoning, model theft, model extraction, adversarial examples, and infrastructure-level attacks routed through an ML pipeline. This lineage draws from MITRE ATLAS, OWASP's LLM-specific application security guidance, and traditional cybersecurity frameworks extended to model-specific attack surfaces.

The practitioner error, stated plainly: Treating all AI risk as "security risk with a chatbot in it," or conversely treating security incidents inside AI systems as purely a governance paperwork exercise. Both misclassifications misroute the incident and misassign the fix.

Dimension	AI Risk	AI Security Risk
Primary concern	What the system does (or fails to do)	What others do to the system
Origin	Internal — model behavior, data, design	External — adversarial actors, vulnerabilities
Nature	Often non-adversarial, probabilistic	Typically adversarial, intentional
Representative examples	Bias, drift, hallucination, misuse, opacity	Prompt injection, model extraction, poisoning, model theft
Governing frameworks (illustrative)	NIST AI RMF, ISO/IEC 42001, EU AI Act	MITRE ATLAS, OWASP LLM-focused guidance
Typical remediation	Retraining, governance review, guardrail redesign	Incident response, containment, patching, credential revocation

3. Model Risk vs AI Security Risk

Model Risk is frequently confused with AI Security Risk, particularly in regulated sectors (banking, insurance) that already run formal Model Risk Management programs. The two require different expertise and often different owners.

Model Risk	AI Security Risk
Focuses on model performance and reliability	Focuses on malicious compromise
May occur without any attacker involved	Requires an adversarial or abuse scenario
Representative issues: drift, bias, poor calibration, hallucination	Representative issues: prompt injection, data poisoning, model theft, model extraction
Managed through validation, benchmarking, and monitoring	Managed through security controls and adversarial/red-team testing
Primarily a statistical discipline	Primarily a security discipline

Worked example: A model's predictive accuracy gradually declines because underlying market conditions shifted. That is Model Risk — no attacker, no compromise. If an attacker instead manipulates training data specifically to degrade that same accuracy, the resulting issue is AI Security Risk, because malicious intent is now present. The observable symptom (declining accuracy) can look identical; the triage question is why it's declining.

4. AI Safety vs AI Security

Safety and security are frequently discussed together but answer different questions:

- **AI Safety** asks: will the system cause harm even when no one is attacking it?
- **AI Security** asks: can the system be intentionally compromised, manipulated, or abused?

Dimension	AI Safety	AI Security
Primary concern	Prevent unintended/accidental harm	Prevent malicious compromise
Threat source	Design flaws, uncertainty, edge cases	Attackers, insiders, malicious users
Typical failure	Unsafe recommendation or output	Prompt injection or extraction attack
Root cause	Insufficient guardrails, poor design	Adversarial manipulation
Primary controls	Guardrails, human oversight, output validation	Authentication, authorization, adversarial defenses
Assurance activity	Safety validation, scenario testing	Security assessment, red teaming

Worked example: An autonomous system recommends an unsafe action because of a reasoning error — that is primarily a Safety Risk. If an attacker deliberately crafts a prompt to force that same unsafe recommendation, the incident becomes an AI Security Risk with downstream safety consequences. As with Model Risk, the same output can sit in either category depending on whether intent and manipulation are present.

5. The Quadrant Matrix

	Low AI-Behavioral Novelty	High AI-Behavioral Novelty
Low Adversarial Involvement	Quadrant A — Classic IT security (patching, access control, network defense)	Quadrant B — Model quality/governance issues: hallucination, bias, drift, unsafe output with no attacker present
High Adversarial Involvement	Quadrant C — AI infrastructure attacked via conventional means: API key theft, container escape into an ML pipeline	Quadrant D — AI-native attack surface: prompt injection, data poisoning, model extraction, adversarial examples

Quadrant D is where most "is this security or AI governance" disputes originate, because the exploit target is the model's behavior, not a conventional software flaw. Quadrant B is where organizations most often misroute — treating a non-adversarial governance failure (Model Risk, Section 3; Safety Risk, Section 4) as if it required a security incident response.

6. Overlap Zone Detail: Quadrant D

Attack Pattern	Traditional Security Analogue	Why It's Different
Prompt injection	Input validation / injection (SQLi-like)	Exploits the model's instruction-following behavior rather than a parser flaw — no malformed syntax is required, because natural language itself is the payload [T3 — controlled academic demonstrations; a production instance was disclosed via a June 2025 zero-click Microsoft 365 Copilot data-exfiltration vulnerability, T2 given vendor and researcher disclosure]
Data poisoning	Supply-chain compromise	Corruption can be introduced at multiple stages — initial training, fine-tuning, retrieval corpora used by RAG systems, or reinforcement-learning feedback — and is often invisible until inference-time anomalies surface [T2 — documented in academic literature and vendor red-team reports]
Model extraction / theft	IP theft via exfiltration	Extraction can occur through legitimate API query patterns alone, with no unauthorized-access event to detect [T3]
Adversarial examples	Malformed input / fuzzing	Perturbations are often imperceptible to humans, evading anomaly detection tuned for structured data [T2]
Supply-chain compromise (model/data layer)	Traditional software supply-chain attack	Compromise can enter through a pretrained model or dataset rather than a code dependency, and detection tooling for the latter often doesn't cover the former [T2]

A structured adversary-technique taxonomy exists for this exact overlap zone — MITRE's ATLAS knowledge base catalogues adversarial tactics and techniques specific to AI systems [T2 — vendor/community-maintained knowledge base, cross-referenced against public incident writeups], serving a comparable function to MITRE ATT&CK for conventional infrastructure. ODA3 references ATLAS as an example of correct precedent for open, identifier-level taxonomy — consistent with ODA3's own open/controlled asset boundary for UAIF™.

7. Framework Landscape (Reference, Not Endorsement)

The frameworks below are cited for practitioner orientation only. ODA3 Institute is not affiliated with NIST, ISO, IEC, IEEE, MITRE, or OWASP, and this table does not constitute an assertion that any ODA3 framework is certified, compliant, or equivalent to any listed standard.

Framework	Primary Domain	Type	Notes
NIST AI RMF	AI risk (broad)	Voluntary risk-management operating model organized around Govern, Map, Measure, Manage functions [T2]	Strategic / governance-oriented
ISO/IEC 42001	AI risk (broad)	Certifiable AI management system standard; risk assessment addressed under Clause 6.1, with methodology left to organizational discretion [T2]	The only framework here that is third-party certifiable
EU AI Act	AI risk (broad, regulatory)	Binding EU regulation requiring documented risk management for high-risk applications [T2 — verify current enforcement posture against EU sources]	Not certifiable; conformity assessment required for in-scope systems
MITRE ATLAS	AI security risk	Non-binding adversarial technique taxonomy for AI systems [T2]	Technical/tactical, not governance-oriented
OWASP (LLM-focused guidance)	AI security risk	Non-binding application-security taxonomy for LLM-based systems [T3 — community-maintained, versioned frequently]	Complements ATLAS at the application layer

Practitioner takeaway: governance-oriented frameworks (NIST AI RMF, ISO 42001, EU AI Act) answer "what risk does this system create," while security-taxonomy frameworks (ATLAS, OWASP) answer "how is this system being attacked." Neither substitutes for the other, and neither substitutes for an organization's own risk register.

8. Governance Controls vs Technical Controls

Organizations frequently overinvest in one and underinvest in the other. Both are required, and they answer different questions.

Governance Controls	Technical Controls
Policy and risk appetite	Authentication and authorization
Decision rights and accountability	Encryption
Committees and approval processes	Network and model access controls
Risk acceptance decisions	Logging and monitoring
Internal audit	Automated enforcement, alerting

Governance determines: what risks are acceptable, who is accountable, which controls are mandatory, and how exceptions are handled. Technical controls implement those decisions inside operational systems. A mature AI security program needs both — a well-governed but technically uncontrolled system is exposed; a technically hardened but ungoverned system has no accountable owner when something goes wrong.

9. Fast Routing Guide

- **Attacker present + exploits code/infra** → Security incident response (Quadrant A/C)
- **Attacker present + exploits model behavior** → Joint AI-security incident response (Quadrant D) — route to AI-IRF™-aligned process
- **No attacker + model behaves badly** → AI governance/quality issue (Quadrant B) — route to model risk owner, not SOC
- **Uncertain which applies** → Default to joint triage; misrouting Quadrant D incidents to pure security teams, or Quadrant B issues to pure security teams, is a commonly observed contributor to delayed resolution in public incident post-mortems [T3]

Decision tree (condensed)

Is there an adversary manipulating the system?

NO → Is the failure a behavioral/output problem (bias, drift, hallucination)?

YES → AI governance issue → Business/Model Risk Owner + AI Governance Lead

NO → Standard operational issue → Engineering/Operations

YES → Is the exploit target the model's behavior itself (prompt injection, poisoning, extraction)?

YES → Joint AI-security incident (Quadrant D) → CISO + AI Governance Lead

jointly

NO → Conventional security incident routed through AI infra (Quadrant C) →

CISO

10. Risk Ownership at a Glance

Ownership should track decision-making authority, not just technical familiarity — the function that can authorize, fund, or accept the risk should own it.

Risk Type	Primary Owner	Key Supporting Functions
AI security risk (Quadrant C/D)	CISO	Engineering, AI Governance Lead, Legal
AI governance/behavioral risk (Quadrant B)	Business Owner / Model Risk Owner	AI Governance Lead, Engineering
Prompt injection	CISO	Engineering, AI Governance
Hallucination affecting a business decision	Business Owner	AI Governance, Engineering
Personal data leakage through an AI system	Privacy Office	CISO, Legal, Compliance
Model drift	Business Owner	Engineering, AI Governance
Poisoned training data	CISO	Engineering, AI Governance
Regulatory investigation	Compliance	Legal, Business Owner
Vendor/third-party model compromise	CISO	Procurement, Vendor Management

Notably absent from most organizational RACI models: a joint escalation path for Quadrant D incidents. Most incident-response runbooks assign a single owner by default (usually the SOC), which is precisely the pattern that produces misrouting in the overlap zone.

11. Illustrative Security Controls by Risk Type

This table is intentionally non-exhaustive and control-family level, not a certification checklist.

Risk Type	Representative Controls
AI security risk	Prompt-injection defenses, model access control, output filtering, model integrity verification, adversarial/red-team testing
AI governance/behavioral risk	Bias evaluation tooling, drift/calibration monitoring, human-in-the-loop review for high-stakes outputs, retraining triggers
Overlap (Quadrant D)	Combined: adversarial testing and behavioral monitoring on the same asset, since a single exploit (e.g., data poisoning) can simultaneously degrade accuracy and constitute a security event

Practical principle: security controls primarily reduce the likelihood of compromise. They do not, by themselves, resolve governance, legal, ethical, or strategic exposure — and vice versa. A model can pass every adversarial red-team test and still produce biased or unsafe outputs with no attacker involved [T2].

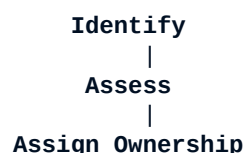
12. Detection Signal Differentiation

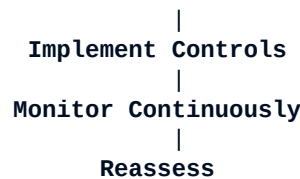
Indicator Type	AI Risk Signal	AI Security Risk Signal
Performance	Accuracy degradation with no known trigger	Accuracy degradation correlated with an identifiable attack pattern
Input	Distribution shift in incoming data	Delimiter-injection or known adversarial-pattern sequences
Output	Hallucination, biased output	Unexpected content consistent with data leakage or extraction
Behavior	Policy violation with no external trigger	Unauthorized tool invocation or credential misuse
Data	Training data statistically shifting	Training data source unverified or provenance broken

Compound signal example: a training-pipeline alert showing both distribution deviation and an unverified data source should be triaged as both AI risk and AI security risk simultaneously, not routed to a single owner [T3 — pattern observed across vendor red-team writeups, not a formally standardized detection rule].

13. After Triage: The Continuous Risk Lifecycle

Routing an incident to the right owner is the beginning of risk management, not the end. AI risk — in both lineages — evolves through retraining, data changes, and new attack techniques, so a one-time assessment does not provide lasting assurance.





Residual risk: the risk remaining after controls are implemented. It is expected, not a sign of failure — but it should be documented, assigned an accountable owner, and reviewed on a defined cycle rather than assumed to be zero.

Risk acceptance: a governance decision, not a technical one. An organization may reasonably accept a residual risk when further mitigation isn't cost-effective or technically feasible — provided the acceptance is documented, has a named accountable owner, and does not bypass a legal or regulatory obligation.

14. Common Organizational Mistakes

- **Treating AI risk as purely a cybersecurity problem.** Security teams often become default owners for all AI-related concerns; appropriate for Quadrant C/D, but this overlooks governance, ethical, and business dimensions that live in Quadrant B.
- **Equating Model Risk with AI Security Risk.** A poorly performing model is not necessarily a compromised one, and a secure model can still produce inaccurate or harmful output (Section 3).
- **Assuming compliance equals security.** Passing a regulatory assessment does not demonstrate adversarial resilience, and a technically secure system can still violate a regulatory obligation.
- **Assigning all AI ownership to one function.** No single function has the authority or expertise to manage every dimension of AI risk; this is a commonly reported contributor to Quadrant D misrouting [T3].
- **Treating AI risk as static.** Models drift through retraining, data changes, and new attack techniques; a one-time assessment does not provide lasting assurance (Section 13).

15. Real-World Incident Reference

These are public, previously disclosed incidents included only to illustrate the classification logic in this cheat sheet — not to assess the organizations involved. Each is labeled by primary risk type per the framework above; several sit closer to the overlap zone than the label alone suggests.

Incident (public, disclosed)	Primary Risk Type	Root Cause Pattern	Practitioner Lesson
Airline customer-service chatbot issuing an incorrect policy commitment that was later held binding (2024) [T2 — widely reported tribunal decision]	AI Risk (Safety / Governance)	Non-adversarial hallucination treated as a routine software bug rather than a governance and legal exposure	Safety and reliability failures need legal and operational sign-off, not just a patch — Section 4
Zero-click prompt-injection vulnerability in a major enterprise Copilot-style assistant enabling data exfiltration via email content (June 2025) [T2 — vendor and researcher disclosure]	AI Security Risk (Quadrant D)	Insufficient input sanitization and missing agent-action monitoring in an agentic workflow	Indirect prompt injection requires the same monitoring rigor as any other remote exploitation path — Section 6
Employees pasting proprietary	AI Security Risk	No network egress	Shadow AI is a security

Incident (public, disclosed)	Primary Risk Type	Root Cause Pattern	Practitioner Lesson
source code into a public LLM interface (2023) [T2 — widely reported]	(Shadow AI)	control or acceptable-use enforcement for unauthorized external AI tools	control gap before it is a governance one — enforce at the network boundary
Automated home-valuation model materially overestimating property values, contributing to large realized losses before the program was shut down (2021) [T2 — company disclosure and financial reporting]	AI Risk (Model Risk)	Insufficient continuous monitoring of model performance against a volatile, shifting environment	Model drift in a live financial process is an operational and Model Risk failure, not a security incident — Section 3

16. Threat Actor Taxonomy vs Risk Type

Not every risk source is an attacker. This taxonomy separates actors and conditions that primarily generate AI Risk from those that primarily generate AI Security Risk — useful when a Quadrant B/D dispute needs a quick sanity check on intent.

Actor / Source	Primary Motivation	Typically Drives	Illustrative Vector
Nation-state or organized APT group	Intelligence gathering, disruption	AI Security Risk	Model extraction, training-data exfiltration, supply-chain compromise
Cybercriminal	Financial gain	AI Security Risk	Prompt injection for data exfiltration, credential theft via AI-integrated tools
Malicious insider	Retaliation, personal gain	AI Security Risk	Model tampering, unauthorized data export
Negligent insider	Convenience, unawareness of policy	Both — commonly mislabeled as pure security	Shadow AI usage, accidental exposure of confidential data to external models
Automated attacker / bot	Scalable, opportunistic exploitation	AI Security Risk	Automated jailbreak attempts, volumetric prompt-injection probing
No actor — dynamic data environment	N/A	AI Risk (Model Risk / Safety)	Concept drift, distributional shift, degraded calibration over time
No actor — regulatory or legal environment	Compliance enforcement	AI Risk (Regulatory)	Non-conformity with a documented risk-management obligation

Reading this table: the bottom two rows have no actor at all — they are included deliberately, because the most common triage error is searching for an attacker where the actual driver is a dynamic environment or a compliance obligation.

17. Assessment Readiness Checklist

A self-assessment aid for organizations preparing to formalize their AI Risk vs AI Security Risk governance. Not a certification checklist — treat gaps as inputs to a roadmap, not as findings against a standard.

- ☐ AI systems are formally inventoried, with an accountable business owner per system
- ☐ AI Risk and AI Security Risk are explicitly defined as distinct categories in the organization's risk taxonomy
- ☐ Both risk types appear as separate, labeled entries in the enterprise risk register — not only in a security-only register
- ☐ A named function (e.g., AI Governance Lead) owns AI Risk distinct from the CISO's ownership of AI Security Risk
- ☐ Model validation / Model Risk Management process exists independent of security testing
- ☐ Adversarial or red-team testing is performed on AI systems classified as high-exposure
- ☐ Sociotechnical or bias/fairness assessment is performed independent of security testing
- ☐ Detection coverage exists for both behavioral drift (AI Risk) and adversarial patterns (AI Security Risk)
- ☐ Incident response runbooks route Quadrant B and Quadrant D incidents differently
- ☐ Residual risk is documented with a named accountable owner and review date
- ☐ A formal risk-acceptance process exists and is distinct from simply not fixing an issue
- ☐ Third-party/vendor AI models undergo a security review and a separate behavioral/output review
- ☐ Executive or Board reporting distinguishes AI Risk exposure from AI Security Risk exposure
- ☐ Continuous monitoring — not an annual point-in-time review — covers both risk types
- ☐ Regulatory mapping exists for both risk types (e.g., broad AI governance obligations vs. security-specific obligations)

18. Maturity Indicators

A rough self-assessment scale for how well an organization currently separates and governs the two risk types. Intended as a conversation-starter, not a scored instrument.

Level	Observable Characteristics
Level 0–1: Initial	AI Risk and AI Security Risk are confused or undifferentiated; risk assessments are ad hoc; no distinct ownership
Level 2: Managed	A written distinction exists between the two risk types, but it isn't consistently applied across risk registers or incident response
Level 3: Defined	Clear ownership assigned per Section 10; risk registers carry both categories separately; routing logic (Section 9) is documented and used
Level 4: Quantitative	Both risk types are monitored continuously with defined indicators (Section 12); residual risk and acceptance decisions are tracked systematically

19. Quick Reference: Question → Domain → Owner

Triage Question	Primary Risk Domain	Typical Owner
Can an attacker manipulate the AI system's behavior?	AI Security Risk (Quadrant D)	CISO + AI Governance Lead
Is the model becoming less accurate over	Model Risk	Business Owner / Model

Triage Question	Primary Risk Domain	Typical Owner
time with no attacker involved?		Risk Owner
Could a person be harmed by an unsafe output?	AI Safety Risk	Business Owner + AI Governance Lead
Does this expose personal information?	Privacy Risk	Privacy Office
Is a vendor or upstream model introducing unacceptable exposure?	Third-Party / Supply Chain Risk	CISO + Procurement / Vendor Management
Does this violate a documented regulatory obligation?	Regulatory & Compliance Risk	Compliance + Legal
Is this a conventional infrastructure attack that happens to touch an AI system?	AI Security Risk (Quadrant C)	CISO
Is the model producing biased or unfair outcomes with no attacker involved?	AI Risk (Governance / Ethical)	AI Governance Lead + Business Owner

20. Five Priority Actions

- **1. Establish a common AI Risk taxonomy.** Define AI Risk, AI Security Risk, Model Risk, and Safety Risk consistently across governance and security functions so incidents stop being triaged twice under different names.
- **2. Assign explicit, non-overlapping ownership.** Name an accountable owner for Quadrant B and a separate one for Quadrant D, with a documented joint path for genuine overlap cases.
- **3. Integrate both risk types into enterprise risk reporting.** Neither should live solely in a security-only register or a siloed AI ethics document — both belong in the same enterprise view (Section 1).
- **4. Build continuous, not point-in-time, monitoring.** Behavioral drift and adversarial activity both evolve after deployment; an annual assessment cannot substitute for ongoing detection (Sections 12–13).
- **5. Establish independent assurance for both domains.** Adversarial testing validates AI Security Risk controls; it does not validate fairness, calibration, or safety — commission both, from independent reviewers where feasible.

21. Notably Absent

This cheat sheet does not include:

- Proprietary incident telemetry or client-derived data (ODA3 Institute, founded March 2026, holds no such datasets)
- Quantified likelihood or frequency estimates for any quadrant — public disclosure volume is not a reliable proxy for actual incidence rate
- Vendor-specific tooling recommendations
- Legal or regulatory determinations of liability across quadrants — this is a triage aid, not a compliance ruling
- Full-depth treatment of Operational, Business, Regulatory, Privacy, Third-Party, and Supply Chain Risk — these are placed in enterprise context (Section 1) but not developed at the depth of Model Risk, Safety, and Security in this publication
- A joint-ownership RACI template for Quadrant D incidents — organizations should build one

- locally rather than adopt a generic template, since escalation authority varies by org structure
- A scored or weighted maturity model — Section 18 is presented as a self-assessment conversation starter, not a validated scoring instrument
- Named organizations behind the incidents in Section 15 beyond what is already public record — this cheat sheet draws only on previously disclosed, publicly reported events

This cheat sheet maps to UAIF™ v1.0 classification categories and GAISSE™ v1.0 assurance guidance. It does not constitute certification, compliance confirmation, or endorsement under any referenced framework.

ODA3 Institute — Where AI Governance meets Operational Reality.