AI SECURITY INCIDENT & RISK ANALYSIS

Q2 2026 AI Security Technical & Compliance Report

AI Security Is Moving from Model Outputs to Authorised Actions

Software & model supply-chain compromise · the collapsing patch gap · adversary techniques · regulatory readiness

TECHNICAL & COMPLIANCE EDITION

Field	Detail
Document ID	ODA3-2026-07-TCR-SEC-004
Status	Final
Edition	Technical & Compliance
Reporting period	1 April – 30 June 2026
Evidence cut-off	30 June 2026
Publication	July 2026
Companion	Q2 2026 AI Security Executive Brief (ODA3-2026-07-EXB-SEC-005)
Publisher	ODA3 Institute
Classification	Public analytical briefing
Distribution	Approved for public release

For: Chief Information Security Officers · Security Architects · AI Governance Leads · Compliance Officers · incident responders · assessors · standards participants.

In this report, *AI* means an artificial-intelligence system and, where the context is an agentic deployment, the model together with the identity, data, tools and execution authority that surround it.

Important Notice

This report is an intelligence analysis based on publicly available information available to ODA3 Institute as of 30 June 2026. It does not constitute legal, regulatory, financial, audit, certification or compliance advice. The control outcomes, detection logic, financial estimates and response priorities represent ODA3 Institute's analytical assessment; they do not guarantee security outcomes or regulatory compliance. Organisations should obtain advice appropriate to their jurisdiction, sector, systems, operating model and circumstances. All financial figures are ODA3 modelled estimates using conservative benchmark assumptions; no single aggregate quarter loss is published.

Evidence basis. The flagship records in this report are corroborated by primary and authoritative public sources — including a U.S. Cybersecurity and Infrastructure Security Agency (CISA) advisory and Known Exploited Vulnerabilities (KEV) listing, a GitHub security advisory, assigned Common Vulnerabilities and Exposures (CVE) identifiers, U.S. Commerce Department action, a filed U.S. class action, and major independent outlets. Sources are listed in Appendix C. Where a claim rests on a single or secondary source, or on a controlled research finding, this is stated and tiered accordingly.

Contents

PART I — ANALYTICAL FOUNDATION

1. Executive Analytical Statement

Q2 2026 shifted analytical emphasis toward the software, development and model supply chains, while the identity, orchestration and observability risks identified in Q1 2026 remained active. The quarter's most consequential events did not turn on what a model generated; they turned on what AI-connected systems and the pipelines that build them were trusted to do — and on how quickly a disclosed weakness could be turned into a working attack.

Principal judgement. Material AI security risk increasingly depends on delegated enterprise authority, supply-chain trust and execution control rather than on model output alone. **Confidence: Moderate.**

- **Basis:** the architecture of agentic and AI-assisted systems; a verified Q2 software-supply-chain campaign chain; research demonstrating rapid weaponisation of disclosed vulnerabilities; the first reported export-control action against a deployed model; and recurrent control deficiencies in provenance, credential hygiene, authorisation and observability.
- **Alternative explanation:** some identified risks may be ordinary software, identity or third-party failures relabelled as AI security.
- **Why the judgement still holds:** AI integration changes the scale, speed, iteration and cross-system reach of those established failure modes.
- **What would change the assessment:** strong evidence that production agents and AI build pipelines are consistently isolated, independently authorised, provenance-verified and fully reconstructable across representative deployments.

This Technical & Compliance Report provides the forensic evidence base for the strategic judgement in the companion **Executive Brief**. The two documents share one dataset, one evidence cut-off and one financial treatment. The throughline across both quarters: **whatever the ecosystem trusts by default becomes the delivery mechanism.**

2. Methodology — Evidence, Classification and Analytical Discipline

ODA3 Institute holds no proprietary incident telemetry, client data or internal datasets. All findings derive from public disclosures, government and vendor advisories, court filings, regulatory filings and named threat-intelligence research.

2.1 Two-Dimension Evidence Model

Each record is assigned **both** a record class (what kind of event it is) and a source tier (how strong the evidence is), plus an analytical confidence. Controlled research can be strong evidence of a *capability* while proving nothing about a *production incident*; separating the two dimensions prevents research from being mistaken for a weak incident report.

Record classes

Code	Record class	Definition
INC	Realised incident	Confirmed adverse security, privacy, safety, financial or operational impact
ATT	Attempted incident	Malicious activity without confirmed material impact
EXP	Exposure event	Data, credentials or capability became accessible without confirmed exploitation
VUL	Vulnerability	A technical weakness, with or without observed exploitation
RES	Research observation	Controlled evaluation, experiment or laboratory finding
REG	Regulatory development	Government or supervisory action affecting security, availability or compliance
SIG	Threat signal	Relevant activity that does not satisfy the incident threshold

Source strength tiers

Tier	Meaning
T1	Primary confirmed — affected organisation, regulator, government directive, court record, authoritative advisory or verifiable technical artefact
T2	Independently corroborated — two or more credible, materially independent sources
T3	Credibly reported — one credible source with detail but incomplete corroboration
T4	Unverified / anecdotal — retained only for monitoring

Confidence: High · Moderate · Low · Indeterminate (analyst judgement, stated separately from tier).

2.2 Analytical Separation

To keep the evidence base auditable and legally defensible, each proposition is labelled by type: **fact** (authoritative/attributable source); **source allegation** (named source, not independently established); **attribution assessment** (judgement about actor/origin/method); **ODA3 inference** (reasoned conclusion from established facts); **ODA3 estimate** (transparent quantitative derivation with range and confidence); **research observation** (controlled finding); **scenario** (forward-looking); **unknown** (not established).

2.3 Naming and Attribution Standard

Organisations, products and open-source projects are named only where the event has been publicly disclosed by the affected party, confirmed by a government body, or assigned a public identifier (for example, a CVE). **Naming an entity reflects the public record of an event and does not constitute an assertion of legal negligence, fault, or liability beyond what is established by the cited evidence.** For open-source projects (for example, LiteLLM and TanStack), the analysis addresses supply-chain mechanics and specific compromised versions; it does not imply that core maintainers acted maliciously. Threat-actor attributions (for example, TeamPCP / UNC6780, and any Lapsus$ association) are reported as attribution assessments made by named third parties and may evolve.

2.4 Financial Estimation Method

Where individual record costs are estimated, ODA3 Institute separates disclosed realised loss, modelled incremental response and recovery cost, and scenario exposure, and applies:

> **Estimated record cost = disclosed non-overlapping realised loss**
> + modelled incremental response and recovery cost
> + probability-weighted regulatory or litigation exposure (capped at the 25th percentile of a named analogue set)

Every estimate carries a low–high range, a confidence assessment, currency, valuation date, sector/jurisdiction assumptions, explicit overlap controls, and the input categories used. Prevented attacks, research demonstrations, vulnerabilities and regulatory developments are never added to realised losses. **No single aggregate Q2 loss and no single incident count are published**, because the record population spans different classes and the disclosed values do not support defensible aggregation.

2.5 Benchmark Qualification

Prevalence and cost benchmarks cited below (for example, breach-cost averages, third-party-share trends, agent-involvement rates, dependency-prevalence percentages) derive from vendor surveys and commercial telemetry with differing populations and methodologies. They are directional context, not general prevalence facts, and are attributed to their source.

3. What Changed Since Q1 2026

Q1 framed the AI control plane — identity, permissions, orchestration, validation and observability inside deployed systems — as the battlefield. Q2 evidence shifted emphasis **up the stack** and **forward in time**; the Q1 risks did not disappear.

Dimension	Q1 2026	Q2 2026	Technical significance
Emphasis of observed attack surface	Deployed AI control plane (identity, permissions, orchestration)	Software build & distribution pipeline (source, dependencies, CI/CD)	The observed objective shifted from operating a deployed system to compromising the pipeline that delivers it; Q1 runtime risks remain active.
Primary technique	OAuth token exploitation; credential theft	Pipeline & distribution-channel poisoning (OpenID Connect token abuse; silent auto-update)	Attackers moved from *theft of stored credentials* toward abuse of federated identity and automated credential-minting pathways.
Key failure mode	Unbounded runtime authority	Inherited, unbounded trust in build provenance	Default trust expanded from runtime permissions to build-time provenance.
Headline loss type	Mass record / personally identifiable information (PII) exposure	Credential, source-code, intellectual-property (IP) and capability events	The loss shape changed; a single aggregate is no longer defensible.
Capability frontier	AI-accelerated reconnaissance	Reported autonomous rapid exploit synthesis; first reported export-control action against a deployed model	The "expert-weeks" assumption for exploit development is reported to be compressing toward hours.
Carried-forward gap	Model Context Protocol (MCP) / agent authentication baseline	Still unmet — folded into CTRL-06 and GAP-06	The Q1 agentic-authentication gap remains open.
Throughline	Trust inheritance is the recurring failure: whatever the ecosystem trusts by default becomes the delivery mechanism.		

4. Notably Absent — Boundaries Explicitly Documented

A disciplined assessment must state what the evidence did **not** establish. These absences constrain the report's conclusions; they do not establish that the corresponding conditions could not occur.

Incident scope and scale

- The evidence did not establish a defensible total number of realised AI security incidents for the quarter, a defensible aggregate exposed-record total, a defensible aggregate Q2 financial-loss figure, or a reliable population-wide AI incident rate.
- No confirmed mass-PII government breach at the scale of Q1's multi-agency intrusion (~195 million records) recurred in Q2 public reporting.

Attribution and autonomy

- The evidence did not establish that a generally available model independently selected a target, obtained access and completed an end-to-end material cyberattack without human direction or enabling infrastructure.
- The evidence did not establish AI systems as generally functioning as independent threat actors, nor that every reported model-assisted exploit was independently discovered by a model.
- The reported advanced-model exploit-development capability (RES-26-Q2-01) is a controlled-research finding; no confirmed in-the-wild criminal use was observed within the period, and no authoritative source establishes that this research triggered the export-control action (REG-26-Q2-01).
- No confirmed exfiltration of the Fable 5 / Mythos 5 model weights occurred; the June action was a government access directive, not a model theft.
- The export-control trigger is contested: Anthropic characterises it as a narrow, non-universal jailbreak available from other public models; the government has not published the order.

Control effectiveness

- The evidence did not establish that conventional security controls are universally ineffective against AI-enabled threats, that human incident response is obsolete, or that model-level safeguards provide reliable authorisation for consequential actions.
- For the GitHub / Nx breach (INC-26-Q2-01), customer repositories, enterprise accounts and user data were reported **not** affected; exposure was confined to GitHub's internal estate.
- No confirmed autonomous-agent runaway in a production environment was observed.

Standards and regulation

- The evidence did not establish that any single current framework provides a complete implementation specification for agent identity, delegated authority, runtime enforcement and action-level forensics.
- No European Union (EU) AI Act high-risk fines were recorded; the high-risk obligations were the subject of a proposed deferral during the quarter (not yet enacted).
- The evidence did not support combining vulnerabilities, attempted attacks, research observations and regulatory developments into a single incident total.

PART II — VERIFIED RECORD PORTFOLIO

5. Record Register

Records are assessed by class and are **not** summed into a single "incident" count. I-x aliases are retained for cross-reference with the Executive Brief and the detection/traceability sections.

Record ID	Alias	Record	Class	Tier	Confidence
INC-26-Q2-01	I-1	GitHub internal-repository breach via poisoned Nx Console extension	INC	T1	High
INC-26-Q2-02	I-2	Mercor data breach via LiteLLM supply-chain compromise (late-Q1 origin; Q2 fallout)	INC	T1 fact / T2 scope	High (fact) / Moderate (scope)
INC-26-Q2-03	I-3	TanStack npm CI/CD compromise (CVE-2026-45321)	INC (also: VUL)	T1	High
INC-26-Q2-04	I-4	Self-propagating npm worm ("Shai-Hulud" / "Mini Shai-Hulud") — campaign code reportedly published	INC (also: SIG)	T2	Moderate

Record ID	Alias	Record	Class	Tier	Confidence
RES-26-Q2-01	(was I-5a)	Advanced-model rapid exploit-development capability research	RES	T2 / T3 (specific figures)	Moderate (capability)
REG-26-Q2-01	(was I-5b)	U.S. export-control directive suspending Fable 5 / Mythos 5	REG	T1	High
REG-26-Q2-02	(new)	Partial restoration of Mythos 5 access for approved U.S. organisations	REG	T1	High
REG-26-Q2-03	(new)	Full withdrawal of Fable 5 / Mythos 5 export controls	REG	T1	High
EXP-26-Q2-01	I-6	Mass exposure of unauthenticated AI infrastructure	EXP	T3	Moderate

6. Realised Incidents (INC)

INC-26-Q2-01 (I-1) — GitHub Internal Repository Breach via Poisoned Nx Console Extension

Class: INC · **Tier:** T1 · **Confidence:** High.

Confirmed (fact). A poisoned build of the Nx Console VS Code extension (nrwl.angular-console v18.95.0) was published to the Visual Studio Marketplace and Open VSX on 18 May 2026 and distributed silently via editor auto-update [T1]. Multiple sources including GitHub's CISO confirmed the compromise of a GitHub employee device and exfiltration of approximately 3,800 internal repositories by the actor tracked as TeamPCP / UNC6780 [T1]. GitHub issued a security advisory (GHSA-c9j4-9m59-847w); CVE-2026-48027 was assigned and added to the CISA KEV catalog (with CVE-2026-45321 for TanStack) on 27 May 2026, with a 10 June 2026 federal remediation deadline under Binding Operational Directive 22-01 [T1]. The malicious build was live on the Marketplace for roughly 18 minutes (≈12:30–12:48 UTC) and approximately 36 minutes on OpenVSX (≈12:33–13:09 UTC) [T1].

Mechanics (fact / attribution). The malicious version, root-caused to the TanStack compromise (INC-26-Q2-03), stole a legitimate Nx contributor's GitHub CLI OAuth token and was published without required manual approval [T2]. On workspace activation it ran a hidden task — disguised as a routine MCP setup step — that fetched an obfuscated payload from a dangling commit on the official nrwl/nx repository, harvested developer and cloud credentials (reported to include npm, GitHub, AWS and 1Password stores, and AI coding-assistant configurations), and established macOS persistence via a LaunchAgent [T2].

Reported (scope, to confirm). Marketplace download counters understate reach because auto-update is not counted; Nx telemetry indicated on the order of ~6,000 activations of the malicious build [T2]. Downstream, tooling from OpenAI, Grafana Labs and Mistral has been named in reporting as potentially affected via harvested secrets [T2]; treat CI/CD credential stores as potentially compromised.

Control failure. No marketplace code-signing or hold period before auto-update push; silent auto-update enabled by default on privileged developer machines; long-lived secrets and AI-assistant configs readable on disk; single-actor extension release without approval.

Indicative technique alignment (MITRE ATT&CK; analyst-assessed): supply-chain compromise (T1195.002); valid-account/credential access; persistence. Mappings are indicative and reflect reported behaviour.

Financial treatment. ODA3 modelled estimate US$20M–US$50M **[Low–Medium]** — inputs: internal incident response; mass credential rotation across npm/GitHub/AWS/1Password; internal source-exposure risk; brand. Excludes customer-data loss (none reported) and ecosystem-wide rotation across other affected installs. Not a disclosed figure.

Significance. The breach required no exploit and no memory-safety vulnerability; it weaponised trust in an official marketplace, and an ~18-minute window sufficed to reach a major code-hosting platform. ODA3 assessment: exposure-time is no longer a reliable primary risk metric for this class of attack.

INC-26-Q2-02 (I-2) — Mercor Data Breach via LiteLLM Supply-Chain Compromise

Class: INC · **Tier:** T1 (fact) / T2 (scope) · **Confidence:** High (fact) / Moderate (scope). **Coverage note:** the originating compromise occurred in **late March 2026 (Q1)**; it is retained here as a carried record because material developments (litigation, continuity fallout) fell within Q2.

Confirmed (fact). Mercor — a data-labelling supplier to multiple frontier labs (reported customers include OpenAI, Anthropic, Meta, Google) — confirmed on 31 March 2026 that it was "one of thousands of companies" affected by a compromise of the open-source LiteLLM AI gateway [T1]. Malicious LiteLLM versions 1.82.7 and 1.82.8 were published to the Python Package Index (PyPI) on 24 March 2026 and were live for roughly 40 minutes [T1]. Attribution to TeamPCP is widely reported [T2]; the Lapsus$ extortion group separately claimed possession of Mercor data (attribution assessment / source allegation) [T2]. A putative class action was filed in the U.S. District Court, Northern District of California (Ananthula v. Mercor.io, No. 3:26-cv-03362) [T1].

Reported (scope, to confirm / conflicting). Threat-actor claims and secondary outlets cite roughly 4 TB exfiltrated and 40,000+ contractors' data exposed (candidate records, interview video/biometrics, source code, API keys); however, the formal class-action pleading specifies approximately 211 GB of candidate records — the discrepancy may reflect source code, biometric video and API-key material not itemised in the legal filing (ODA3 inference — the available evidence does not fully establish the explanation). Meta reportedly paused work with Mercor. LiteLLM is reported at ~97M monthly downloads and present in an estimated ~36% of cloud environments (vendor/research figures; treat as directional). The precise entry vector varies by source and full scope rests on secondary research and pleadings.

Control failure. A trusted open-source AI dependency was consumed without provenance gating or a Software Bill of Materials (SBOM) for AI; unpinned dependencies propagated the tainted release into build and runtime environments.

Financial treatment. ODA3 modelled estimate US$25M–US$70M **[Low]** — inputs: notification; incident response; contract loss; IP-exposure remediation. Systemic exposure across the LiteLLM dependent base is **not reliably quantifiable**.

Significance. A single AI-gateway dependency became a shared point of failure across the frontier-lab training supply chain. TeamPCP's reported intent to work with extortion groups invites comparison to the 2023 MOVEit campaign (~100M individuals), but that comparison is a scenario, not a realised outcome.

INC-26-Q2-03 (I-3) — TanStack npm CI/CD Compromise ("Mini Shai-Hulud" wave)

Class: INC (also: VUL) · **Tier:** T1 · **Confidence:** High.

Confirmed (fact). The TanStack npm compromise was assigned CVE-2026-45321 (reported CVSS 9.6) and added to the CISA KEV catalog on 27 May 2026 [T1]. Reporting (Wiz, OX Security, Phoenix Security) describes a chain through TanStack's GitHub Actions pipeline — a pull_request_target workflow misconfiguration, Actions cache poisoning, and runtime extraction of an OpenID Connect (OIDC) token — used to obtain fresh npm publishing credentials that carried valid Sigstore / Supply-chain Levels for Software Artifacts (SLSA) Build Level 3 provenance [T2]. The malicious package @tanstack/zod-adapter@1.166.15 seeded the contributor-token theft behind INC-26-Q2-01 [T2].

Reported (scope, to confirm). Dozens of malicious package versions were published within minutes, all carrying valid cryptographic provenance; downstream exposure was reported across dependents.

Control failure. A valid provenance attestation does not certify benign intent when the signing pipeline itself is compromised; no minimum-age hold before downstream consumption.

Indicative technique alignment: supply-chain compromise — software dependencies (T1195.001); OIDC token extraction and credential-minting abuse (no exact ATT&CK sub-technique; see §12). Indicative.

Financial treatment. ODA3 modelled estimate US$15M–US$40M **[Low]** — inputs: mass credential rotation and downstream incident response across dependents.

Significance. Signed-provenance checks (SLSA/Sigstore), a cornerstone of recent supply-chain defence, are bypassed when the pipeline minting the signature is compromised.

INC-26-Q2-04 (I-4) — Self-Propagating npm Worm ("Shai-Hulud" / "Mini Shai-Hulud") — Campaign Code Reportedly Published

Class: INC (also: SIG) · **Tier:** T2 · **Confidence:** Moderate.

Reported (corroborated). A self-propagating npm worm using install-time lifecycle scripts (binding.gyp / node-gyp hooks) and editor tasks.json auto-run as detonation primitives was documented by multiple research labs as part of the same campaign, and its framework was reportedly published publicly, lowering the barrier for copycats. Propagation counts (e.g., 1,000+ repositories within 24 hours; later multi-package bursts) are research estimates.

Control failure. Package lifecycle scripts execute automatically at install with no sandbox or review; editor task files run on folder-open. Both are "features by design" that depend on a trust assumption attackers violate.

Indicative technique alignment: supply-chain compromise (T1195.002); user-execution / lifecycle-script execution; worm propagation. Indicative.

Financial treatment. Not separately quantified; folds into ecosystem remediation. Public availability of reusable campaign code may lower the barrier for to lower-skilled actors, raising future frequency.

Significance. A nation-state-style supply-chain technique reportedly moved to commodity crimeware and was then reportedly made publicly available, compressing the proliferation timeline.

7. Research Observation (RES)

RES-26-Q2-01 — Advanced-Model Rapid Exploit-Development Capability

Class: RES · **Tier:** T2 (capability direction) / T3 (specific figures) · **Confidence:** Moderate. **Not an incident.**

Research observation. Frontier red-team research reported that an advanced ("Mythos-class") model could turn disclosed, patched vulnerabilities into working exploits in hours rather than the "expert-weeks" the patch-cadence model assumes — reportedly producing multiple working exploits from recent browser and operating-system patches, at low per-exploit cost, including for issues rated low-likelihood by vendors. Public models with safeguards removed reportedly produced working exploits at lower success. Related reporting notes advanced models identifying large numbers of vulnerabilities.

ODA3 assessment. This is credible evidence of a *capability trajectory* ("N-day → N-hour"), not evidence of a production incident or of in-the-wild criminal use. The specific counts and per-exploit costs are reported figures that ODA3 has not independently reproduced; they are presented as reported and tiered accordingly. This capability is offensive-enabling regardless of operator and is a defensive concern for slow-to-patch, internet-exposed and fixed-cadence assets (industrial control systems, medical devices, Internet-of-Things estates).

Required control outcome. Patch prioritisation and compensating controls should be calibrated to a credible scenario in which selected high-exposure disclosed vulnerabilities may be weaponised within hours (see CTRL-04, GAP-05). No causal link between this research and REG-26-Q2-01 should be inferred.

8. Regulatory Developments (REG)

REG-26-Q2-01 — U.S. Export-Control Directive Suspending Fable 5 / Mythos 5

Class: REG · **Tier:** T1 · **Confidence:** High. **Not a cyber incident.**

Confirmed (fact). On 12 June 2026 the U.S. Commerce Department's Bureau of Industry and Security issued an export-control directive (Lutnick letter to Anthropic) [T1] requiring an export license for any foreign national — including foreign nationals inside the United States and Anthropic's own non-citizen employees — to access Claude Fable 5 and Mythos 5, citing national-security authorities under the Export Control Reform Act / Export Administration Regulations. Unable to screen users by nationality in real time, Anthropic disabled both models for all customers worldwide [T1]; less-capable models (e.g., Claude Opus 4.8) were unaffected. Fable 5 had been publicly available for three days (released 9 June). The order has not been made public.

Contested (attribution / rationale). The government cited a reported method of "jailbreaking" Fable 5, reportedly flagged by Amazon. Anthropic disputes the severity, characterising it as a narrow, non-universal jailbreak available from other deployed models (it named OpenAI's GPT-5.5), and says it received no detailed technical evidence. Analysts (e.g., CSIS) note the worldwide "is informed" control basis had reportedly not been used before, which supports describing this as **reported to be among the first / most aggressive** export-control actions against a commercially deployed model — a qualified, not absolute, "first."

Control implication. The June 2026 sequence demonstrates that access to a frontier model can be suspended, conditionally restored and fully reinstated through rapidly changing government and provider decisions. The restrictions remained in effect for approximately 18 days, with partial Mythos access restored before full withdrawal; broader service restoration began after the reporting-period cut-off. Material AI-enabled processes require documented fallback and continuity (see CTRL-05, GAP-06).

REG-26-Q2-02 — Partial Restoration of Mythos 5 Access

Class: REG · **Tier:** T1 · **Confidence:** High. **Update to REG-26-Q2-01.**

Confirmed (fact). On 26 June 2026 the Commerce Secretary issued a letter (reported by Axios) permitting a license-free path for approved U.S. organisations (entities named in an annex, and their and Anthropic's foreign-national employees) to access **Mythos 5**; restrictions on **Fable 5** were not changed, and controls remain for all organisations not explicitly approved. The letter reserves the right to adjust scope.

ODA3 assessment. Model-access restrictions are granular, conditional and jurisdiction-specific rather than binary; continuity plans should not assume a simple available/unavailable state (see CTRL-05).

REG-26-Q2-03 — Full Withdrawal of Fable 5 / Mythos 5 Export Controls

Class: REG · **Tier:** T1 · **Confidence:** High. **Update to REG-26-Q2-01 and REG-26-Q2-02.**

Confirmed (fact). On 30 June 2026 — the evidence cut-off date — the U.S. Commerce Department withdrew the export controls applying to both Claude Fable 5 and Mythos 5 [T1]. Commerce Secretary Lutnick stated the controls in the 12 June letter were "withdrawn" following the Bureau of Industry and Security's evaluation that appropriate safeguards were in place. A licence is no longer required for the models under the withdrawn controls. Anthropic confirmed it would begin restoring access from 1 July and stated it had agreed to proactively detect and address security risks, work with the government on standards for upcoming models, and inform the government of malicious activity. Fable 5 will be restored with additional cybersecurity safeguards.

ODA3 assessment. The full June sequence — imposition (12 June), partial conditional restoration (26 June), full withdrawal with commitments (30 June) — demonstrates that access to frontier-model capability can be suspended, conditionally restored and reinstated through rapidly changing government and provider decisions within a single quarter. The episode demonstrates **regulatory and provider-access volatility** rather than a continuing prohibition. Business-critical deployments should treat legal authority, provider policy and jurisdiction-specific availability as continuity dependencies that may change with limited notice. The control outcome (CTRL-05) remains necessary: even though controls were withdrawn for these models, the precedent and mechanism remain available.

9. Exposure Finding (EXP)

EXP-26-Q2-01 (I-6) — Mass Exposure of Unauthenticated AI Infrastructure

Class: EXP · **Tier:** T3 · **Confidence:** Moderate. **Not a realised incident.**

Reported (corroborated). A security-research scan (reported via The Hacker News) of roughly one million internet-exposed AI services found agent-management platforms (n8n, Flowise) and Model Context Protocol (MCP) servers reachable without authentication; some exposed chatbot business logic, credential lists, and file-write or code-interpretation tools. This extends a Q1 finding of hundreds of unauthenticated MCP servers to a larger base.

Control failure. Absent access-management controls in AI tooling: access to a bot integrated with third-party systems often means access to everything it touches; weak sandboxing enlarges blast radius; infrastructure frequently sits outside any demilitarised zone (DMZ).

Financial treatment. Not quantified; standing latent exposure rather than realised loss.

Significance. The AI-tooling layer is shipping faster than its security review, reproducing an "exposed admin panel" failure mode at AI scale (see CTRL-06, GAP-06).

PART III — ANALYSIS & RESPONSE

10. Financial Treatment — Bounded Estimation

Q2's disclosed losses are smaller than Q1's; the systemic exposure is larger and is deliberately left unbounded. **No single aggregate Q2 loss is published.**

Metric	Value	Confidence	Basis
Modelled incident-response and recovery costs, named supply-chain incidents (INC-01 to INC-03)	US$60M – US$160M	Low	Sum of per-record ODA3 estimates in §6 (inputs listed there)
Systemic exposure (dependency base + patch-gap collapse)	Not reliably quantifiable	—	No countable denominator in public data
Average supply-chain breach cost (benchmark)	US$4.91M / incident	—	IBM 2025 (survey)
Average resolution time (benchmark)	267 days	—	IBM 2025 (survey)

Why we do not publish a single Q2 aggregate. Q1 2026's aggregate was anchored to a countable population of exposed records. Q2's dominant exposures — a dependency reported in ~36% of cloud environments, an extension distributed by auto-update, and a reported exploit-development capability that applies to every unpatched system — have no countable denominator. A precise total would imply confidence the evidence does not support. The US$60M–US$160M figure is a **bounded sum of directly estimable modelled losses for three named incidents**, not a quarter total; it excludes the systemic exposure above. Per-record input categories are listed in §6; full numerical workpapers (organisation counts, response hours/rates, rotation scope, notification population, litigation probability, overlap exclusions, analogue set) are maintained in the internal financial ledger and should be attached before any figure is relied upon externally. Speculative ceiling scenarios are **not** published.

11. Supply-Chain & Capability Trust — Architecture and Standards Crosswalk

Definition — software & model supply-chain trust boundary. The assurance layer responsible for the integrity and provenance of every dependency, build artifact, distribution channel, developer credential and model an organisation consumes — before it is trusted to execute or to act. It should enforce provenance verification, channel integrity, least-privilege secret handling, and continuity of model access independent of any single provider.

Layer	Function	Q2 2026 failure	Coverage gap	Records
1. Source & dependency integrity	Verify packages/dependencies are unmodified and from trusted maintainers	LiteLLM & TanStack package poisoning; orphan-commit injection into nrwl/nx	AI-dependency provenance guidance fragmented	INC-01/02/03
2. Distribution & update channel	Review marketplace/registry updates before they reach clients	Poisoned extension auto-updated silently; live ~18 min	No marketplace signing / hold-period requirement	INC-01/04

Layer	Function	Q2 2026 failure	Coverage gap	Records
3. Developer credential & secret hygiene	Keep secrets off developer disk; scope and rotate	Vault, 1Password, AWS, CI/CD and AI-assistant configs harvested	Secret-handling guidance for AI dev tools incomplete	INC-01/03
4. Build / CI-CD pipeline trust	Bind signatures to benign builds	OIDC token theft; valid SLSA L3 provenance on malicious artifacts	Provenance ≠ intent; compromised-pipeline control incomplete	INC-03
5. Model provenance & access continuity	Maintain verifiable, resilient access to model capability	Export-control directive disabled deployed models for all customers; controls subsequently withdrawn with provider commitments	Model-failover / continuity guidance incomplete	REG-01/02/03

Standards crosswalk (partial / principle-level coverage exists; implementation detail is fragmented). Source & dependency: NIST SSDF (PS.1–PS.3), SP 800-161 (C-SCRM), SLSA, ISO/IEC 42001 §8.1. Distribution: SP 800-161; SBOM/NTIA minimum elements (software only). Credential hygiene: SP 800-207 (Zero Trust); ISO/IEC 42001 §6.1.2; ISO/IEC 27001 A.5.17. Pipeline: SSDF (PO.5); SLSA Build L3 (insufficient under pipeline compromise). Model continuity: NIST AI RMF (MG-2, MG-4); ISO/IEC 42001 §8.1/§9.1 (no continuity clause for provider revocation).

12. Adversary Technique Mapping (indicative) and Detection Data Sources

Technique identifiers reference MITRE ATT&CK (and, for AI-specific behaviour, MITRE ATLAS). **Mappings are indicative, analyst-assessed, and provided only where the behaviour is reported and observable**; they are not a validated per-artefact attribution, and no non-existent identifiers are used.

Stage	Indicative technique	Reported Q2 behaviour	Records
Initial access	Supply-chain: software dependencies (T1195.001)	Compromise of trusted npm/PyPI dependency via legitimate registry	INC-02/03/04
Initial access	Supply-chain: software supply chain (T1195.002)	Compromise of editor extension distributed via marketplace auto-update	INC-01
Execution	User execution (T1204) / scripting (T1059)	Lifecycle-script & editor-task auto-run; orphan-commit payload on startup	INC-01/04
Credential access	Credentials from stores (T1555) / unsecured credentials (T1552)	Harvest of cloud, CI/CD, secret-manager and AI-assistant secrets	INC-01/02/03
Credential access	Steal or abuse federated identity tokens (no exact ATT&CK sub-technique assigned)	Runtime OIDC-token extraction used to obtain npm publishing credentials; combines token theft with abuse of automated credential-minting authority	INC-03
Context-dependent	Valid accounts (T1078)	Reuse of harvested credentials across code-hosting, cloud or registry services	INC-01/02
Exposure condition	(Not mapped to an adversary technique)	Internet-reachable AI infrastructure without authentication	EXP-01

Stage	Indicative technique	Reported Q2 behaviour	Records
Capability (research)	ATLAS (no specific current identifier maps precisely; indicative only)	Reported rapid exploit synthesis from disclosed patches	RES-01

Detection data sources by tactic. Initial access → endpoint detection and response (EDR) process telemetry + extension/package-update logs (child processes after an update; consumption of a version published minutes earlier). Execution → EDR process-tree (lifecycle scripts/task runners launching shells at install or folder-open). Credential access → file-access auditing + secret-manager logs (editor/gateway reading credential stores it never previously touched). Credential-minting abuse → CI/CD + cloud OIDC issuance logs (tokens to unexpected audiences). Valid-account reuse → cloud audit + registry/VCS logs (valid creds from a new location; clone-scale repo access). External exposure → external attack-surface management (unauthenticated n8n/Flowise/MCP endpoints).

13. Regulatory Exposure

Two structural shifts: a proposed deferral of the EU AI Act high-risk deadline, and the first reported export-control action against a deployed model.

13.1 Legal-Status Vocabulary

Status	Meaning
Enacted	Legally operative instrument
Applicable	Obligation currently applies
Political agreement	Negotiated, not yet enacted
Proposed amendment	Subject to legislative completion and Official Journal publication
Reported target date	Not yet legally operative

No date is described as legally binding unless the operative instrument is verified as enacted and effective at the assessment date.

13.2 EU AI Act — Digital Omnibus (as reported)

A "Digital Omnibus" simplification was the subject of a reported provisional agreement (7 May 2026) and European Parliament endorsement (16 June 2026). As **reported target dates** subject to formal adoption and Official Journal publication:

Obligation	Original date	Reported revised date	Legal status
Prohibited practices (Art. 5)	2 Feb 2025	Unchanged	Applicable
General-purpose AI (GPAI) obligations (Ch. V)	2 Aug 2025	Unchanged	Applicable
Transparency / synthetic-content incl. watermarking (Art. 50)	2 Aug 2026	Reported 2 Dec 2026	Proposed amendment / reported target

Obligation	Original date	Reported revised date	Legal status
New Art. 5 prohibition — non-consensual intimate imagery (NCII) / child sexual abuse material (CSAM)	— (new)	Reported 2 Dec 2026	Proposed amendment
High-risk Annex III	2 Aug 2026	Reported 2 Dec 2027	Proposed amendment (deferral)
High-risk Annex I (embedded products)	2 Aug 2027	Reported 2 Aug 2028	Proposed amendment (deferral)

Compliance implication. Any deferral is relief on timeline, not on architecture. AI inventory and classification gate every downstream obligation. **Action:** until the amending instrument is enacted and published in the Official Journal, the original 2 August 2026 date remains legally applicable — organisations should plan for both, and prioritise AI inventory as the foundational step that enables classification and conformity assessment.

13.3 Record-to-Regulation Mapping

INC-01: NIST C-SCRM; CISA KEV / BOD 22-01 (federal remediation — applicable). INC-02: General Data Protection Regulation (GDPR) Art. 28/32; processor obligations; active U.S. litigation. INC-03: NIST SSDF; EU Cyber Resilience Act (phasing 2026–2027). REG-01/02/03: U.S. Export Administration Regulations (export-control authority). EXP-01: GDPR Art. 32; EU Network and Information Security Directive 2 (NIS2).

14. Standards Gaps

Implementation gaps, not total absence. For each, existing frameworks provide partial or principle-level coverage, but implementation-specific guidance for this failure mode is fragmented or incomplete.

GAP-01 — Editor / IDE extension & marketplace security. *Covered:* general supply-chain and asset-management principles (SP 800-161; ISO/IEC 27001 A.8.19). *Fragmented:* mandatory code signing, publication-to-availability hold periods, behavioural review before auto-update, and CVE-feed representation of hijacked extensions/leaked publisher tokens. *Gap type:* technical + normative. **Evidence:** INC-01.

GAP-02 — AI coding-assistant credential protection. *Covered:* Zero-Trust and secret-management principles (SP 800-207). *Fragmented:* how AI dev tools store, scope and protect configuration files and secrets, now a named harvesting target. *Gap type:* technical. **Evidence:** INC-01, INC-03.

GAP-03 — AI dependency provenance & SBOM-for-AI. *Covered:* SLSA / SBOM for software. *Fragmented:* provenance for AI gateways, model artifacts and training-data lineage; vetting criteria for AI dependencies. *Gap type:* technical + adoption. **Evidence:** INC-02, INC-03.

GAP-04 — Compromised-pipeline / provenance-intent control. *Covered:* SLSA Build L3 (verifies origin). *Fragmented:* controls that fail closed when the signing pipeline is compromised; runtime OIDC-token protection. *Gap type:* technical. **Evidence:** INC-03.

GAP-05 — Patch-gap / "N-hour" response. *Covered:* patch-management frameworks (SP 800-40). *Fragmented:* out-of-band response expectations and compensating-control baselines calibrated to sub-day weaponisation. *Gap type:* normative + technical. **Evidence:** RES-01.

GAP-06 — Model access continuity & provider failover. *Covered:* business-continuity standards (ISO 22301). *Fragmented:* documented fallback models, tested switch-over, and continuity of AI capability independent of any single provider or jurisdiction; agent/MCP authentication baseline (carried from Q1). *Gap type:* normative + technical. **Evidence:** REG-01/02/03, EXP-01.

15. Recommended Control Outcomes

These are ODA3 **recommended implementation outcomes**, not a normative standard. They are intended for mapping to the authoritative GAISSF control register before use as a conformance reference; the GAISSF control ID for each is **to be inserted from the GAISSF normative corpus** and must not be inferred. Applicability depends on system architecture and risk; each carries applicability conditions and residual limitations.

CTRL-01 — Dependency & extension provenance with hold. *Outcome:* third-party packages, dependencies and editor/IDE extensions are provenance-verified and subject to a minimum-age hold before entering a privileged build, CI/CD or developer environment; privileged pipelines pin trusted versions. *Applicability:* environments consuming external packages/extensions. *Residual limitation:* does not stop a compromise that survives the hold window undetected. *GAISSF: [to be mapped]. Reference: NIST SSDF PS.1–PS.3; SP 800-161; SLSA.* (INC-01/02/03/04; GAP-01/03/04)

CTRL-02 — No silent extension auto-update on privileged endpoints. *Outcome:* silent auto-update is disabled on machines with access to source, secrets or production; updates pass review or a hold; an extension/permission inventory is maintained. *Applicability:* developer and privileged workstations. *Residual limitation:* depends on inventory completeness. *GAISSF: [to be mapped]. Reference: SP 800-161; ISO/IEC 27001 A.8.19.* (INC-01/04; GAP-01)

CTRL-03 — Short-lived, task-scoped developer/AI-tool secrets. *Outcome:* secrets accessible to AI developer tools are managed, short-lived and task-scoped, not long-lived on disk; after any developer-tooling compromise, all reachable credentials are treated as compromised and rotated. *Applicability:* any AI-assisted development environment. *Residual limitation:* rotation latency. *GAISSF: [to be mapped]. Reference: SP 800-207; ISO/IEC 42001 §6.1.2.* (INC-01/02/03; GAP-02)

CTRL-04 — Out-of-band patch capability for the N-hour reality. *Outcome:* an out-of-band patch path and documented compensating controls exist for internet-exposed and slow-to-patch assets, calibrated to a credible scenario in which selected high-exposure disclosed vulnerabilities are weaponised within hours; prioritisation weights exposure and weaponisation speed over vendor likelihood ratings. *Applicability:* internet-exposed, ICS/medical/IoT and fixed-cadence estates. *Residual limitation:* cannot control the underlying capability, only the exposure window. *GAISSF: [to be mapped]. Reference: SP 800-40; NIST CSF (Respond).* (RES-01; GAP-05)

CTRL-05 — Model-access continuity & failover. *Outcome:* every business-critical AI integration has a documented fallback model and a tested switch-over path; model-access continuity is a business-continuity item; single-provider dependencies are identified and risk-accepted at a named level. *Applicability:* business-critical AI-enabled processes. *Residual limitation:* fallback may not match performance/safety. *GAISSF: [to be mapped]. Reference: NIST AI RMF MG-2/MG-4; ISO 22301.* (REG-01/02/03; GAP-06)

CTRL-06 — Authenticated agent / MCP surfaces. *Outcome:* MCP servers, agent endpoints and orchestration interfaces enforce mutual authentication, least-privilege per-task grants and immutable invocation logging; no agent platform is exposed to untrusted networks without authentication. *Applicability:* any deployed agent/MCP infrastructure. *Residual limitation:* does not address model-layer manipulation. *GAISSF: [to be mapped]. Reference: OWASP Agentic guidance; ISO/IEC 42001 §9.1.* (EXP-01; GAP-06; carried unmet from Q1)

16. Detection & Threat Hunting

Behavioural hunts, not signatures. Q2's events repeatedly defeated signature- and provenance-based controls; detection that survives the next variant watches what trusted software *does* after it is trusted.

INC-01 — Poisoned editor extension → credential harvest. *Hunt:* correlate editor extension-update events with the following process tree; flag an extension host that, within seconds of an update, spawns an interpreter/shell, reaches a non-allowlisted host, or reads credential stores; hunt unexpected macOS LaunchAgents after an extension event. *Maps to CTRL-02, CTRL-03; addresses GAP-01.*

INC-02 — Malicious AI-gateway dependency. *Hunt:* enumerate where the AI gateway runs; compare versions against the known-bad range; watch for credential reads and egress outside documented model endpoints; check SBOMs/lockfiles for affected versions. *Maps to CTRL-01; addresses GAP-03.*

INC-03 — Pipeline-forged signed packages. *Hunt:* review CI/CD for pull_request_target workflows with publish scope and cross-boundary cache restores; alert on OIDC issuance to unexpected audiences and publish credentials minted outside the release job; do not treat valid provenance as benign intent. *Maps to CTRL-01; addresses GAP-04.*

INC-04 — Self-propagating npm worm. *Hunt:* lifecycle scripts (preinstall/postinstall, node-gyp hooks) performing egress/credential access rather than compilation; editor tasks.json auto-running shells on folder-open; identical payloads across unrelated packages. *Maps to CTRL-01, CTRL-02; addresses GAP-01.*

RES-01 — Patch-gap collapse (posture, not signature). *Hunt for exposure:* maintain a live view of internet-exposed / slow-to-patch assets and the disclosure-to-patch window; treat sub-hour windows as unsafe; do not rely on vendor "exploitation unlikely" ratings for internet-reachable systems. *Maps to CTRL-04; addresses GAP-05.*

EXP-01 — Unauthenticated AI infrastructure. *Hunt:* external attack-surface discovery for agent-management and MCP endpoints answering without authentication; confirm no AI tooling sits outside a DMZ. *Maps to CTRL-06; addresses GAP-06.*

Generic agentic detections (durable, authority-centric). DET-AI-01 tool use outside *task-specific* entitlement; DET-AI-02 untrusted external/public content followed by a consequential action without deterministic approval (indirect prompt injection); DET-AI-03 delegated-credential use exceeding a *task-specific* baseline; DET-AI-04 behavioural-configuration change lacking an approved change record; DET-AI-05 unapproved model/provider substitution; DET-AI-06 denied consequential action followed by an alternate tool path to the same objective; DET-AI-07 session exceeding the approved duration/step envelope; DET-AI-08 cross-agent delegation outside the approved trust graph.

17. Traceability — Control / Gap / Record Matrix

Control \ Record	INC-01	INC-02	INC-03	INC-04	RES-01	REG-01/02/03	EXP-01	Closes gap
CTRL-01 Dependency provenance + hold	✓	✓	✓	✓				GAP-01/03/04
CTRL-02 No silent extension auto-update	✓			✓				GAP-01
CTRL-03 Short-lived dev-tool secrets	✓	✓	✓					GAP-02
CTRL-04 Out-of-band patch / N-hour					✓			GAP-05
CTRL-05 Model-access continuity						✓		GAP-06
CTRL-06 Authenticated agent / MCP							✓	GAP-06
Residual exposure	Low	Med	Low	Med	High	Med	Med	—

Residual exposure reflects risk persisting after mapped controls are implemented. RES-01 stays **High** because the underlying capability shift cannot be controlled at the consumer level. REG-01/02/03 moved from High to **Medium** because the controls were withdrawn — the mechanism and precedent persist but the immediate disruption is resolved.

18. Remediation Roadmap — Sequenced 90-Day Plan

NOW (contain — assume you may already be exposed). (1) Inventory and disable silent auto-update for editor/IDE extensions on privileged endpoints (CTRL-02 / GAP-01). (2) SBOM/lockfile sweep for the known-bad AI-gateway and package versions (LiteLLM 1.82.7/1.82.8; TanStack wave; Nx Console 18.95.0); rotate reachable credentials (CTRL-01/03). (3) External attack-surface scan for unauthenticated n8n/Flowise/MCP endpoints (CTRL-06 / GAP-06).

30–60 DAY (structural). (4) Minimum-age hold + provenance check before any new dependency/extension enters a privileged pipeline (CTRL-01 / GAP-04). (5) Convert long-lived developer/AI-tool secrets to short-lived, task-scoped credentials (CTRL-03 / GAP-02). (6) Audit CI/CD for pull_request_target misuse, cache poisoning and runtime OIDC exposure; fail closed on pipeline compromise (GAP-04).

60–90 DAY (resilience). (7) Re-prioritise patching for internet-exposed/fixed-cadence assets on an hours-not-weeks assumption; stand up an out-of-band path (CTRL-04 / GAP-05). (8) Document fallback models and a tested switch-over path for every business-critical AI integration; name accountable owners for single-provider dependencies (CTRL-05 / GAP-06). (9) Confirm AI inventory and high-risk classification ahead of the reported EU transparency/watermarking target date (2 Dec 2026), treating dates as planning assumptions pending enactment (§13).

None of these requires proprietary tooling; all map to control outcomes in §15 and gaps in §14.

PART IV — GOVERNANCE

19. Source & Claim Traceability

This report is designed for claim-level traceability. Publication-grade traceability depends on completion and independent review of the associated claim ledger and evidence archive. The public source list (Appendix C) provides titles and outlets; the internal claim ledger maps each material proposition to a source with a retrieval date and, where relevant, page/paragraph and an archived copy. Ledger fields: claim ID · record ID · claim text · claim class (fact / allegation / attribution / inference / estimate / scenario) · source ID + URL · retrieval date · tier · confidence · contradictory source · analyst · review status. The ledger, and the financial workpapers referenced in §10, should be attached or made available on request before external reliance.

20. Corrections Policy and Q3 2026 Collection Priorities

Corrections. Material corrections (a wrong fact, a withdrawn source claim, a corrected identifier, a changed classification or legal status, or a materially revised estimate) are recorded in a public corrections entry stating the publication affected, the date, the original and corrected wording, the reason, and the impact on conclusions. Editorial corrections are non-substantive.

Q3 2026 collection priorities (intelligence questions, not predictions). (1) Will the EU AI Act amendments be formally adopted and published, and which obligations and dates become operative? (2) Will public disclosures distinguish model-assisted exploit development from independent model discovery? (3) Will organisations report agent identity, permission and tool-use failures with sufficient detail for cross-incident analysis? (4) Will model/provider restrictions evolve further — will the withdrawn Fable 5 / Mythos 5 controls be reimposed, will conditions tighten, and will organisations adopt tested substitution/continuity despite the current resolution? (5) Will supply-chain controls expand to behavioural configuration (prompts, tool descriptions, agent manifests)? (6) Will the reported "N-hour" exploit-development capability be observed in the wild, or remain research-bound?

21. Appendices

Appendix A — Evidence-Tier Register

Record	Class	Tier	Basis
INC-26-Q2-01	INC	T1	GitHub advisory (GHSA-c9j4-9m59-847w) + CISA KEV + CVE-2026-48027 + multiple independent analyses
INC-26-Q2-02	INC	T1 fact / T2 scope	Mercor confirmation to primary press + filed class action; full scope on secondary research
INC-26-Q2-03	INC (also: VUL)	T1	CVE-2026-45321 + CISA KEV + multiple research firms
INC-26-Q2-04	INC (also: SIG)	T2	Several research labs/outlets; propagation counts are estimates
RES-26-Q2-01	RES	T2 / T3	Research reporting on capability; specific figures reported, not independently reproduced
REG-26-Q2-01	REG	T1	Commerce/BIS action; Anthropic statements; Fortune/CNN/CSIS/Just Security. Controls subsequently withdrawn (REG-26-Q2-03).
REG-26-Q2-02	REG	T1	Commerce letter reported by Axios; Anthropic confirmation. Superseded by full withdrawal (REG-26-Q2-03).

Record	Class	Tier	Basis
REG-26-Q2-03	REG	T1	Commerce withdrawal 30 Jun 2026; Lutnick statement; CNBC/CNN/Washington Post/Al Jazeera; Anthropic confirmation.
EXP-26-Q2-01	EXP	T3	Single research scan relayed through one outlet; not independently reproduced

Appendix B — Glossary & Acronyms

AI (artificial intelligence) · API (Application Programming Interface) · ATT&CK / ATLAS (MITRE technique knowledge bases; ATLAS covers AI-specific techniques) · BCP (business continuity plan) · BIS (U.S. Bureau of Industry and Security) · BOD 22-01 (CISA Binding Operational Directive) · C-SCRM (Cyber Supply Chain Risk Management, NIST SP 800-161) · CI/CD (Continuous Integration and Continuous Delivery) · CISA (U.S. Cybersecurity and Infrastructure Security Agency) · CRA (EU Cyber Resilience Act) · CSAM (child sexual abuse material) · CVE (Common Vulnerabilities and Exposures) · DMZ (demilitarised zone) · EAR (U.S. Export Administration Regulations) · ECRA (Export Control Reform Act) · EDR (endpoint detection and response) · GDPR (General Data Protection Regulation) · GPAI (general-purpose AI) · ICS (industrial control systems) · IoT (Internet of Things) · IP (intellectual property) · KEV (Known Exploited Vulnerabilities catalogue) · MCP (Model Context Protocol) · NCII (non-consensual intimate imagery) · N-day / N-hour (a disclosed-and-patched vulnerability; "N-hour" denotes disclosure-to-working-exploit compression from weeks to hours) · NIS2 (EU Network and Information Security Directive 2) · OIDC (OpenID Connect) · PII (personally identifiable information) · PyPI (Python Package Index) · SBOM (Software Bill of Materials) · SLSA (Supply-chain Levels for Software Artifacts) · SSDF (NIST Secure Software Development Framework) · TeamPCP / UNC6780 (threat-actor designations reported for the Q2 supply-chain campaign) · VCS (version control system).

Appendix C — Sources (public)

Supply-chain campaign (INC-01/03/04): CISA, "Supply Chain Compromises Impact Nx Console and GitHub Repositories," KEV advisory, CVE-2026-48027 & CVE-2026-45321 (cisa.gov, 28 May 2026); GitHub Security Advisory GHSA-c9j4-9m59-847w, "Compromised Nx Console version 18.95.0" (github.com); The Hacker News, "GitHub Internal Repositories Breached via Malicious Nx Console VS Code Extension" (thehackernews.com, May 2026); Infosecurity Magazine (infosecurity-magazine.com, 21 May 2026); Cybersecurity Dive (cybersecuritydive.com, 29 May 2026); Tech Times (techtimes.com, 21 May 2026); Corgea research (corgea.com, 20 May 2026).

Mercor / LiteLLM (INC-02): TechCrunch, "Mercor says it was hit by cyberattack tied to compromise of open source LiteLLM project" (techcrunch.com, 31 Mar 2026); The Register (theregister.com, 2 Apr 2026); The Next Web (thenextweb.com, 4 Apr 2026); Cybernews (cybernews.com, 1 Apr 2026); Hall Attorneys / court filing, Ananthula v. Mercor.io, N.D. Cal. No. 3:26-cv-03362 (21 Apr 2026); Strikegraph analysis (strikegraph.com, Apr 2026).

Export control (RES-01, REG-01/02/03): Fortune, "Anthropic disables Fable and Mythos AI models following U.S. government export ban" (fortune.com, 13 Jun 2026); CNN Business (cnn.com, 21 Jun 2026); CSIS, "The Department of Commerce Restricted Access to Anthropic's Latest Models" (csis.org, Jun 2026); TechPolicy.Press (techpolicy.press, Jun 2026); Just Security (justsecurity.org, Jun 2026); Axios, "Commerce Department greenlights partial return of Anthropic's Mythos" (axios.com, 27 Jun 2026); Forbes (forbes.com, 16 Jun 2026); CNBC, "Anthropic says Trump admin has lifted export controls on Claude Fable 5 and Mythos 5" (cnbc.com, 30 Jun 2026); CNN, "White House lifts export control on Anthropic" (cnn.com, 30 Jun 2026); Washington Post (washingtonpost.com, 30 Jun 2026); Al Jazeera (aljazeera.com, 1 Jul 2026); Reuters coverage of the 26 Jun limited restoration and 30 Jun full withdrawal.

Exposed AI infrastructure (EXP-01): The Hacker News reporting of an Intruder scan of exposed AI services (thehackernews.com, May 2026).

Benchmarks / frameworks (directional): IBM Cost of a Data Breach Report 2025; HiddenLayer 2026 AI Threat Landscape Report; NIST AI RMF 1.0, SSDF, SP 800-161; ISO/IEC 42001:2023; OWASP guidance for agentic applications; SLSA framework documentation.

Retrieval dates, persistent identifiers and archived copies for each source are held in the internal claim ledger (§19).

Companion publication: Q2 2026 AI Security Executive Brief (ODA3-2026-07-EXB-SEC-005).

Companion publication: Q2 2026 AI Security Executive Brief (ODA3-2026-07-EXB-SEC-005).

© 2026 ODA3 Pvt Ltd. All financial figures are conservative ODA3 modelled estimates. This report does not constitute legal, regulatory or professional advice.