CSAI Foundation

Cloud Security Alliance AI Safety Initiative

Attributing AI Attacks: When Cyber Coverage Becomes Conditional

War Exclusions, Attribution Gaps, and the Insurance Reckoning in the AI Era

Unofficial AI-assisted Research

2026-04-10

Key Takeaways

- The landmark Merck/NotPetya litigation and its January 2024 settlement reshaped war exclusion language industrywide, yet the resulting LMA5567A clauses introduce attribution standards that are increasingly tested in an era of AI-assisted attacks.
- Lloyd's mandated state-backed attack exclusions across all standalone cyber policies from March 2023; the threshold trigger — "major detrimental impact" on state functioning — leaves a substantial coverage gray zone for corporate-targeted destructive attacks that fall short of systemic warfare-scale damage.
- The March 2026 Stryker cyberattack exemplifies this coverage gray zone: a destructive wiper operation carried out by Handala — formally attributed by the Justice Department to Iran's Ministry of Intelligence and Security on March 20, 2026 — yet unlikely to satisfy the "major detrimental impact on state functioning" threshold that most war exclusion clauses require to exclude coverage.
- AI-native malware — including PROMPTFLUX (assessed as still in development and testing) and PROMPTSTEAL (deployed operationally by APT28), both identified by Google's Threat Intelligence Group — can dynamically generate obfuscated payloads and impersonate state-actor TTPs, making forensic attribution materially harder and coverage disputes more likely to be protracted.
- Security teams should treat cyber insurance policy language as a technical control, reviewing attribution evidence requirements, negotiating carve-backs for state-adjacent destructive attacks, and maintaining incident documentation that maps to policy threshold language.

Background

War exclusions have been a standard feature of property and casualty insurance since the early twentieth century, but their application to cyber incidents remained largely theoretical until 2017. The NotPetya attack — a destructive wiper deployed by Russian military intelligence that originated in Ukraine and cascaded globally, causing an estimated $10 billion in damages across the pharmaceutical, shipping, and manufacturing sectors — forced the question into courtrooms [1]. Pharmaceutical giant Merck, which

sustained approximately $1.4 billion in losses, filed claims under its all-risk property insurance. Its insurer, Ace American, invoked a traditional "hostile or warlike action" exclusion, arguing that because the attack originated in an active conflict zone, coverage was voided.

A New Jersey Superior Court and its appellate court rejected this defense, finding that policy language designed for conventional armed conflict was insufficient to exclude cyberattacks that spread indiscriminately beyond any theater of war and struck civilian commercial enterprises with no connection to the underlying hostilities [1][3]. Merck and its insurers reached a confidential settlement in January 2024, reportedly resolving approximately $1.4 billion in losses claimed under its policy [2]. The outcome, while favorable to policyholders, sent a clear signal to underwriters: existing war exclusion frameworks were not fit for purpose in cyberspace.

Lloyd's of London responded with Market Bulletin Y5381, issued in August 2022, mandating that all standalone cyber policies incepted or renewed after March 31, 2023 include explicit exclusions for state-backed cyberattacks [4]. The Lloyd's Market Association developed model clauses — most prominently LMA5567A — to give underwriters standardized language. Rather than a blanket exclusion for any nation-state activity, LMA5567A adopts a threshold approach: the exclusion applies when a cyber operation causes, or is reasonably likely to cause, "major detrimental impact" to the functioning of a state's essential services or security capabilities [5]. This threshold approach appears designed to ringfence catastrophic systemic events while preserving coverage for ordinary ransomware incidents, even those carrying state-actor fingerprints.

That calibration made strategic sense in 2022. By 2026, the framework is under significant stress: the attribution standard it requires has become harder to satisfy as AI-assisted attack tooling blurs the forensic indicators on which attribution has traditionally rested, and an expanding category of destructive, state-adjacent incidents falls awkwardly between the clause's exclusions and protections.

Security Analysis

The Attribution Paradox at the Heart of Modern Cyber Policy

Attribution of state-sponsored cyberattacks has always been contested, but the legal and financial stakes attached to attribution determinations have increased significantly since LMA5567A took effect. Under the clause, insureds and insurers are required to assess "objectively reasonable evidence" to determine whether an operation is attributable to a state [6]. In practice, this standard exposes a fundamental asymmetry: governments rarely share the classified intelligence sources that underpin their

attribution judgments, while private insurers and policyholders must interpret public statements, commercially available threat intelligence, and forensic artifacts — evidence that is often incomplete, delayed by months, and diplomatically qualified.

The political dimensions compound the problem. As widely documented in analyses of state attribution practice, governments frequently decline to make formal designations even when intelligence confidence is high, preferring to characterize incidents as "consistent with" a particular actor rather than "conducted by" that state [6]. This creates a zone of strategic or genuine ambiguity — in some cases reflecting deliberate diplomatic caution, in others reflecting authentic analytical uncertainty — that is now a material financial risk variable for any enterprise relying on a standalone cyber policy. Carriers wishing to invoke LMA5567A must establish both that an attack is state-backed and that it meets the severity threshold — requirements that formal government channels may satisfy only on a timeline far longer than claim resolution demands.

The LMA has attempted to address attribution mechanics by distinguishing between LMA5567A (which includes agreed attribution procedures) and LMA5567B (which omits attribution language and requires prior Lloyd's approval) [7][8]. But neither variant resolves the underlying problem: the evidentiary record available to private parties in a cyber insurance dispute is arguably structurally insufficient for the attribution determinations that LMA5567A contemplates — a gap coverage counsel have increasingly identified as a material systemic weakness. Updates published in 2025 refined the clause language without fundamentally altering this dynamic [5][19].

AI as a Systematic Obfuscation Layer

The maturation of AI-assisted attack tooling has introduced a new and qualitatively different dimension to the attribution challenge. Google's Threat Intelligence Group has identified novel malware families — including PROMPTFLUX, assessed as still under active development and testing, and PROMPTSTEAL, deployed operationally by APT28 — that incorporate large language models as active components during execution, dynamically generating obfuscated malicious scripts, adapting payloads to evade detection, and impersonating the tooling signatures associated with other threat actor groups [14]. Traditional forensic attribution relies heavily on consistent behavioral indicators: IP ranges, malware hashes, command-and-control infrastructure patterns, and code style characteristics that have historically been distinctive enough to support group attribution. AI-generated and AI-adapted malware introduces material challenges to traditional attribution models, degrading confidence in behavioral indicator analysis.

Research from Deep Instinct notes that AI tools enable threat actors to mimic the attack patterns, tools, and techniques of entirely different groups — effectively manufacturing false attribution indicators [16]. A sophisticated criminal organization can deploy infrastructure that resembles a state-sponsored

Advanced Persistent Threat, and a state-sponsored group can deliberately adopt criminal ransomware tradecraft to complicate attribution and invoke plausible deniability. Munich Re's 2026 risk reporting characterizes this convergence explicitly, noting that "the dividing line between state-sponsored APTs and state-tolerated groups and criminals is becoming blurred," with groups "operating in scalable, specialised and agile ecosystems" that serve geopolitical ends without constituting formal state operations [12]. Recorded Future has separately examined the technical maturity of AI-native malware within this evolving threat environment, noting that the gap between attacker capability and defender detectability continues to widen [15].

For claims adjusters and coverage counsel, this convergence is acute. LMA5567A's attribution standard asks parties to assess evidence that is frequently ambiguous, contested, politically sensitive, and — now — potentially fabricated by adversarial AI systems specifically designed to impersonate other actors. As agentic AI further automates attack orchestration, the human intent behind any given incident will grow progressively harder to establish with the precision that insurance disputes require. Munich Re has separately warned that agentic AI is poised to affect the frequency of attacks materially, driving higher claims volumes across multiple coverage lines simultaneously [13].

The Stryker Incident: War Exclusions Under Stress

The March 2026 attack on Stryker Corporation illustrates each of these structural tensions in a concrete claims context. On March 11, 2026, Iran-linked hacktivist group Handala executed a destructive wiper operation against Stryker, a US medical device manufacturer holding a $450 million contract with the Department of Defense [9][10][21]. The attackers claimed responsibility for wiping more than 200,000 systems and exfiltrating 50 terabytes of data — figures that have not been independently verified but were widely reported in trade and insurance press, with some independent reporting suggesting a substantially lower number of devices were forensically confirmed affected [9][10]. No ransom demand was issued. On March 20, 2026, the Justice Department formally attributed Handala's operations to Iran's Ministry of Intelligence and Security, describing the group as a state-operated fake hacktivist persona [21]. The formal attribution, while significant, did not resolve the insurance coverage question: even with a government designation on record, the central LMA5567A threshold — whether the attack caused "major detrimental impact" to state functioning — remained unanswered, and a single-company wiper attack, however severe, falls well short of systemic infrastructure disruption.

The insurance implications were instructive even though Stryker had chosen not to purchase cyber insurance [9]. Enterprises with elevated exposure to state-sponsored targeting — including defense contractors, healthcare organizations with federal contracts, and critical infrastructure operators — may face analogous coverage ambiguities under LMA5567A. For such entities, the Stryker scenario maps to the gap the clause leaves unaddressed: an attack targeted enough to be catastrophic for the enterprise

yet not systemic enough to satisfy the exclusion's threshold, and attributed to a state-controlled actor in a manner that formal government channels did confirm — but only nine days after the attack, well outside most claims-initiation timelines.

Insurance market coverage analysis of the post-Stryker landscape identifies precisely this challenge: an attack too targeted to constitute systemic warfare-scale damage, too politically sensitive to attract formal government attribution on a claims-relevant timeline, and too destructive to be processed through the conventional ransomware coverage framework [11]. In the absence of a coverage category designed for this scenario, the practical resolution may increasingly follow the Merck pattern — protracted dispute followed by negotiated settlement — applied at greater frequency and across a wider range of claim sizes as state-adjacent destructive attacks proliferate.

Market Response and the Emerging Coverage Architecture

Munich Re reported the global cyber insurance market at $15.3 billion in 2024, while warning that first-party claims exposure from agentic AI-driven incidents remains materially underestimated by current actuarial models [12][13]. The market's response to this uncertainty has been additive exclusion rather than coverage innovation: a number of leading carriers have begun introducing AI-specific exclusions targeting hallucination-related losses, algorithmic failures, and autonomous AI system behavior — a trend documented across both primary insurers and specialty markets [22]. These exclusions layer onto an already exclusion-laden cyber policy landscape, narrowing effective coverage from multiple directions simultaneously.

For enterprises operating in sectors with elevated state-sponsored targeting exposure, the cumulative effect is significant. A destructive AI-assisted attack attributed to a state-linked group could simultaneously trigger war exclusion defenses (if the carrier views the attribution as sufficient) and AI-related exclusions (if the attacker employed AI-generated malware in the delivery mechanism). Captive insurance structures have emerged as one market response, enabling larger enterprises with dedicated risk management resources to retain coverage for state-backed attack scenarios that the commercial market excludes, but captives are economically practical primarily for that narrower segment of the enterprise market [13].

Recommendations

Immediate Actions

Security and legal teams should conduct a targeted review of cyber insurance policy language before the next renewal, focusing on three specific provisions: the definition of "state-backed" attack and whether it requires formal government attribution or permits insurer-led evidentiary assessment; the threshold language governing what severity of impact triggers exclusion; and any AI-specific exclusions introduced since the policy's last renewal cycle. Policyholders should request written clarification from brokers on how their specific carrier has interpreted each provision following the Stryker incident and the LMA5567A updates published in 2025.

Incident response plans should be updated immediately to include insurance-specific documentation protocols. When a destructive or state-adjacent attack occurs, the organization should begin contemporaneous documentation of: the absence of a ransom demand (indicating destructive rather than financially motivated intent), any indicators of state-actor tooling or infrastructure visible in forensic analysis, the scope of operational continuity impact, and all available threat intelligence reports linking observed TTPs to known APT groups. This contemporaneous record will be the primary evidentiary basis for any subsequent coverage dispute, and evidence quality degrades rapidly in the days following an incident.

Short-Term Mitigations

At the next policy renewal, organizations with elevated exposure to state-sponsored targeting should negotiate explicit carve-backs for attacks attributed to state-linked actors that do not produce "major detrimental impact" at the systemic level. Brokers specializing in cyber risk can often secure endorsements or supplemental coverage addressing the coverage gray zone illustrated by the Stryker scenario, though the pricing environment for state-actor adjacent coverage has tightened considerably in the current market [22].

Organizations in the defense industrial base, critical infrastructure, healthcare with government contracts, and financial services should engage specialist insurance counsel prior to renewal. The distinction between LMA5567A (which includes attribution mechanics) and LMA5567B (which does not and requires separate Lloyd's approval) carries material coverage implications that non-specialist brokers may not adequately surface during the underwriting process [7][19]. Understanding which clause variant applies and whether the policy's attribution procedure aligns with the likely evidence profile of a state-adjacent incident should be a standard renewal checklist item.

Strategic Considerations

The cyber insurance market's evolving approach to war exclusions signals a structural tension that extends well beyond policy language: the industry faces significant challenges pricing and underwriting coverage reliably for a threat category in which the boundary between crime and warfare is systematically blurred by technology and by deliberate adversary strategy. Over the medium term, organizations should not treat cyber insurance as a primary backstop for state-sponsored attack scenarios. Rather, the strategic posture should position insurance as a complement to — not a substitute for — direct investment in detection, response, and resilience capabilities.

As the Government Accountability Office observed in 2022 — before LMA5567A's framework took effect — the federal government had not clearly defined its role in responding to catastrophic cyberattacks and the question of a federal backstop program analogous to the Terrorism Risk Insurance Act remained unresolved [20]. As AI-assisted attacks make attribution more contested and LMA5567A's threshold framework harder to apply consistently across the spectrum of state-adjacent incidents, legislative and regulatory pressure for a federal cyber insurance backstop may intensify. Security and risk leaders should monitor this space and engage government affairs counterparts proactively, as the policy design of any backstop program will directly affect the underwriting landscape for state-linked cyber risk.

CSA Resource Alignment

Several CSA frameworks and initiatives address the threat landscape and control requirements relevant to this analysis.

The **AI Controls Matrix (AICM)** provides a comprehensive framework for managing AI-specific risks, including those introduced by adversarial AI use in offensive operations. Organizations should map their AI security controls against the AICM to identify gaps that affect both security posture and insurability, as carriers are beginning to incorporate AI-specific control standards into underwriting discussions — a trend the AICM is well-positioned to address [18]. The AICM's coverage of model integrity, supply chain risks, and agentic AI governance directly corresponds to the AI obfuscation and attribution challenges described in this note.

MAESTRO, CSA's Agentic AI Threat Modeling framework, addresses the threat surface introduced by autonomous AI agents — including their exploitation by state and non-state threat actors conducting multi-stage, adaptive operations. The attribution challenges described here, particularly around AI malware that dynamically mimics other actors' TTPs, are within MAESTRO's scope. Organizations using

MAESTRO-aligned threat models may find that systematic documentation of attack vectors — including AI agent behaviors — provides a cleaner evidentiary foundation for insurance claims assessments, particularly where adjusters must evaluate the sophistication and attribution of an incident.

The **STAR (Security Trust Assurance and Risk) program** and **Cloud Controls Matrix (CCM)** provide structured mechanisms for documenting an organization's security control environment. In the cyber insurance context, STAR certification and CCM-aligned control inventories serve as credible evidence of risk management maturity, enabling more favorable exclusion negotiations and reducing the likelihood of blanket state-actor exclusion language being applied at renewal.

CSA's annual **State of Cloud and AI Security** reports provide longitudinal benchmarking data on industry preparedness relative to the evolving AI threat landscape [17]. With over half of organizations already deploying AI systems and more than a third reporting AI-related breaches, the gap between adoption and security maturity identified in CSA's research maps directly to the underwriting risk that carriers are attempting to manage through expanded exclusion language.

References

[1] Insurance Journal. "Merck Settles Coverage Dispute With Insurers Over War Exclusion in NotPetya Attack." Insurance Journal, January 5, 2024.

[2] Pro Policyholder. "Merck Settlement of $1.4 Billion Coverage Dispute Over NotPetya Cyberattack." Pro Policyholder, January 2024.

[3] Cybersecurity Dive. "Merck reaches settlement in closely watched NotPetya insurance case." Cybersecurity Dive, January 2024.

[4] Lloyd's of London. "Market Bulletin Y5381: Cyber-attack Exclusions." Lloyd's, August 2022.

[5] Lloyd's Market Association. "War and Cyber Operation Exclusion No. 4: LMA5567A." LMA, September 2025.

[6] DAC Beachcroft. "War exclusions in cyber policies: an overview." DAC Beachcroft, 2023.

[7] Clifford Chance. "Lloyd's cyber war exclusion." Clifford Chance Insurance Insights, September 2023.

[8] DWF Group. "Lloyd's requirements for state-backed cyber-attack exclusions." DWF Group, October 2022.

[9] Insurance Journal. "Stryker Attack Mirrors Tactics Used in Iran-Aligned Hacks." Insurance Journal, March 15, 2026.

[10] Claims Pages. "Stryker Cyberattack Tests War Exclusion Clauses in Cyber Insurance Policies." Claims Pages, March 23, 2026.

[11] Ryan Specialty. "Iran, Cyber War and Your Policy: War Exclusions, Coverage Implications and What Policyholders Need to Know Now." Ryan Specialty, 2026.

[12] Industrial Cyber. "Munich Re Sees Untapped Potential in $15.3B Cyber Insurance Market Amid Rising Threats and Evolving Risks." Industrial Cyber, 2026.

[13] The Insurer. "Munich Re warns agentic AI-driven cyberattacks could spur first-party claims." The Insurer, March 26, 2026.

[14] Google Cloud / GTIG. "GTIG AI Threat Tracker: Advances in Threat Actor Usage of AI Tools." Google Cloud Blog, 2025.

[15] Recorded Future. "AI Malware: Hype vs. Reality." Recorded Future Blog, 2025.

[16] Deep Instinct. "The Rise of AI-Driven Cyber Attacks: How LLMs Are Reshaping the Threat Landscape." Deep Instinct, 2025.

[17] Cloud Security Alliance. "The State of Cloud and AI Security 2025." CSA, 2025.

[18] Cloud Security Alliance. "AI Controls Matrix." CSA.

[19] Cyber Insurance Academy. "LMA5567A/B: A 2026 Market Update." Cyber Insurance Academy, 2026.

[20] US Government Accountability Office. "Cybersecurity: Federal Agencies Made Progress, but Need to Fully Implement Incident Response Requirements." GAO-22-104256, April 2022.

[21] TechCrunch. "US accuses Iran's government of operating hacktivist group that hacked Stryker." TechCrunch, March 20, 2026.

[22] Business Insurance. "Insurers, Brokers Adjust as AI Exclusions Emerge." Business Insurance, 2026.