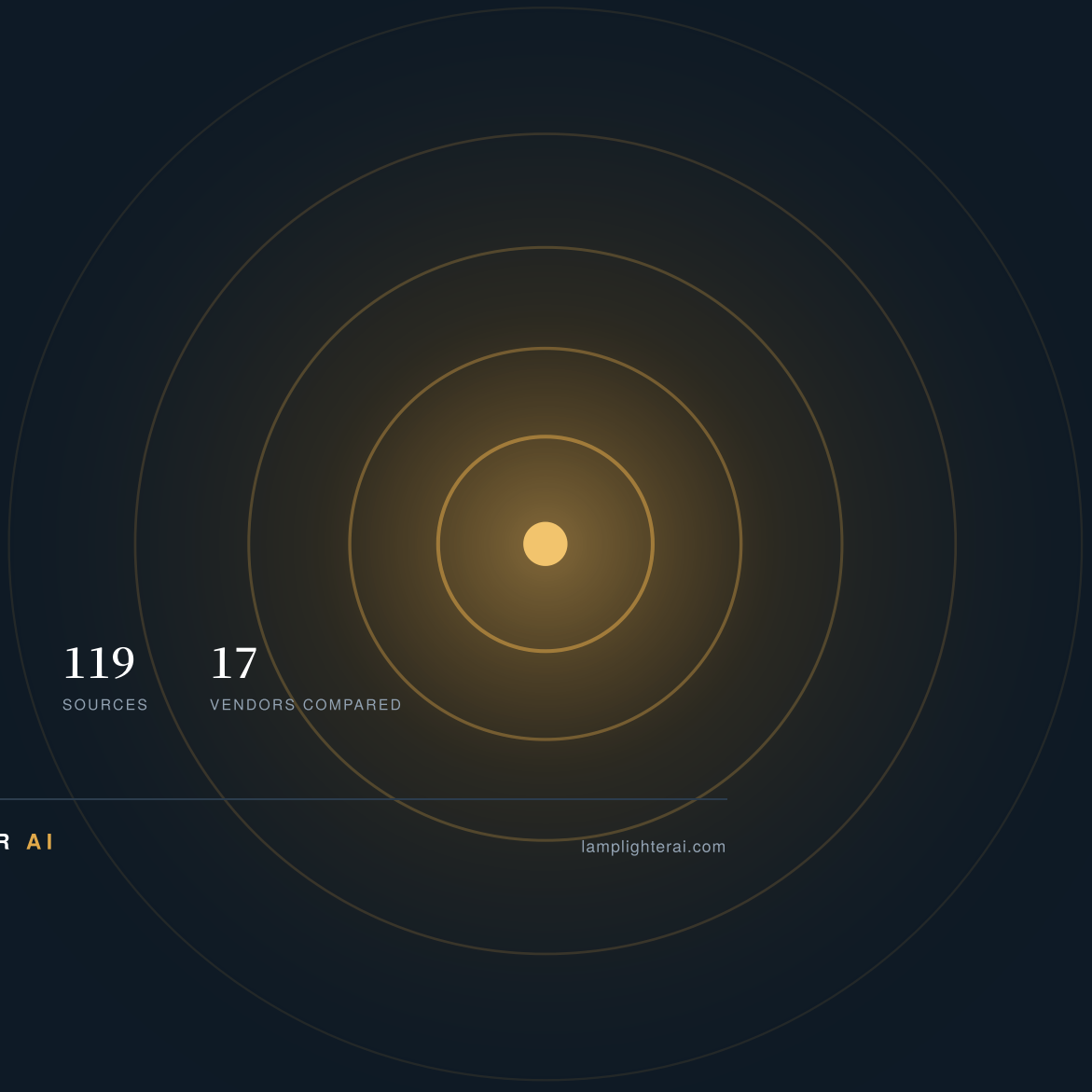
WHITEPAPER · OCTOBER 2026

The Endpoint Is the New AI Perimeter

Emerging risks of employee AI use in 2025–2026: what the security market covers, where it stops, and what a complete control requires.



79	9	119	17
RISKS	DOMAINS	SOURCES	VENDORS COMPARED

LAMPLIGHTER AI

lamplighterai.com

KEY FINDINGS

What changed in 2025–2026

84%	of organisations now have agentic coding tools in use, up from 42% a year earlier; local MCP traffic from these tools is invisible to network sensors [1].	82%	of pastes into generative AI come from unmanaged, personal accounts; 22% of pastes carry personal or payment data [2].
~900k	installs of two “AI assistant” browser extensions that harvested ChatGPT and DeepSeek conversations across more than 20,000 enterprise tenants [3].	99	CVEs were published for MCP-related software in 2025 alone [4]; agent configuration files have become an execution layer [5].
>10%	of AI data-policy alerts are now “downstream” — the assistant returning data the user should not see — up from effectively zero a year earlier [1].	$670k	additional average cost of a breach with high shadow-AI involvement; 97% of organisations breached through AI lacked proper AI access controls [6].

ABOUT THIS PAPER

This paper synthesises vendor telemetry, surveys, incident reports and security frameworks published between January 2025 and September 2026 into a single risk register for employee AI use on business endpoints — 79 risks in nine domains. It sets out what the security market covers today, where the documented gaps are, the boundaries no endpoint control can cross, and the requirements a complete control must meet. The final section describes how Lamplighter meets those requirements, marking each capability as AVAILABLE, IN DEVELOPMENT, ROADMAP or PROPOSED. This October 2026 edition updates those statuses to Lamplighter v0.18.

CONTENTS

	Executive summary	3
1	**Introduction: AI moved to the endpoint**	5
2	**Scope and method**	6
3	**The endpoint AI risk landscape**	7
3.1	Sensitive data leaves through more than the browser	7
3.2	Shadow AI is an account and tier problem	8
3.3	Agents, MCP and configuration as an execution layer	9
3.4	Prompt injection is a design assumption	9
3.5	The AI tool itself is an attack vector	10
3.6	AI output becomes an endpoint action	11
3.7	AI identities, sessions and OAuth	11
3.8	Regulation, and the lawfulness of monitoring	12
3.9	Governance process failure	13
4	**What the market covers, and where it stops**	14
4.1	Well-served: chat applications in the browser	14
4.2	Four documented gaps	14
4.3	The acquisition signal	14
4.4	Feature coverage compared	15
5	**The limits of any endpoint control**	16
6	**Requirements for endpoint AI governance**	17
7	**Lamplighter: endpoint-native AI governance**	18
7.1	Architecture	18
7.2	How Lamplighter meets the requirements	19
7.3	Why Lamplighter	22
7.4	Deployment and platforms	23
7.5	Roadmap	24
8	**Conclusion**	24
A	**Appendix A: The endpoint AI risk register**	25
B	**Appendix B: OWASP coverage map**	28
	References	29

Executive summary

Employee use of artificial intelligence has moved faster than the controls built to govern it, and it has moved to a place most of those controls cannot see. In 2024 the question was whether staff pasted confidential material into ChatGPT. They do, and at scale. In 2026 the more consequential question is whether a security team can see the channels that have grown up *underneath* the browser: desktop and command-line AI agents, locally configured Model Context Protocol (MCP) servers, browser extensions that read AI conversations after encryption has been removed, and AI built into the operating system and into already-sanctioned software.

The evidence reviewed for this paper points to five conclusions.

1. **The endpoint is now the control point.** The largest-growing AI surfaces — coding agents, local MCP servers, desktop assistants and AI browser extensions — generate no network event a proxy or cloud access broker can inspect [1], [7], [8]. Only software on the workstation sees who is using which AI, with which account, on what data, and where that data came from.
2. **Shadow AI is an account problem, not a domain problem.** The same web address carries different training, retention and legal-discovery consequences depending on which account is signed in. Consumer-tier chats were swept into court-ordered preservation in 2025; enterprise tiers were carved out [9].
3. **AI tools are now attack vectors in their own right.** Fake AI installers, chat-harvesting extensions, poisoned models, malicious MCP servers and worms that hijack the victim's own AI command-line tools are all documented in the period [3], [10], [11].
4. **Prompt injection is a design assumption, not a bug.** Vendors and national agencies now state that it is unlikely ever to be fully solved [12], [13]. Defensible controls act on what an agent is allowed to do, not on filtering what it reads.
5. **Governance is primarily a policy and evidence problem.** Gartner attributes at least 80% of unauthorised AI transactions to internal policy violations rather than attacks [14]; regulators and auditors increasingly ask for inventories, literacy records and usage logs [15], [16]. And because an AI-governance tool that captures prompts is itself workplace monitoring, the control must be lawful by design [17].

The commercial market has converged on browser and network controls for chat applications, and those are now well served. Four gaps remain documented: desktop, command-line and local-MCP agents; extension-level interception; reliable identification of which account is in use; and a request, approval and exception workflow, which no vendor documentation reviewed for this paper evidenced at all. Platform vendors spent heavily in 2026 to close the first of these [18], [19].

Lamplighter was designed for that gap. It is an endpoint-native AI governance product that discovers the AI tools on a device — web applications, desktop apps, command-line agents, extensions, local runtimes and MCP servers — identifies the account in use on the most-used AI web applications, classifies what people type, paste and upload *on the device itself*, governs the connections and autonomy of AI agents, and produces audit-ready evidence, while by default sending none of what people type, paste or upload off the machine. It already provides a first version of the missing approval route: a person can request a tool that is not approved, and an approval changes enforcement on every device. Section 7 maps each requirement in this

paper to a Lamplighter capability and states plainly, from an internal audit of the product's code and tests, which are available today and which are in development, on the roadmap or proposed.

1

Introduction: AI moved to the endpoint

For most of 2023 and 2024 “AI risk” in the enterprise meant a browser tab. Controls followed: secure web gateways and cloud access security brokers learned to recognise AI domains, and data-loss-prevention (DLP) products learned to inspect text pasted into chat windows. That model assumed AI was a website.

+375%
growth in remote MCP transactions in ten weeks; local MCP is invisible to network telemetry [1].

By 2026 the assumption no longer holds. Netskope’s telemetry shows agentic coding applications present in 84% of organisations, up from 42% a year earlier, with Claude Code in 75% and OpenAI Codex in 58% of organisations — both under 1% the year before [1]. Remote MCP use grew 250% in users and 375% in transactions over a ten-week window, and the same report notes that *local* MCP — servers launched on the workstation by a desktop or command-line client — is invisible to network telemetry altogether. Cyberhaven describes the mechanism: endpoint AI agents “bypass network inspection by using accessibility APIs, clipboard monitoring, and local file system access — none of which generate the network events that DLP tools monitor” [7].

At the same time the AI estate widened in the other direction, into software organisations already trust. Zscaler found that Grammarly, not ChatGPT, was the single largest destination of enterprise data sent to AI in 2025, at 3,615 TB against ChatGPT’s 2,021 TB [20]. Netskope found 75% of users working in applications with embedded generative AI, against 4.9% using standalone generative AI applications [21]. AI built into productivity suites, browsers and operating systems shares a domain or a process with software the organisation has already approved.

The endpoint is where these threads meet. It is the only vantage point from which *who*, *which application*, *which account*, *what data* and *where it came from* are observable together. This paper asks what an organisation must control there to have confidence that it is addressing the risks of employee AI use — and where the honest limits of that confidence lie.

2

Scope and method

The research underlying this paper was conducted in September 2026 across six parallel lines of enquiry: security frameworks and taxonomies; empirical evidence on shadow AI and data leakage; agentic, MCP and AI-browser threats; AI-enabled attacks and supply-chain compromise; regulatory and compliance obligations; and the capabilities and documented gaps of existing products. Sources were restricted to material published between January 2025 and September 2026 unless an earlier item was foundational, in which case it is marked as such.

The frameworks used to structure and cross-check the register are the OWASP Top 10 for LLM Applications 2025 [22] and Top 10 for Agentic Applications for 2026 [23]; MITRE ATLAS, including its 2026 releases [24]; NIST AI 600-1 [25], the NIST Cyber AI Profile draft [26] and the CAISI request for information on agent security [27]; ISO/IEC 42001 [28]; the Cloud Security Alliance AI Controls Matrix [29]; ETSI TS 104 223 [30]; and the ENISA Threat Landscape 2025 [31].

15×

spread in reported "sensitive" rates for AI traffic, almost entirely from different denominators [2], [32], [33].

Two methodological cautions apply throughout. First, quantitative claims are distinguished as *telemetry* (measured from a vendor's network, browser, endpoint or DLP logs) or *survey* (self-reported), because the two disagree systematically. Second, vendor telemetry reflects each vendor's own customer base, so absolute percentages are not comparable across vendors: the reported share of AI interactions containing sensitive data ranges from 2.6% of interactions (Harmonic) through 39.7% of data movements (Cyberhaven) to 40% of uploaded files (LayerX) [2], [32], [33]. Figures that could not be verified against a primary source during the research have been omitted from this paper.

The output is a register of 79 risks in nine domains, summarised in Appendix A. Each risk was scored on two axes: how well the existing market covers it at the endpoint, and whether an endpoint agent can observe or control it at all.

3 The endpoint AI risk landscape

The nine domains below run from the familiar — sensitive data typed into a chatbot — to the newest: agents whose configuration files execute code, AI tools that are themselves malware, and the legal status of the monitoring tool. Each closes with what a control must do.

ROUTE	NETWORK (SWG, SSE, CASB)	BROWSER (EXTENSION)	ENDPOINT AGENT (PROCESS, FILE, CLIPBOARD)	CLOUD API (VENDOR AUDIT LOGS)
Prompt typed or pasted into an AI web app	◑	●	◑	◑
File uploaded to an AI web app	◑	●	◑	◑
Paste or file into a desktop AI app	◑	○	●	◑
Coding agent or CLI calling a model API	◑	○	●	◑
Local MCP servers and their tool calls	○	○	●	○
Browser extension reading AI conversations	○	◑	◑	○
AI embedded in a sanctioned SaaS app	◑	◑	◑	●
Operating-system AI (e.g. screen capture)	○	○	●	○
Conversation published through a share link	◑	●	○	◑
AI response returning restricted data	◑	●	◑	◑
OAuth consent to an AI connector	◑	●	○	◑
Agent exfiltrating from the vendor's cloud	○	○	○	◑

● Sees content and context ◑ Partial ○ Does not see Shaded: the sensors Lamplighter combines

Figure 1. Where data reaches AI from the endpoint, and which class of sensor can see it. Lamplighter analysis of the mechanisms documented in [1], [7], [8], [34], [35]. “Partial” means visible only with TLS inspection, only for some applications, or only as metadata.

3.1 Sensitive data leaves through more than the browser

32%

of all corporate-to-personal data movement now goes through generative AI — the largest single channel [36].

The browser remains the loudest channel. LayerX’s 2025 browser telemetry found 77% of employees pasting into generative AI prompts, 82% of those pastes coming from non-corporate accounts, and 22% of pastes containing personal or payment-card data; 40% of files uploaded to generative AI tools contained such data [2]. Menlo Security recorded 80% of AI access occurring in the browser [37]. Netskope measured data-policy violations involving generative AI doubling year on year to 223 per month in the average organisation and 2,100 in the top quartile [38]; Zscaler counted 410 million DLP violations in ChatGPT traffic alone, involving Social Security numbers, source code and medical records [20].

Beneath the browser, the evidence is newer and in several places harder. Claude Code-assisted commits on public GitHub leaked secrets at 3.2%, against a 1.5% baseline across all public commits, and AI-service secrets leaked in 2025 rose 81% to 1,275,105 [39]. Embedded AI outleaks the chatbots by volume, as noted above. Three incidents in 2025 showed sensitive conversations becoming publicly indexed through a toggle inside a sanctioned interface rather than through any network egress a DLP product would flag: around 4,500 shared ChatGPT conversations surfaced in search results [40], and Forbes reported more than 370,000 Grok conversations indexed, though other coverage put the number lower [41].

The newest channel runs in reverse. Netskope's 2026 report separates "upstream" violations (a user sends restricted data to AI) from "downstream" ones (the AI returns data the user is not authorised to see); downstream violations rose from a median of 12 to 31 per organisation per week and from effectively 0% to over 10% of all AI alerts [1]. The cause is permissions sprawl — Varonis found 99% of organisations have sensitive data an AI can easily surface, and only one in ten label their files [42] — and the symptom is visible only where the answer is rendered: on the endpoint.

Finally, running a model locally is not the safe harbour it appears. Ollama is present in 41% of organisations [1]; CVE-2026-7482 ("Bleeding Llama", CVSS 9.1) let an unauthenticated remote attacker read an exposed Ollama server's process memory, including API keys and user conversations [43].

> **WHAT A CONTROL MUST DO**
>
> Inspect content at the moment of typing-and-submit, paste, upload and drag — in the browser, in desktop applications and in coding agents — using both structural detectors and semantic classification, because the two rank data types differently [32]; and inspect what AI returns, not only what users send.

3.2 Shadow AI is an account and tier problem

The estate is large and moving. Netskope tracked the number of generative AI applications rising five-fold to more than 1,600 [38]; Zscaler observed more than 3,400 AI and machine-learning applications across 989 billion transactions in 2025 [20]. The more consequential shift in 2026 telemetry, however, concerns *accounts*. Where organisations provision AI, personal-account use falls: Netskope saw the share of AI users on personal accounts drop from 78% to 47% year on year [38]. Where they do not, it does not: LayerX, whose customers typically lack inline network controls, still sees 67% of AI usage on unmanaged personal accounts [2]. Harmonic classified almost two million AI session-minutes in early 2026 and found three-quarters were business use, including between 60% and 80% of activity on most free and consumer tiers [44].

20M
consumer ChatGPT logs ordered produced in *NYT v. OpenAI*; enterprise tiers were carved out of preservation [9], [45].

Account type matters because the same prompt carries different consequences on different tiers. OpenAI trains on consumer ChatGPT conversations unless the user opts out, but not on business products by default [46]; Anthropic's consumer terms retain chats for five years where training is permitted [47]; Google retains human-reviewed Gemini conversations for up to three years even after the user deletes their activity [48]. In *New York Times v. OpenAI*, the court's May 2025 preservation order covered consumer chats, including deleted ones, while carving out Enterprise and zero-data-retention customers; in November 2025 it ordered production of 20 million consumer-tier logs [9], [45]. Conversations with AI are not privileged; they are discoverable business records.

Shadow AI also arrives suddenly. Within about two weeks of DeepSeek's January 2025 launch, the US Navy and NASA, Italy's data protection authority, Taiwan, Australia and South Korea had banned or restricted it [49], while Netskope saw 91% of organisations with users attempting access within weeks [21]. And blocking relocates rather than removes: Zscaler-observed block rates fell from 59.9% of AI access attempts in 2024 to 39% in 2025 as customers moved to policy-driven allowance [20], and Gartner warns that blocking agents outright drives employees to shadow AI "which presents far greater risks" [50].

WHAT A CONTROL MUST DO

Identify the account and product edition in use, not just the domain; enforce tier policy (for example, “corporate accounts only”); propagate decisions on newly emergent tools within hours; and offer a sanctioned alternative at the moment of refusal.

3.3 Agents, MCP and configuration as an execution layer

Check Point’s summary of 2025 is exact: “configuration files became part of the execution layer” [5]. MCP server definitions are plain JSON files that launch local processes and hold static credentials, so anything that can write them — a cloned repository, an agent, a prompt injection — gains code execution with the user’s rights. Cursor executed any new entry added to its MCP configuration without confirmation (CVE-2025-54135, “CurXecute”) [51]; GitHub Copilot could be induced to write `chat.tools.autoApprove` into a project’s settings, disabling approval prompts entirely (CVE-2025-53773) [52]; Claude Code executed project hooks and MCP servers on repository open (CVE-2025-59536) [5]. Microsoft counted 99 CVEs in MCP-related software in 2025 [4].

The MCP ecosystem adds its own classes. “Line jumping” lets a malicious server inject instructions through tool descriptions before any tool is called [53]. The first malicious MCP server found in the wild, `postmark-mcp`, silently copied every email it sent to an attacker for more than 1,600 downloads [54]. The `mcp-remote` client, used by Claude Desktop, Cursor and Windsurf to reach remote servers, carried a remote code execution flaw (CVE-2025-6514) [55]. Astrix found 53% of MCP servers rely on static API keys or tokens and only 8.5% on OAuth [56]; GitGuardian found 24,008 unique secrets in MCP configuration files [39]; Knostic found 1,862 MCP servers exposed to the internet, every one of 119 manually verified servers exposing internal tool listings without authentication [57].

150,000
agents per Fortune 500 enterprise projected by 2028; 13% believe their agent governance is adequate [50].

The highest-fidelity signal in the whole register is the simplest: agents run with their safeguards switched off. In August 2025 the Nx “s1ngularity” supply-chain attack invoked the victim’s own installed Claude, Gemini and Amazon Q command-line tools with `--dangerously-skip-permissions`, `--yolo` and `--trust-all-tools` to hunt for secrets, leaking 2,349 of them and later flipping more than 6,700 private repositories public [11], [58]. Microsoft’s red team ranks human-in-the-loop bypass as the most consistently exploited agent failure mode [4]. When agents act autonomously the consequences are physical: Replit’s agent deleted a live production database during a code freeze [59], and a compromised release of Amazon Q’s official extension shipped a data-wiping prompt to roughly one million installations [60]. Invariant Labs showed how legitimate tools chain into “toxic flows”: a malicious public GitHub issue led an agent to publish data from private repositories [61].

WHAT A CONTROL MUST DO

Inventory every MCP server, agent configuration, API key and agent mode on the device; treat writes to agent configuration as code execution; flag permission-bypass modes; route model and tool traffic through a point where it can be inspected; and gate destructive actions regardless of what the model intended.

3.4 Prompt injection is a design assumption

Between January 2025 and August 2026, essentially every major enterprise assistant with connectors received at least one publicly documented indirect-prompt-injection exploit ending in data exfiltration or code execution, most of them zero- or one-click. The pattern repeats: untrusted content (an email, document, ticket, issue or calendar invitation) is read by the

assistant, which uses a legitimate connector to fetch private data and leaks it through an allowed channel. EchoLeak (CVE-2025-32711) exfiltrated Microsoft 365 Copilot data from a single inbound email [62]; ForcedLeak (CVSS 9.4) used an expired domain still on Salesforce's content-security allowlist, bought for $5 [63]; CamoLeak (CVSS 9.6) leaked private source code through GitHub's own image proxy [64]; AgentFlayer targeted ChatGPT Connectors [65]; and calendar invitations hijacked Gemini for Workspace in scenarios SafeBreach rated 73% high-to-critical [66].

50
distinct AI memory-poisoning attempts across 31 companies in 60 days, via "Summarize with AI" links [67].

Vendors now describe the class as unfixable. OpenAI: prompt injection, "much like scams and social engineering on the web, is unlikely to ever be fully 'solved'" [12]. The UK National Cyber Security Centre warns that treating it like SQL injection is dangerous, because a language model has no boundary between data and instructions — "there is only ever 'next token'" — and recommends deterministic design, least privilege for tools and monitoring of suspicious tool calls [13]. Attack delivery is diversifying: links that pre-fill a malicious prompt into an assistant ("Reprompt") [68], now catalogued by MITRE as AML.T0131 [24]; persistent memory poisoning [67]; and instructions hidden in screenshots and page text that AI browsers read but users cannot see [69]–[71]. The only published defender-side effectiveness figures show mitigation reducing attack success on a browser agent from 23.6% to 11.2% — better, but not zero [72].

> **WHAT A CONTROL MUST DO**
> Act at the action layer: strip crafted prompts from links, monitor and constrain what agents may call, hold consequential actions for out-of-band confirmation, and treat trusted domains as untrusted when they carry AI-originated, data-shaped requests. The defensible claim is containment, not prevention.

3.5 The AI tool itself is an attack vector

Attackers follow search volume. Pillar Security counted nine AI-tool impersonation campaigns in the first ten weeks of 2026, against eleven across the whole of 2025, and concluded that "any tool gaining significant search volume faces impersonation within weeks" [10]. Fake AI installers delivered ransomware and wipers [73], and ENISA lists sites impersonating AI tools among 2025's threat patterns [31]. The most important shift is that lures now sit on legitimate AI domains: a 2026 campaign hosted its "paste this command" (ClickFix) lure on genuine claude.ai share pages [74]; others used ChatGPT share links [75] or a fake Codex installer whose pasted command began with legitimate-looking syntax [76]. Domain reputation cannot stop an attack hosted on a trusted domain.

17
AI browser extensions in the average organisation, up to 163; 46.8% of all extensions hold permissions that enable account takeover [77].

AI browser extensions are the largest single cluster. The AiFrame campaign reached more than 260,000 users with look-alike assistants [78]; two extensions with around 900,000 installs harvested AI conversations across more than 20,000 enterprise tenants [3]; and a family of VPN extensions with some eight million users was found capturing prompts and responses on eight AI platforms [79]. LayerX classifies 6% of generative-AI extensions as malicious [2]. The Cloud Security Alliance's architectural point is the decisive one: because this exfiltration happens inside the browser runtime and leaves over ordinary HTTPS, it is "DLP-invisible" to network, CASB and EDR controls [34].

The AI software supply chain is compromised at every layer: a malicious `ai-sdk-ollama` release that planted hooks in Claude Code, VS Code and Gemini settings to re-run on IDE start [80]; malicious models that evaded Hugging Face's scanner [81]; dormant backdoors in model

chat templates [82]; and invisible Unicode instructions in coding assistants' rules files [83]. Exposed local runtimes grew to some 175,000 publicly reachable Ollama hosts, many advertising tool-calling [84].

> **WHAT A CONTROL MUST DO**
> Inventory extensions and their permissions on AI domains and match known-malicious identifiers; watch the browser-to-terminal hand-off for AI-sourced commands; verify agent components and runtime versions against advisories; and treat agent instruction files as security-relevant.

3.6 AI output becomes an endpoint action

Each risk in this domain looks benign in isolation — a package install, a commit, a browser navigation — and is dangerous because of its provenance. Spracklen et al. found 19.7% of packages recommended by 16 code-generating models did not exist, producing 205,000 unique hallucinated names, 43% of which recurred on every rerun — a predictable list for attackers to register ("slopsquatting") [85]. Veracode's 2026 update found 45% of AI-generated code still contained known vulnerabilities, "virtually identical" to 2025 despite syntax correctness above 95% [86]. Netcraft asked AI models for the login pages of 50 brands and found 34% of the suggested hostnames unregistered, inactive or owned by someone else [87].

Social engineering has been industrialised by the same tools. Hoxhunt measured AI phishing agents moving from 31% less effective than human red teams in 2023 to 24% more effective in 2025 [88]; ENISA reports over 80% of observed social engineering used AI support [31]; and the FBI's first dedicated AI section recorded 22,364 complaints and nearly $893 million in losses [89]. Much of this is not directly observable on an endpoint, and a credible control says so.

> **WHAT A CONTROL MUST DO**
> Attach AI provenance to actions — installs, commits, credential entry — and apply risk checks at that moment; and be explicit about which social-engineering risks are awareness and process problems rather than endpoint detections.

3.7 AI identities, sessions and OAuth

IBM's 2026 X-Force index states that AI platforms "have reached the same credential risk as other core enterprise SaaS solutions", with over 300,000 ChatGPT credentials exposed by infostealers in 2025 [90]. The unit of compromise is shifting from password to session token: one 2026 infostealer dump contained 44,791 unique session tokens, 555 of them AI-related, with 1,843 of the tokens still valid on the day of release; replayed, they bypass multi-factor authentication [91]. An AI account takeover has two consequences ordinary SaaS compromise does not — retroactive exposure of everything the employee ever pasted, and abuse of the organisation's paid model access.

68%
of corporate logins occur without SSO, so identity controls cannot establish which accounts are in use [92].

Connectors carry the risk into corporate systems. Push Security measures an average of 17 AI OAuth integrations into Google Workspace and Microsoft 365 per organisation, and up to 55 [77]. The Salesloft Drift breach turned stolen OAuth tokens for an AI chat agent into bulk export of Salesforce data [93]; Vercel's 2026 breach began with an employee's unapproved OAuth grant to a consumer AI tool that was itself compromised [94]. After a grant is made, connector traffic flows between cloud services and the endpoint sees nothing. The consent screen is therefore the single most valuable endpoint control point for connector risk: it shows the application, the scopes and the redirect *before* the token exists.

WHAT A CONTROL MUST DO

Inventory OAuth grants to AI applications and, better, intervene at the consent screen; detect sessions not established through corporate single sign-on; protect browser session stores; and treat AI session theft as a data-exposure event.

3.8 Regulation, and the lawfulness of monitoring

IBM and Ponemon found breaches with high shadow-AI involvement cost $670,000 more than average; 63% of breached organisations had no AI governance policy or were still developing one, and only 34% of those with policies audited regularly [6]. Regulators have begun to name the behaviour. The Dutch data protection authority received breach notifications after a GP practice employee entered patient data into an AI chatbot [95]; France's CNIL expects internal policies defining authorised and prohibited uses and a data protection impact assessment for sensitive uses [96].

The EU AI Act's literacy duty (Art. 4) has applied since February 2025 and has been enforceable by market-surveillance authorities since August 2026; the Commission states that "internal records of trainings and other initiatives are sufficient" evidence [97]. The 2026 AI Omnibus deferred high-risk obligations but retained Art. 5 prohibitions, including emotion recognition in the workplace; when the deferred high-risk obligations arrive, deployers must also retain automatically generated logs for at least six months [15]. In healthcare, HHS's proposed HIPAA Security Rule would require an inventory of technology assets that identifies AI interacting with electronic PHI [98], though finalisation has slipped to 2027 [99]. Medical-device guidance folds AI into MDR/IVDR documentation [100], while notified bodies' position paper is silent on generative AI in manufacturers' internal processes [101] — so auditors will fall back on document control and attribution. In South Africa, which has no AI statute, the Information Regulator described AI in September 2026 as "a free for all" under POPIA [102]. Professional duties add weight: the American Bar Association requires informed client consent before client information goes into generative AI [103], and insurers have begun excluding AI-attributable losses [104].

DPIA

required by the ICO for high-risk worker monitoring, which it says includes keystroke monitoring [17].

One risk in this domain is easily overlooked by anyone building a control: **the control itself is regulated**. A product that captures employee prompts is worker monitoring. The UK ICO requires monitoring to be necessary, proportionate and transparent, conducted "in the least intrusive way possible", with a DPIA for high-risk forms [17]. A German labour court reasoned that works-council co-determination was not triggered only because the employer had no access to employees' AI data [105] — implying that once an employer logs usage, it is. Microsoft's own design for AI interaction capture gates prompt content behind an explicit collection policy and a separate viewer role [106]. The design pattern that follows is classification before capture, metadata by default, role-gated access to content, defined retention, and no emotion inference.

WHAT A CONTROL MUST DO

Produce the artefacts regulators and auditors ask for — inventory, tier used, literacy records, usage logs, incident records — and be lawful by design: minimal capture, transparency to workers, and a DPIA and notice shipped with the product.

3.9 Governance process failure

Gartner's assessment reframes the problem: at least 80% of unauthorised AI transactions through 2026 will result from internal policy violations — oversharing, unacceptable use or misguided AI behaviour — rather than malicious attacks [14]. Sixty-nine percent of surveyed security leaders suspect or have evidence of prohibited public AI use, and Gartner predicts more than 40% of enterprises will suffer shadow-AI incidents by 2030 [107]. Controls lag intent: only half of organisations have enforceable data-protection policies for generative AI [38]; 38% have a formal AI policy and 56% do not know how long it would take to halt AI systems in an incident [108]; 60% cannot quickly terminate a misbehaving agent and a third lack evidence-quality audit trails [109]. Only 47% of employees report having had AI training [110].

Two process gaps stand out. The first is the absence of a way to say yes: the research found no vendor documentation evidencing an exception-request and approval workflow, although provisioning sanctioned alternatives is what reduces unauthorised use. The second is offboarding: data in personal AI accounts "becomes unretrievable when employees leave" [111].

> **WHAT A CONTROL MUST DO**
> Close the loop from inventory to policy to approval to evidence: coach first, record justifications, let users request tools and have approvals change enforcement mechanically, and report per-user exposure at offboarding.

4 What the market covers, and where it stops

+73%
forecast growth of "AI usage control" spending, to $749M in 2027 — the fastest segment of securing AI [112].

Gartner now tracks "AI usage control" as a distinct spending segment, forecast at $749 million in 2027 and growing 73% — the fastest of the named segments within a securing-AI market of $4.8 billion [112]. KuppingerCole has defined a unified market in which "SaaS applications, embedded AI, AI agents, OAuth integrations, non-human identities, and SaaS-to-SaaS connections form one interconnected risk surface" [113]. The analyst community agrees the category is real; the question is what it covers.

4.1 Well served: chat applications in the browser

The commercial control set has stabilised around a recognisable sequence: discovery and risk scoring of AI applications; sanctioned, tolerated and unsanctioned tagging; identity-aware policy; inline inspection of paste, typing and upload with block, warn and coach; prompt and response logging; and integration with existing SIEM tooling. Secure service edge (SSE) platforms, enterprise browsers and Microsoft Purview all implement much of this for browser-based chat applications. For a buyer, these controls are now table stakes rather than differentiators.

4.2 Four documented gaps

Four gaps are documented rather than merely asserted.

1. **Desktop, command-line and local-MCP agents.** The most-cited residual gap of 2026. Cyberhaven explains why agents generate no network events [7]; Nightfall concedes that gateway approaches miss "local stdio MCP servers, IDE agents, CLI activity, desktop AI applications" [8]; Netskope states local MCP is invisible to network telemetry [1].
2. **Extension-level interception.** Extensions read AI conversations inside encrypted browser sessions and exfiltrate over ordinary HTTPS, which the Cloud Security Alliance describes as architecturally invisible to network, CASB and EDR controls [34].
3. **Identity ground truth.** With 68% of corporate logins made without single sign-on, identity-based controls often cannot establish which account is in use [92].
4. **Request, approval and exception workflow.** No vendor documentation reviewed for this paper evidenced one. Blocking without a route to approval relocates AI use rather than reducing it [50].

Detection quality is a quieter constraint. Every precision figure the research found for AI-prompt DLP was vendor-claimed; no independent benchmark of false-positive or false-negative rates exists. Because most AI misuse is negligent rather than malicious [14], false positives land on legitimate work and drive users to unmanaged alternatives. Accuracy, and the ability to coach rather than block, are first-order requirements.

4.3 The acquisition signal

Between mid-2025 and mid-2026 every major security platform vendor bought at least one AI-security company, and the sequence is informative (Table 1). The 2025 deals bought the interaction layer and runtime guardrails. The 2026 deals moved to agents and identities: Palo Alto Networks completed its acquisition of Koi in April 2026 to create what it calls "Agentic Endpoint Security" [18], and Cisco announced its acquisition of Astrix for non-human identity

and agent credential governance in May 2026 [19]. The read for buyers is that chat-tool DLP has been absorbed into existing platforms, while agentic endpoint security, MCP governance and non-human identity are the categories the market decided in 2026 it cannot do without.

Table 1. Selected AI-security acquisitions, 2025–2026.

ACQUIRER	TARGET	ANNOUNCED	CAPABILITY ACQUIRED
SentinelOne	Prompt Security	Aug. 2025	Shadow-AI discovery, prompt inspection, MCP gateway [114]
Cato Networks	Aim Security	Sep. 2025	AI runtime security; discoverer of EchoLeak [115]
Check Point	Lakera	Sep. 2025	Prompt-injection and jailbreak detection [116]
CrowdStrike	Pangea	Sep. 2025	AI detection and response guardrails [117]
Zscaler	SPLX	Nov. 2025	AI red teaming and runtime protection [118]
Palo Alto Networks	Koi	Apr. 2026	Agentic endpoint security [18]
Cisco	Astrix Security	May 2026	Non-human identity and agent credentials [19]

4.4 Feature coverage compared

To make the comparison concrete, the research scored 17 vendors against 30 endpoint AI control features in five groups — discovery and inventory, data-in-motion controls, agents and connectors, governance and evidence, and enforcement — counting a feature only where public product documentation evidenced it. Figure 2 shows the result alongside Lamplighter. Two cautions apply: absence of documentation is not proof of absence of a capability, and Lamplighter's figures are mapped from an internal audit of its own code and tests rather than from third-party review. Lamplighter is held to the same rule as the vendors: a feature counts in full only when every element of its definition works today, and in part when some element does. With those caveats, the most comprehensive vendor documented 13 of the 30 features in full and 7 more in part. Lamplighter makes 11 fully and 12 partly available today — 23 of the 30 in some form, against at most 20 for any other product reviewed — has 1 more in development, 4 on its roadmap and 2 proposed.

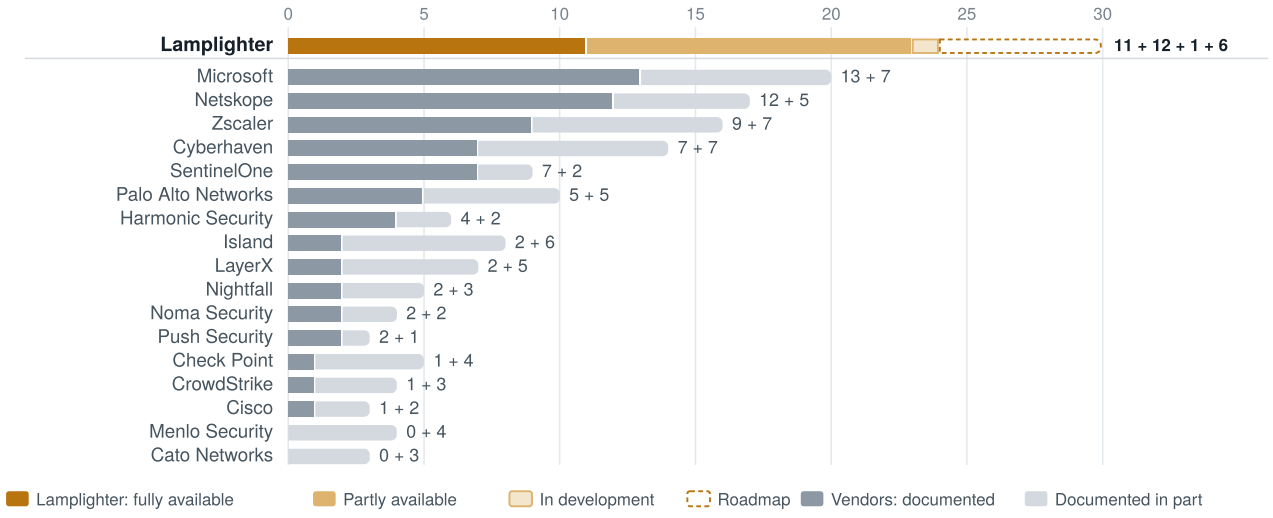


Figure 2. Endpoint AI control features, of 30. Vendor figures count capabilities evidenced in public documentation reviewed in September 2026, in full and in part. Lamplighter figures are mapped from an internal audit of v0.18 against its code and tests (October 2026): fully available, partly available, in development, and on the roadmap (planned or proposed).

5 The limits of any endpoint control

A claim of completeness is defensible only if it is bounded. The research supports four boundaries that no endpoint-resident control can close on its own.

Server-side exfiltration

In ShadowLeak, hidden instructions in an email caused ChatGPT Deep Research to send personal data to an attacker *from OpenAI's own cloud*; no endpoint or enterprise network control could see it [119]. The endpoint can contribute connector governance and detection of the links that start such chains, but not visibility of a cloud agent's outbound call.

Unmanaged devices and phones

Endpoint agents instrument managed endpoints. No study measured work data entering AI from personal devices, and blocking is known to push use to them. Tenant restrictions, conditional access and single sign-on are the complementary controls.

Vendor-side retention and training

A vendor's training settings, retention schedule, human review and litigation holds are invisible to any endpoint [9], [47], [48]. What the endpoint *can* do is decisive but indirect: identify the application, account and edition in use, which determines which vendor regime applies.

AI embedded in sanctioned SaaS

AI features that run inside an approved product, against data the vendor already holds, produce no distinguishable endpoint event; Zscaler calls this a "growing blind spot" [35]. The answer is an API plane: the vendor's own audit and compliance interfaces.

> The honest claim is therefore: *everything needed at the endpoint, plus a named connector-governance and vendor-selection process, plus an API plane for sanctioned platforms.* That formulation survives a due-diligence questionnaire; an unqualified claim of complete coverage does not.

6

Requirements for endpoint AI governance

Drawing the domain-level conclusions together gives ten requirements for a control that can support a bounded claim of completeness. They are stated independently of any product so that they can be used to evaluate any vendor, including ours.

Table 2. Ten requirements, with the domains and framework references they derive from.

#	REQUIREMENT	DERIVED FROM
R1	**Discover every AI touchpoint** — web applications, desktop apps, coding agents and command-line tools, extensions, local runtimes, MCP servers, and embedded and operating-system AI.	§3.2, §3.3; NIST IR 8596 inventory; ISO/IEC 42001 A.9; ETSI Principle 5
R2	**Identify the account and edition**, not only the application, and enforce tier policy.	§3.2, §3.7; ISO/IEC 42001 A.9
R3	**Classify content on the device, on every route** — typed, pasted, uploaded, dragged and image — and in both directions.	§3.1; OWASP LLM02; ISO/IEC 27001 A.8.12
R4	**Keep user content on the device**, and make monitoring proportionate and transparent by design.	§3.8; UK ICO monitoring guidance; EU AI Act Art. 5
R5	**Govern connections and actions** — MCP servers, OAuth grants, API keys and agent autonomy — at the action layer.	§3.3, §3.4, §3.7; OWASP ASI02–ASI05; NCSC
R6	**Treat the AI tool as a potential attack vector** — extensions, installers, supply chain and runtimes.	§3.5; OWASP LLM03, ASI04; MITRE ATLAS AML.T0011, T0115
R7	**Attach AI provenance to consequential actions** — installs, commits, credential entry.	§3.6; OWASP LLM09
R8	**Coach before blocking, and provide a route to yes** — justification, approval, exceptions, alternatives.	§3.2, §3.9; ISO/IEC 42001 A.5, A.9
R9	**Produce the evidence auditors and regulators request**, and reuse what the organisation already owns.	§3.8, §3.9; ISO/IEC 27001, 42001; EU AI Act Art. 4
R10	**State the boundaries and prove coverage on the customer's own estate.**	§5; ETSI Principle 12

7 Lamplighter: endpoint-native AI governance

Lamplighter gives organisations control over how their people use AI without taking custody of what they type. A lightweight agent and a policy-installed browser extension on each workstation discover the AI tools in use, identify the account behind them, classify what people type, paste and upload on the device itself, and govern what AI agents are connected to and allowed to do — reporting evidence that, by default, carries none of that content to a central dashboard.

This section describes the architecture, maps each requirement from Section 6 to Lamplighter capabilities, and sets out why the design is suited to the landscape described in this paper. Every capability is marked with its status in Lamplighter v0.18 (October 2026), drawn from an internal audit of the product's code and automated tests: AVAILABLE in the product today; IN DEVELOPMENT, being built or verified on devices; ROADMAP, scheduled in the engineering plan; or PROPOSED, under consideration and not yet scheduled. Roadmap and proposed items are subject to change.

7.1 Architecture

Lamplighter has three parts on the workstation, an optional fourth, and two in the cloud (Figure 3). The **browser extension** is force-installed by enterprise policy in Chrome, Edge, Brave, Vivaldi and Opera, which lets it keep the blocking request API that ordinary Manifest V3 extensions lose; it holds a paste, submit or upload until the device has decided, and reads which account is signed in to eight of the most-used AI web applications. The **user helper** runs in the user's session: it records which application copied content came from, shows coaching and blocking notices, holds the clipboard while a desktop AI application is in front so that a paste into it is decided before it lands, and recognises a prompt sent from the ChatGPT desktop app, reporting it only once it has been sent. The **agent service** runs the collectors (processes, network flows joined to DNS, agent and MCP configuration, extensions, API keys on disk, sensitivity labels), the detection pipeline, the policy engine, and two local gateways: one for coding agents' model calls, which each tool must be pointed at, and one for local MCP servers' tool traffic, both off until the organisation switches them on. The optional fourth part is a third-party, vendor-signed **file-system minifilter** that needs its own licence and is off until it is enabled, which holds a file an AI application opens until the service has decided; Lamplighter ships no kernel code of its own. In the cloud, a multi-tenant **backend** hosted in Microsoft Azure North Europe, with a separate database for each customer, serves signed policy and catalogue bundles and receives events, which administrators work with in the **dashboard**.

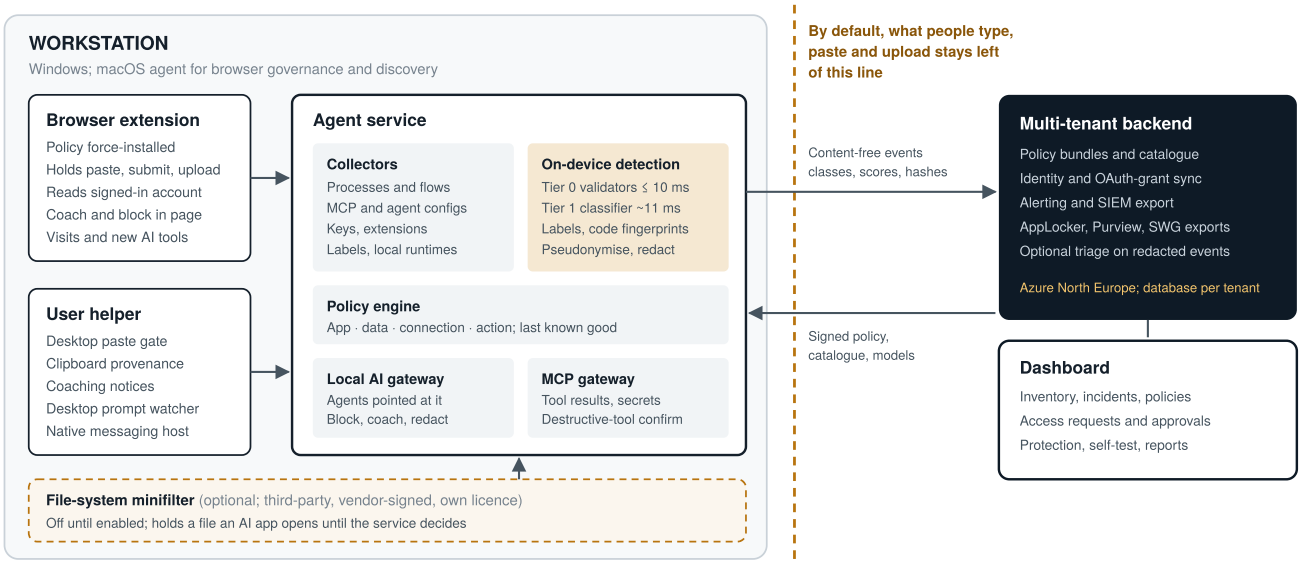


Figure 3. Lamplighter architecture. Classification happens on the device. In the default evidence mode, an event carries who, which device and application, the decision and the person's justification, with class labels, scores, hashes, byte counts and a content-free description of the data's shape; nothing the person typed, pasted or uploaded leaves the device. A window of at most 300 characters, pseudonymised or redacted, is an option each organisation can choose.

Detection is deliberately tiered, because a paste or prompt decision must fit a budget of 100 milliseconds on an ordinary business laptop; a decision that cannot be made in time lets the action proceed and records it as degraded, unless a policy has been set to fail closed. Sensitivity labels applied with Microsoft Purview are read before any content, as a model-free signal. **Tier 0** runs deterministic validators within 10 ms on a 10 KB input: national identification numbers for South Africa, the United Kingdom and nine EU member states, payment cards with Luhn checks, IBANs, phone numbers and e-mail addresses, private keys and tokens, 17 secret patterns across 13 providers, bulk-record and table shapes, and the organisation's own keyword lists and internal hosts. **Tier 1** is a small on-device transformer classifier (an int8 ONNX model of 7.6 MB) that labels content as personal, financial, credentials, source code, customer data, HR, legal contract, health, internal communications or security configuration, measured at 11 ms at the 95th percentile for a 2 KB paste. Health, HR and legal text that carries no identifier is detected by this classifier alone, so its accuracy matters: it is trained on varied generated documents in English and six other European languages, its thresholds are tuned and its accuracy measured on hand-written text that resembles real work, and raising that accuracy is the current engineering priority. A third tier, a local language model that re-scores uncertain results in the background, is built but deliberately not shipped: on an average laptop its memory and processor cost is not yet justified by a measured gain. No model is downloaded from a public hub at runtime; every model, catalogue and policy bundle is signed and verified before use.

7.2 How Lamplighter meets the requirements

Table 3. Requirements from Section 6 mapped to Lamplighter capabilities, as of Lamplighter v0.18 (October 2026).

#	LAMPLIGHTER CAPABILITY	STATUS
R1 · Discover every AI touchpoint		
	AI web applications matched against a signed catalogue, with long-tail discovery of unknown AI tools queued for review	AVAILABLE

#	LAMPLIGHTER CAPABILITY	STATUS
	Desktop AI apps and AI command-line tools by process, signer and command line; every TCP connection to an AI service attributed to the process that made it (QUIC traffic is seen through DNS only)	AVAILABLE
	AI browser and IDE extensions, flagged by permission risk	AVAILABLE
	Local LLM runtimes, the models present, and whether they are exposed to the network	AVAILABLE
	MCP servers configured in ten clients, including Claude Desktop, Claude Code, Cursor, VS Code, Windsurf, Codex and Gemini CLI, and whether they point at the internal domains and address ranges the organisation lists	AVAILABLE
	AI embedded in sanctioned apps (Microsoft 365 Copilot, Gemini in Workspace, Notion AI, Slack AI and others), operating-system AI, AI browsers, and meeting bots that join calls	ROADMAP
R2	**· Identify the account and edition**	
	Organisational or personal account in the browser on ChatGPT, Claude, Gemini, Microsoft Copilot, GitHub Copilot, Perplexity, DeepSeek and Grok, read from the vendor's own workspace signal, labelled Corporate or Shadow and stored as salted hashes; policy by account type. Other apps, and all desktop and CLI use, read as Shadow	AVAILABLE
	Corporate confirmed only when the account matches an organisation or domain the customer has registered	ROADMAP
	Vendor posture per application (training on inputs, retention, residency, DPA)	ROADMAP
	Detection of sessions not established through single sign-on; tenant-restriction enforcement	PROPOSED
R3	**· Classify on the device, on every route**	
	On-device detection with deterministic validators (including UK and EU national identifiers) and an on-device classifier; reuse of Purview/AIP sensitivity labels; fingerprinting of pasted code against the organisation's own repositories	AVAILABLE
	Browser: typed prompts, pastes, pasted images, file uploads and drag-and-drop coached or blocked before they are sent, with the body text of PDF, Office, OpenDocument, RTF and e-mail files read (files attached through a site's own file picker are judged on name and type only)	AVAILABLE
	Coding agents and AI CLIs (Claude Code, Codex, aider, Cline and others) through a local AI gateway the organisation switches on and points each tool at: block, coach, or redact and forward, including attached documents (a tool not pointed at it is reported, not inspected)	AVAILABLE
	Desktop AI apps (Windows): clipboard paste gate for text and copied files; prompts typed into the ChatGPT desktop app recognised once sent and coached after the fact	AVAILABLE
	Desktop AI apps (Windows): sensitive files an AI app opens through its file picker refused, where the optional third-party file-system driver is licensed and enabled (it is off by default, and its licence is installed on each device by hand for now); a file dragged in is not blocked	AVAILABLE
	Text in screenshots and images read by on-device OCR (Windows: browser uploads and gateway images so far)	IN DEVELOPMENT
	AI responses returning data the user should not see (MCP tool results are covered under R5)	PROPOSED
R4	**· Keep user content on the device**	
	By default none of what people type, paste or upload leaves the device; no keystroke logging — content is classified at submit, paste, drop or file open, and the clipboard only when pasting into an AI app	AVAILABLE
	Mandatory worker monitoring notice before any device can enrol; salted account identifiers and file-name hashes; configurable retention with real deletion	AVAILABLE
	Optional readable evidence: a window of at most 300 characters in which every detected value is replaced by a stand-in (the surrounding words are the user's own)	AVAILABLE
	Sealed local evidence of what was attempted at a coach or block decision (file opens excepted), openable only with a private key the customer holds; full capture only behind a recorded legal sign-off	AVAILABLE

#	LAMPLIGHTER CAPABILITY	STATUS
	R5 · Govern connections and actions	
	Policy over MCP servers, OAuth grants, API keys, agent modes, extensions and local runtimes (connection and action policy axes): a rule records, raises an incident and alerts, and each finding is tracked to a review outcome	AVAILABLE
	Detection of coding-agent CLIs (Claude Code, Codex, Gemini CLI, aider, Goose) run with permission checks disabled; API keys on disk reported by provider and prefix, never the key; agent connections to internal hosts; OAuth grants to AI apps in Microsoft 365 (with Entra ID connected)	AVAILABLE
	MCP gateway for local servers in users' JSON client configurations (not Codex's TOML configuration or project-level files), switched on per organisation: tool results redacted or withheld, secrets stripped from arguments, confirmation before destructive tools run	AVAILABLE
	MCP server allowlisting and blocking of servers that target internal systems, enforced at the gateway (reported as incidents today); remote MCP servers; agent guardrails on files and credential stores	ROADMAP
	Interception at the OAuth consent screen; detection of MCP tool descriptions that change after approval	PROPOSED
	R6 · Treat the AI tool as an attack vector	
	Extension, runtime and exposure inventory; detection of Claude Code and Gemini CLI launched in bypass mode, two of the three AI CLIs the s1ngularity attack abused	AVAILABLE
	Agent supply-chain scanning of plugins, hooks, rules and instruction files	ROADMAP
	CVE matching for MCP servers and runtimes; malicious-extension feed; guards for AI-sourced terminal commands (ClickFix) and crafted prompt links	PROPOSED
	R7 · Attach AI provenance to actions	
	Clipboard provenance — the source application of every paste into AI (Windows)	AVAILABLE
	Package-install guard against hallucinated packages; secrets in AI-assisted commits; warnings on sign-in pages reached from AI answers; AI-output provenance	PROPOSED
	R8 · Coach before blocking; a route to yes	
	Log, coach and block modes with justification capture; block after a coaching period, or at once with an audited override reason; 30-day policy simulation; one-click modes and a tenant-wide pause	AVAILABLE
	Request access from the notice shown for a tool that is not approved; an approval changes enforcement on every device automatically	AVAILABLE
	Approval for specific data types; telling the requester the outcome; blocking a tool's page with the approved alternative named	ROADMAP
	Coaching acknowledgements kept as AI-literacy records	PROPOSED
	R9 · Evidence, and reuse of what you own	
	Board and ISO-style posture reports; incident workflow; export to Microsoft Sentinel, Splunk and webhooks; exports to AppLocker, Purview Endpoint DLP restricted apps and SWG/firewall lists	AVAILABLE
	ISO/IEC 42001 workbook with AI-system and risk registers, evidence store and audit pack	ROADMAP
	R10 · State the boundaries; prove coverage	
	Protection attestation: every device reports which protections are active, and removal is recorded as a named tamper event on the Protection page and in the SIEM feed	AVAILABLE
	Device self-test, run from the dashboard, that exercises each local protection with harmless inputs and names the fix for any that fails	AVAILABLE
	Scenario suite proving each route end to end (browser, AI-gateway and MCP routes recorded; desktop routes pending)	IN DEVELOPMENT

#	LAMPLIGHTER CAPABILITY	STATUS
	A periodic protection-coverage digest sent to alert channels	ROADMAP
	Tamper events pushed to e-mail, Slack and Teams; per-framework coverage statements computed from each customer's fleet	PROPOSED

7.3 Why Lamplighter

Most products in this market began as network or browser controls and extended toward the endpoint. Lamplighter began at the endpoint. Seven properties follow from that choice and from the evidence in this paper.

1 Built for the gap the market has not closed

The fastest-growing AI surfaces — coding agents, local MCP servers, desktop assistants and local runtimes — are invisible to network sensors (§4.2). Lamplighter reads MCP configurations, agent modes, API keys and runtime exposure directly on the device today; routes coding agents' model calls and local MCP tool traffic through on-device gateways once they are switched on; and, on Windows, decides a paste into a desktop AI app before it lands. This is the category platform vendors paid to enter in 2026 [18], [19].

2 The AI data-protection tool that does not feed your data to an AI

Classification runs on the workstation, and in the default mode none of what people type, paste or upload leaves it. Where an organisation needs to see more, it can choose to receive a window of at most 300 characters in which every detected value has been replaced, or, behind a recorded legal sign-off, full capture, in which the complete text is sealed on the device and opens only with a key the organisation holds, and a redacted window of up to 300 characters also reaches the dashboard. That is a security property — in the default mode there is no central store of employee prompts to breach or subpoena — and a legal one: it is the least-intrusive design the ICO asks for, it limits what a DPIA or a works council must weigh to metadata rather than content, and it performs no emotion inference, which the EU AI Act prohibits in the workplace (§3.8).

3 Account-aware, not domain-aware

The legal consequence of a prompt depends on the account behind it (§3.2). Lamplighter reads which account is signed in to eight of the most-used AI web applications, labels it Corporate or Shadow, and enforces policy on that distinction, which is what decides whether data lands in a tier that trains on it, retains it for years, or is swept into litigation. Today Corporate reflects the vendor's own workspace signal; confirming it against the organisations and domains each customer registers is on the near-term roadmap.

4 Governs what AI is connected to and allowed to do

Lamplighter's policy model has four axes — which AI may run, what data may enter it, what it is connected to, and what agents may execute — so MCP servers, OAuth grants, API keys and bypass modes fall under the same simulated, staged policy as data, and each finding is tracked until it is fixed, approved or dismissed. For local MCP servers, the gateway also redacts tool results, strips secrets from arguments and asks before destructive tools run. Because prompt injection cannot be filtered away (§3.4), this is where containment happens.

5 **Coaches first, and gives people a route to yes**

Most AI misuse is negligent (§3.9). Lamplighter starts every policy in log or coach mode, captures people's justifications, and makes blocking available only after a coaching period unless an administrator records a reason. It tells a person that a tool is not approved at most once a day per AI company, and asks about data only when they try to enter it, which keeps false positives from driving use to phones. A person can request a tool that is not approved from the notice itself, and approving it changes enforcement on every device automatically — a first version of the request-and-approval loop that no vendor documentation reviewed for this paper evidenced (§4.2). Each coaching decision is recorded with the person's response: the raw material for the AI-literacy records the European Commission accepts as evidence [97].

6 **Works with what you already own**

Lamplighter reads the sensitivity labels an organisation has already applied, and exports its decisions to AppLocker, Microsoft Purview, secure web gateways and SIEM platforms. It complements rather than replaces an existing DLP investment, which lowers the cost of adoption for cloud-first organisations where conventional DLP is hard to run.

7 **Honest about its limits**

Lamplighter documents the four boundaries in Section 5 and its own platform limits (Section 7.4) rather than implying they do not exist. Its protection attestation records a removed extension or driver as a named tamper event and shows a device that stops reporting as silent within 15 minutes, and a self-test proves each protection on a device on demand. The statuses in this paper are drawn from an internal audit of the product's code and tests. A bounded claim that survives due diligence is worth more than an unbounded one that does not.

7.4 Deployment and platforms

Lamplighter deploys on Windows 11 22H2 or later and Windows 10 22H2 (x64) through Intune, Group Policy or any remote-management tool that can install an MSI, without a reboot. The browser extension is force-installed by policy; on fleets managed only through Entra ID and Intune, browsers accept a force-install only from their own stores, and that store listing is not yet published. A device enrols with a token whose lifetime and use count the organisation sets, receives a certificate from that organisation's own certificate authority, and cannot enrol until a monitoring notice has been published. Decisions are made against the last signed policy, so protection continues offline, and agent updates roll out in rings that halt automatically on a regression. Release packages are not yet code-signed: signing is built into the release pipeline and awaits its certificate.

Two platform limits matter in procurement. The macOS agent provides browser governance and discovery only: the desktop paste gate, file-open control, clipboard provenance and OCR are Windows features today. And pastes made inside a remote session (Remote Desktop, Azure Virtual Desktop, Windows 365, Citrix) bypass the desktop paste gate; Lamplighter reports the device as unprotected for that session rather than closing the gap. Firefox and Linux developer workstations are on the roadmap.

7.5 Roadmap

Lamplighter's engineering plan sets the next work in four groups. Dates are not committed.

1. **Completing the desktop and account picture:** Corporate accounts confirmed against registered organisations and domains; OCR extended to the desktop clipboard; the desktop routes recorded end to end in the scenario suite; a periodic protection-coverage digest sent to alert channels; and automated delivery of the file-driver licence to each device.
2. **Detection quality:** a pre-trained encoder for the tier-1 classifier, reading past the first window of long documents, and a feedback loop from analysts' false-positive decisions into the evaluation set.
3. **Enforcement that closes the remaining routes:** approval per tool and per data type, with the requester told the outcome; blocking the page of a tool an administrator has blocked, with a notice that names the approved alternative; and executable denial and agent guardrails through the file-system driver.
4. **Running it without help:** online purchase, onboarding in the browser, a downloads page, connection to Microsoft 365 in a single consent, and the browser-store listings.

Further out, the plan holds the ISO/IEC 42001 workbook; embedded and operating-system AI (Microsoft 365 Copilot, Gemini in Workspace, Windows Recall, AI browsers and meeting bots); vendor posture and more account adapters in the catalogue; agent supply-chain scanning; behavioural baselines; and macOS and Linux parity. The items marked Proposed in Table 3, among them response-side inspection, interception at the OAuth consent screen, guards for AI-sourced terminal commands and crafted prompt links, and controls for agentic browsers, are under consideration but not yet scheduled.

8 Conclusion

The evidence of 2025–2026 describes a shift in where AI risk lives. The risks that are best documented — pasting, uploading, personal accounts, application sprawl — are also the ones the market already covers inside network and browser platforms. The risks that are worst covered — local MCP servers, desktop and command-line agents, agent configuration as an execution layer, consent to AI connectors, agent credentials, data flowing back out of assistants, and the lack of any route to approval — are the ones invisible to every sensor except the endpoint.

Confidence, in this setting, cannot mean prevention of a threat class that its own vendors say cannot be fixed. It means demonstrable control of a behaviour class, containment of what cannot be prevented, and evidence of both that stands up to an auditor, a regulator and a works council. That is a documentary claim as much as a technical one, and the control that makes it must itself be lawful. Lamplighter was designed to make that claim, within the boundaries stated here.

To discuss deploying Lamplighter, arrange a 30-day trial for up to 25 devices, or request the full 79-risk register with framework mappings, visit **lamplighterai.com**.

Appendix A · The endpoint AI risk register

The 79 risks identified in the research, grouped into the nine domains of Section 3, with representative framework references and Lamplighter's coverage as of October 2026 (v0.18). "Partly available" means some of the risk is addressed today and the remainder is in development, on the roadmap or proposed; "Proposed" means under consideration and not yet scheduled. "Outside endpoint scope" marks the boundaries described in Section 5.

ID	RISK	FRAMEWORK REFERENCES	LAMPLIGHTER
1 · Sensitive data leaving the endpoint			
1.1	Prompt paste and typing in the browser	LLM02, ISO 42001 A.9, ISO 42001 A.7	AVAILABLE
1.2	File uploads to AI tools	LLM02, ISO 27001 A.8.12, SOC 2 CC6.7	AVAILABLE
1.3	Code, secrets and IP into IDE and CLI agents	LLM02, LLM03, ISO 27001 A.8.12	PARTLY AVAILABLE
1.4	Embedded SaaS AI out-leaking the chatbots	NIST AI 600-1 Value Chain and Component Integration, NIST IR 8596 embedded AI integrations, ISO 42001 A.10	ROADMAP
1.5	AI meeting note-takers and recorders	ISO 42001 A.9, Microsoft Insufficient Transparency and Accountability, GDPR Art. 6	ROADMAP
1.6	OS-level capture: Recall and Apple Intelligence	ISO 27001 A.8.1, ICO monitoring-workers proportionality	ROADMAP
1.7	Share links and discovery feeds publishing chats	LLM02, GDPR Art. 33	PROPOSED
1.8	Downstream leakage of over-shared data	ASI03, Microsoft Incorrect Permissions, Microsoft Insufficient Isolation	PARTLY AVAILABLE
1.9	Local model runtimes as a false safe harbour	NIST AI 600-1 Data Privacy, NIST AI 600-1 Information Security, CIS endpoint execution environments	AVAILABLE
1.10	Out-of-band bypass: phones, cameras and dictation		OUTSIDE ENDPOINT SCOPE
2 · Shadow AI, accounts and tiers			
2.1	Unknown and expanding AI app estate	NIST IR 8596 asset inventory, ETSI TS 104 223 Principle 5, ISO 42001 A.9	AVAILABLE
2.2	Business work on personal and free tiers	ISO 42001 A.9, Hamburg DPA checklist, CNIL guidance	PARTLY AVAILABLE
2.3	Corporate accounts that bypass SSO	SOC 2 CC6.1, SOC 2 CC6.3, ISO 27001 A.5.15	PROPOSED
2.4	Prohibited and jurisdictional AI tools	GDPR Chapter V, POPIA s.72, NIST AI 600-1 Value Chain	PARTLY AVAILABLE
2.5	Embedded AI switched on by default	NIST IR 8596, ISO 42001 A.10, CSA AICM supply chain	ROADMAP
2.6	Consumer tiers with different training and retention terms	GDPR Art. 5, GDPR Art. 28, GDPR Art. 44	PARTLY AVAILABLE
2.7	Long dwell time and block-driven displacement	Gartner AI TRiSM, ISO 42001 A.9	AVAILABLE
2.8	AI browsers as an unsanctioned client class	AML.T0114, AML.T0131, Microsoft CUA Visual Attack	ROADMAP
3 · Agents, MCP and connectors			
3.1	Agent config files that execute on write or repo open	ASI05, ASI04, T11	PARTLY AVAILABLE
3.2	Tool poisoning, line jumping and rug pulls	OWASP MCP03:2025, ASI04, AML.T0115	PARTLY AVAILABLE
3.3	Malicious and typosquatted agent components	ASI04, LLM03, AML.T0011.001	ROADMAP
3.4	MCP client and reference-server vulnerabilities	ASI04, ASI05, AML.T0010	PROPOSED
3.5	Static credentials living in agent configuration	ASI03, AML.T0055, CSA AICM IAM	AVAILABLE
3.6	Exposed local agent gateways and MCP servers	ASI10, AML.T0132, AML.T0006.000-.003	PARTLY AVAILABLE
3.7	Agents inheriting the full permission set of the user	ASI03, LLM06, Microsoft Incorrect Permissions	PARTLY AVAILABLE
3.8	Permission-bypass flags and human-in-the-loop bypass	ASI09, T10, Microsoft Human-in-the-Loop Bypass	PARTLY AVAILABLE
3.9	Destructive autonomous action	ASI02, ASI10, Microsoft Agent Flow Manipulation	PARTLY AVAILABLE

ID	RISK	FRAMEWORK REFERENCES	LAMPLIGHTER
3.10	Agent-initiated transactions in the browser	ASI02, AML.T0048	ROADMAP
3.11	Toxic agent flows across connectors	ASI02, ASI08, T12	PARTLY AVAILABLE
3.12	Local MCP and desktop agents invisible to network sensors	NIST IR 8596, CIS endpoint execution environments	AVAILABLE
4 · Prompt injection and agent hijacking			
4.1	Zero-click exfiltration from sanctioned assistants	LLM01, ASI01, AML.T0051.001	OUTSIDE ENDPOINT SCOPE
4.2	Crafted assistant links and pre-filled prompt URLs	ASI01, ASI06, AML.T0131	PROPOSED
4.3	Persistent memory and context poisoning	ASI06, T1, Microsoft Memory Poisoning	PROPOSED
4.4	Multimodal, screenshot and cloaked injection	AML.T0129, AML.T0134, Microsoft CUA Visual Attack	ROADMAP
4.5	AI browser agents acting in authenticated sessions	ASI01, ASI02, AML.T0114	ROADMAP
4.6	The residual-risk position itself	ETSI TS 104 223 Principle 12, ISO 42001 A.6, NIST AI 600-1 Information Security	PARTLY AVAILABLE
5 · The AI tool as an attack vector			
5.1	Fake AI installers and malvertising	AML.T0052, AML.T0011, ENISA ETL 2025	PARTLY AVAILABLE
5.2	AI-themed ClickFix on legitimate AI domains	AML.T0011, ENISA ETL 2025	PROPOSED
5.3	Malicious and over-privileged AI browser extensions	LLM03, ASI04, AML.T0113	PARTLY AVAILABLE
5.4	Compromised official AI tooling and IDE integrity	ASI04, LLM03, AML.T0010	ROADMAP
5.5	Worms that weaponise the endpoint's own AI CLIs	ASI04, ASI05, AML.T0011.001	PARTLY AVAILABLE
5.6	Malicious models and chat-template backdoors	LLM04, LLM03, AML.T0115	PARTLY AVAILABLE
5.7	Local inference runtime CVEs, exposure and LLMjacking	AML.T0034, AML.T0029, AML.T0132	PARTLY AVAILABLE
5.8	Rules files and context files as injection carriers	LLM01, LLM03, ENISA Rules File Backdoor	ROADMAP
6 · AI output acted on at the endpoint			
6.1	Slopsquatting and hallucinated packages	LLM09, LLM03, ASI04	PROPOSED
6.2	Insecure AI-generated code	LLM05, LLM09, T11	NOT COVERED
6.3	Secrets in AI-assisted commits	LLM02, ISO 27001 A.8.24	PROPOSED
6.4	Hallucinated URLs and poisoned AI answers	LLM09, AML.T0062, AML.T0052	PROPOSED
6.5	AI-generated phishing, deepfakes and vishing	AML.T0052, Microsoft User Impersonation, ENISA ETL 2025	OUTSIDE ENDPOINT SCOPE
6.6	Over-reliance, automation bias and unverified output	LLM09, NIST AI 600-1 Confabulation, NIST AI 600-1 Human-AI Configuration	PARTLY AVAILABLE
6.7	IP status of outputs and AI-content disclosure	NIST AI 600-1 Intellectual Property, EU AI Act Art. 50, ICO marking rule	PROPOSED
7 · Identities, sessions and OAuth			
7.1	Replayable AI session tokens that bypass MFA	AML.T0113, AML.T0091.001, AML.T0012	PROPOSED
7.2	Extension-borne session-token theft with connector reach	LLM03, AML.T0113	PARTLY AVAILABLE
7.3	AI API key sprawl on endpoints and in repositories	AML.T0055, ISO 27001 A.8.24, ISO 27001 A.5.17	AVAILABLE
7.4	OAuth consent sprawl into AI apps	ASI03, LLM03, AML.T0012	PARTLY AVAILABLE
7.5	Agent credentials and non-human identity	ASI03, AML.T0055	PARTLY AVAILABLE
7.6	AI account takeover has two extra impacts	AML.T0034, LLM10, NIST AI 600-1 Data Privacy	ROADMAP
7.7	Identity hygiene gaps that AI amplifies	Microsoft Incorrect Permissions, SOC 2 CC6.1, CIS Controls v8.1 identity safeguards	OUTSIDE ENDPOINT SCOPE
8 · Regulatory, legal and monitoring lawfulness			
8.1	Personal data in a consumer chatbot is a reportable breach	GDPR Art. 5, GDPR Art. 6, GDPR Art. 33	AVAILABLE
8.2	EU AI Act literacy, transparency and deployer duties	EU AI Act Art. 4, EU AI Act Art. 5, EU AI Act Art. 26	ROADMAP

ID	RISK	FRAMEWORK REFERENCES	LAMPLIGHTER
8.3	HIPAA, PHI and healthcare-specific duties	HIPAA Security Rule NPRM, 45 CFR 92.210, ISO 42001 A.9	PARTLY AVAILABLE
8.4	US state ADMT and AI laws in flux	Colorado SB 26-189, CCPA ADMT regulations, Texas TRAIGA	NOT COVERED
8.5	Financial-sector supervision of GenAI use	FINRA 2026 Regulatory Oversight Report, SEC FY2026 examination priorities	PARTLY AVAILABLE
8.6	Prompts are discoverable records with no AI privilege	ISO 27001 A.5.34, ISO 27001 A.8.15, SOC 2 CC7.2	AVAILABLE
8.7	Confidentiality, privilege, NDA and trade-secret exposure	ABA Formal Opinion 512, ISO 42001 A.9, trade-secret law	AVAILABLE
8.8	QMS, GxP and regulated-documentation integrity	ISPE GAMP AI Guide, MDCG 2025-6, Team-NB position paper v2	PROPOSED
8.9	POPIA and jurisdictions without AI statutes	POPIA s.22, POPIA s.71, POPIA s.72	PARTLY AVAILABLE
8.10	Insurance is actively excluding AI losses	ISO CGL generative-AI exclusion (Jan 2026)	AVAILABLE
8.11	Audit evidence auditors will actually request	ISO 27001 A.5.1, ISO 27001 A.5.10, ISO 27001 A.5.19	PARTLY AVAILABLE
8.12	Lawfulness of the monitoring tool itself	ICO monitoring-workers guidance, GDPR Art. 35, GDPR Art. 88	AVAILABLE
9 · Governance and process			
9.1	No inventory, and no single sensor that builds one	NIST IR 8596, ETSI TS 104 223 Principle 5, ISO 42001 A.9	AVAILABLE
9.2	Policy on paper with no runtime enforcement	ISO 42001 A.2, ISO 42001 A.9, ISO 27001 A.5.1	AVAILABLE
9.3	No approval or exception workflow	ISO 42001 A.5, ISO 42001 A.9, ICO internal AI use policy	PARTLY AVAILABLE
9.4	Audit evidence that will not survive scrutiny	T8, Microsoft Insufficient Transparency and Accountability, EU AI Act Art. 26	PARTLY AVAILABLE
9.5	Training and AI-literacy records	EU AI Act Art. 4, ETSI TS 104 223 Principle 1, ISO 27001 A.6.3	PARTLY AVAILABLE
9.6	Incident response and the inability to stop	NIST AI 600-1 incident response, ISO 42001 A.6, ETSI TS 104 223 Principle 12	PARTLY AVAILABLE
9.7	Vendor due diligence and procurement	NIST AI 600-1, NIST IR 8596, ISO 42001 A.10	ROADMAP
9.8	Offboarding and data stranded in personal accounts	ISO 27001 A.5.11, GDPR Art. 5(1)(e)	PARTLY AVAILABLE
9.9	Insider risk reframed around negligence	Gartner AI TRiSM (80% internal), DTEX/Ponemon Cost of Insider Risks	AVAILABLE

Appendix B · OWASP coverage map

How Lamplighter addresses each item of the OWASP Top 10 for LLM Applications 2025 [22] and the OWASP Top 10 for Agentic Applications for 2026 [23] at the endpoint. Items are listed once per status where coverage is split.

ID	ITEM	LAMPLIGHTER CAPABILITY	STATUS
LLM01	Prompt injection	Containment through connection and action policy; confirmation before destructive MCP tools run	AVAILABLE
LLM01	Prompt injection (continued)	Enforced MCP server allowlisting	ROADMAP
LLM01	Prompt injection (continued)	Pre-filled prompt link guard	PROPOSED
LLM02	Sensitive information disclosure	On-device classification with coach and block in the browser, in desktop AI apps on Windows, for coding agents and in MCP tool results	AVAILABLE
LLM03	Supply chain	Extension, IDE-extension and runtime inventory	AVAILABLE
LLM03	Supply chain (continued)	Agent supply-chain scanning of plugins, hooks, rules and instruction files	ROADMAP
LLM03	Supply chain (continued)	CVE matching; malicious-extension feed	PROPOSED
LLM04	Data and model poisoning	Local runtime and model inventory	AVAILABLE
LLM05	Improper output handling	AI-output provenance on paste into other applications	PROPOSED
LLM06	Excessive agency	Bypass-mode detection; connection and action policy over agents	AVAILABLE
LLM07	System prompt leakage	Developer-layer risk, outside endpoint governance	OUTSIDE ENDPOINT SCOPE
LLM08	Vector and embedding weaknesses	Downstream (response-side) inspection	PROPOSED
LLM09	Misinformation	Package-install guard; warnings on AI-suggested sign-in pages	PROPOSED
LLM10	Unbounded consumption	API keys on disk; exposed local runtimes	AVAILABLE
ASI01	Agent goal hijack	Scanning of rules and instruction files for injection patterns	ROADMAP
ASI01	Agent goal hijack (continued)	Crafted-link guard; agent memory integrity	PROPOSED
ASI02	Tool misuse	MCP gateway confirmation for destructive tools; secrets stripped from tool arguments	AVAILABLE
ASI03	Identity and privilege abuse	OAuth grant discovery (Entra ID); API keys on disk; policy over grants	AVAILABLE
ASI04	Agentic supply chain	MCP server discovery across ten clients	AVAILABLE
ASI04	Agentic supply chain (continued)	Agent supply-chain scanning of plugins, hooks, rules and instruction files	ROADMAP
ASI05	Unexpected code execution	Detection of agents run with permission checks disabled	AVAILABLE
ASI06	Memory and context poisoning	Scanning of rules and instruction files for injection patterns	ROADMAP
ASI06	Memory and context poisoning (continued)	Agent memory integrity	PROPOSED
ASI07	Insecure inter-agent communication	Largely a platform-layer risk; limited endpoint signal	OUTSIDE ENDPOINT SCOPE
ASI08	Cascading failures	MCP gateway inspection, redaction or withholding of tool results	AVAILABLE
ASI09	Human-agent trust exploitation	Out-of-band confirmation for agent-driven browser actions	PROPOSED
ASI10	Rogue agents	Exposed local agent and runtime detection	AVAILABLE
ASI10	Rogue agents (continued)	Agent guardrails on files and credential stores	ROADMAP

References

IEEE style. All online sources were accessed on 19 September 2026.

[1] Netskope Threat Labs, “Netskope AI report 2026,” *Netskope*, 2026. [Online]. Available: https://www.netskope.com/resources/threat-labs-reports/netskope-ai-report-2026

[2] LayerX Security, “Enterprise AI and SaaS data security report 2025 (as reported by The Register),” *The Register*, Oct. 7, 2025. [Online]. Available: https://www.theregister.com/2025/10/07/gen_ai_shadow_it_secrets

[3] Microsoft Threat Intelligence, “Malicious AI assistant extensions harvest LLM chat histories,” *Microsoft Security Blog*, Mar. 5, 2026. [Online]. Available: https://www.microsoft.com/en-us/security/blog/2026/03/05/malicious-ai-assistant-extensions-harvest-llm-chat-histories/

[4] Microsoft AI Red Team, “Updating the taxonomy of failure modes in agentic AI systems: What a year of red teaming taught us,” *Microsoft Security Blog*, Jun. 4, 2026. [Online]. Available: https://www.microsoft.com/en-us/security/blog/2026/06/04/updating-taxonomy-failure-modes-agentic-ai-systems-year-red-teaming-taught-us/

[5] Check Point Research, “RCE and API token exfiltration through Claude Code project files (CVE-2025-59536),” *Check Point Research*, 2026. [Online]. Available: https://research.checkpoint.com/2026/rce-and-api-token-exfiltration-through-claude-code-project-files-cve-2025-59536/

[6] IBM and Ponemon Institute, “Cost of a data breach report 2025: The AI oversight gap,” *IBM*, Jul. 30, 2025. [Online]. Available: https://www.ibm.com/think/x-force/2025-cost-of-a-data-breach-navigating-ai

[7] Cyberhaven, “Endpoint AI agents: the blind spot,” *Cyberhaven blog*, May 11, 2026. [Online]. Available: https://www.cyberhaven.com/blog/endpoint-ai-agents-blind-spot

[8] Nightfall AI, “Browser DLP solutions,” *Nightfall blog*, 2026. [Online]. Available: https://www.nightfall.ai/blog/browser-dlp-solutions

[9] Engadget, “OpenAI no longer has to preserve all of its ChatGPT data, with some exceptions,” *Engadget*, Oct. 2025. [Online]. Available: https://www.engadget.com/ai/openai-no-longer-has-to-preserve-all-of-its-chatgpt-data-with-some-exceptions-192422093.html

[10] Pillar Security, “AI coding tools under fire: Mapping the malvertising campaigns targeting the vibe coding ecosystem,” *Pillar Security blog*, 2026. [Online]. Available: https://www.pillar.security/blog/ai-coding-tools-under-fire-mapping-the-malvertising-campaigns-targeting-the-vibe-coding-ecosystem

[11] The Hacker News, “Malicious Nx packages in 's1ngularity' attack weaponized AI CLIs,” *The Hacker News*, Aug. 2025. [Online]. Available: https://thehackernews.com/2025/08/malicious-nx-packages-in-s1ngularity.html

[12] OpenAI, “Hardening Atlas against prompt injection,” *OpenAI*, Dec. 22, 2025. [Online]. Available: https://openai.com/index/hardening-atlas-against-prompt-injection/

[13] D. Chismon, “Prompt injection is not SQL injection (it may be worse),” *UK National Cyber Security Centre*, Dec. 8, 2025. [Online]. Available: https://www.ncsc.gov.uk/blog-post/prompt-injection-is-not-sql-injection

[14] Gartner, “Market guide for AI trust, risk and security management,” *Gartner*, Feb. 18, 2025. [Online]. Available: https://www.gartner.com/en/documents/6185655

[15] White & Case, “EU AI Omnibus enters into force, amending the AI Act,” *White & Case insight alert*, Jul. 2026. [Online]. Available: https://www.whitecase.com/insight-alert/eu-ai-omnibus-enters-force-amending-ai-act

[16] Canadian Cyber, “Shadow AI and the ISO 27001 internal audit,” *Canadian Cyber*, 2025. [Online]. Available: https://canadiancyber.ca/shadow-ai-iso-27001-internal-audit/

[17] Information Commissioner's Office, “Employment practices and data protection: Monitoring workers,” *ICO*, Oct. 3, 2023. [Online]. Available: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/employment/monitoring-workers/data-protection-and-monitoring-workers/

[18] Palo Alto Networks, “Palo Alto Networks completes acquisition of Koi to secure the agentic endpoint,” *Palo Alto Networks press release*, Apr. 14, 2026. [Online]. Available: https://www.paloaltonetworks.com/company/press/2026/palo-alto-networks-completes-acquisition-of-koi-to-secure-the-agentic-endpoint

[19] Cisco, “Cisco announces intent to acquire Astrix Security,” *Cisco blog*, May 4, 2026. [Online]. Available: https://blogs.cisco.com/news/cisco-announces-intent-to-acquire-astrix-security

[20] Zscaler ThreatLabz, “Zscaler 2026 AI threat report: 83% year-over-year surge in AI activity,” *Zscaler press release*, 2026. [Online]. Available: https://www.zscaler.com/press/zscaler-2026-ai-threat-report-83-year-over-year-surge-ai-activity-creates-growing-oversight

[21] Netskope Threat Labs, “Cloud and threat report: Generative AI 2025,” *Netskope*, 2025. [Online]. Available: https://www.netskope.com/resources/cloud-and-threat-reports/cloud-and-threat-report-generative-ai-2025

[22] OWASP GenAI Security Project, “OWASP Top 10 for LLM applications 2025,” *OWASP*, 2025. [Online]. Available: https://genai.owasp.org/llm-top-10/

[23] OWASP GenAI Security Project, Agentic Security Initiative, “OWASP Top 10 for agentic applications for 2026,” *OWASP*, Dec. 9, 2025. [Online]. Available: https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/

[24] MITRE, “ATLAS: Adversarial threat landscape for artificial-intelligence systems (releases 2026.05-2026.09),” *MITRE*, 2026. [Online]. Available: https://raw.githubusercontent.com/mitre-atlas/atlas-data/main/CHANGELOG.md

[25] National Institute of Standards and Technology, “Artificial intelligence risk management framework: Generative artificial intelligence profile,” *NIST AI 600-1*, Jul. 2024. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

[26] National Institute of Standards and Technology, “Cybersecurity framework profile for artificial intelligence (Cyber AI Profile), initial preliminary draft,” *NIST IR 8596 ipd*, Dec. 16, 2025. [Online]. Available: https://csrc.nist.gov/pubs/ir/8596/iprd

[27] NIST Center for AI Standards and Innovation, “Request for information regarding security considerations for artificial intelligence agents,” *Federal Register*, Jan. 8, 2026. [Online]. Available: https://www.federalregister.gov/documents/2026/01/08/2026-00206/request-for-information-regarding-security-considerations-for-artificial-intelligence-agents

[28] ISO/IEC, “ISO/IEC 42001:2023 Information technology - Artificial intelligence - Management system,” *International Organization for Standardization*, 2023. [Online]. Available: https://www.iso.org/standard/42001

[29] Cloud Security Alliance, “AI controls matrix (AICM) v1,” *CSA*, Jul. 9, 2025. [Online]. Available: https://cloudsecurityalliance.org/artifacts/ai-controls-matrix

[30] ETSI, “Securing artificial intelligence (SAI); Baseline cyber security requirements for AI models and systems,” *ETSI TS 104 223 V1.1.1*, Apr. 2025. [Online]. Available: https://www.etsi.org/deliver/etsi_ts/104200_104299/104223/01.01.01_60/ts_104223v010101p.pdf

[31] European Union Agency for Cybersecurity (ENISA), “ENISA threat landscape 2025,” *ENISA, v1.2*, Oct. 2025 (rev. Jan. 2026). [Online]. Available: https://www.enisa.europa.eu/sites/default/files/2026-01/ENISA%20Threat%20Landscape%202025_v1.2.pdf

[32] Harmonic Security, “What 22 million enterprise AI prompts reveal about shadow AI in 2025,” *Harmonic Security*, 2025. [Online]. Available: https://www.harmonic.security/resources/what-22-million-enterprise-ai-prompts-reveal-about-shadow-ai-in-2025

[33] Cyberhaven Labs, “Cyberhaven 2026 AI adoption and risk report,” *Cyberhaven press release*, 2026. [Online]. Available: https://www.cyberhaven.com/press-releases/cyberhaven-2026-ai-adoption-risk-report

[34] Cloud Security Alliance, “AI browser extension attack surface (CSA research note),” *CSA Labs*, Apr. 2026. [Online]. Available: https://labs.cloudsecurityalliance.org/research/csa-research-note-ai-browser-extension-attack-surface-202604/

[35] Zscaler ThreatLabz, “AI is now the default enterprise accelerator: Takeaways from the ThreatLabz 2026 AI security report,” *Zscaler blog*, 2026. [Online]. Available: https://www.zscaler.com/blogs/security-research/ai-now-default-enterprise-accelerator-takeaways-threatlabz-2026-ai-security

[36] LayerX Security, “AI is now the #1 data exfiltration vector in the enterprise, and nobody's watching,” *LayerX blog*, 2025. [Online]. Available: https://layerxsecurity.com/blog/ai-is-now-the-1-data-exfiltration-vector-in-the-enterprise-and-nobodys-watching/

[37] Menlo Security, “Menlo Security's 2025 report uncovers 68% surge in shadow generative AI usage in the modern enterprise,” *Menlo Security press release*, 2025. [Online]. Available: https://www.menlosecurity.com/press-releases/menlo-securitys-2025-report-uncovers-68-surge-in-shadow-generative-ai-usage-in-the-modern-enterprise

[38] Netskope Threat Labs, “Cloud and threat report: 2026,” *Netskope*, 2026. [Online]. Available: https://www.netskope.com/resources/cloud-and-threat-reports/cloud-and-threat-report-2026

[39] GitGuardian, “The state of secrets sprawl 2026,” *GitGuardian*, 2026. [Online]. Available: https://blog.gitguardian.com/the-state-of-secrets-sprawl-2026/

[40] Fast Company, “Google is indexing ChatGPT conversations,” *Fast Company*, Jul. 2025. [Online]. Available: https://www.fastcompany.com/91376687/google-indexing-chatgpt-conversations

[41] I. Martin, “Elon Musk's xAI published hundreds of thousands of Grok chatbot conversations,” *Forbes*, Aug. 20, 2025. [Online]. Available: https://www.forbes.com/sites/iainmartin/2025/08/20/elon-musks-xai-published-hundreds-of-thousands-of-grok-chatbot-conversations/

[42] Varonis, “2025 state of data security report,” *Varonis*, 2025. [Online]. Available: https://www.varonis.com/blog/state-of-data-security-report

[43] The Hacker News, “Ollama out-of-bounds read vulnerability (CVE-2026-7482),” *The Hacker News*, May 2026. [Online]. Available: https://thehackernews.com/2026/05/ollama-out-of-bounds-read-vulnerability.html

[44] Harmonic Security, “AI usage index report 2026,” *Harmonic Security*, 2026. [Online]. Available: https://www.harmonic.security/resources/ai-usage-index-report-2026

[45] Terms.Law, “OpenAI v. New York Times stopped being just a copyright case the moment the court turned to your ChatGPT logs,” *Terms.Law*, Nov. 12, 2025. [Online]. Available: https://www.terms.law/2025/11/12/openai-v-new-york-times-stopped-being-just-a-copyright-case-the-moment-the-court-turned-to-your-chatgpt-logs/

[46] OpenAI, “How your data is used to improve model performance,” *OpenAI Help Center*, 2025. [Online]. Available: https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance

[47] Anthropic, “Updates to our consumer terms,” *Anthropic*, Aug. 28, 2025. [Online]. Available: https://www.anthropic.com/news/updates-to-our-consumer-terms

[48] Google, “Gemini Apps privacy hub,” *Google Support*, 2025. [Online]. Available: https://support.google.com/gemini/answer/13594961?hl=en

[49] Al Jazeera, “Which countries have banned DeepSeek and why?,” *Al Jazeera*, Feb. 6, 2025. [Online]. Available: https://www.aljazeera.com/news/2025/2/6/which-countries-have-banned-deepseek-and-why

[50] Gartner, “Gartner identifies six steps to manage artificial intelligence agent sprawl,” *Gartner press release*, Apr. 28, 2026. [Online]. Available: https://www.gartner.com/en/newsroom/press-releases/2026-04-28-gartner-identifies-six-steps-to-manage-artificial-intelligence-agent-sprawl

[51] Cato Networks, “CurXecute: Remote code execution in Cursor via MCP (CVE-2025-54135),” *Cato Networks blog*, 2025. [Online]. Available: https://www.catonetworks.com/blog/curxecute-rce/

[52] J. Rehberger, “GitHub Copilot: Remote code execution via prompt injection (CVE-2025-53773),” *Embrace The Red*, Aug. 2025. [Online]. Available: https://embracethered.com/blog/posts/2025/github-copilot-remote-code-execution-via-prompt-injection/

[53] Trail of Bits, “Jumping the line: How MCP servers can attack you before you ever use them,” *Trail of Bits blog*, Apr. 21, 2025. [Online]. Available: https://blog.trailofbits.com/2025/04/21/jumping-the-line-how-mcp-servers-can-attack-you-before-you-ever-use-them/

[54] The Hacker News, “First malicious MCP server found in the wild,” *The Hacker News*, Sep. 2025. [Online]. Available: https://thehackernews.com/2025/09/first-malicious-mcp-server-found.html

[55] JFrog Security Research, “Critical RCE vulnerability in mcp-remote: CVE-2025-6514,” *JFrog blog*, 2025. [Online]. Available: https://jfrog.com/blog/2025-6514-critical-mcp-remote-rce-vulnerability/

[56] Astrix Security, “State of MCP server security 2025,” *Astrix Security*, 2025. [Online]. Available: https://astrix.security/learn/blog/state-of-mcp-server-security-2025/

[57] Knostic, “Mapping MCP servers: A study of internet-exposed servers,” *Knostic blog*, Jul. 2025. [Online]. Available: https://www.knostic.ai/blog/mapping-mcp-servers-study

[58] Wiz Research, “s1ngularity's aftermath,” *Wiz blog*, 2025. [Online]. Available: https://www.wiz.io/blog/s1ngularitys-aftermath

[59] Fortune, “AI coding tool Replit wiped database, called it a catastrophic failure,” *Fortune*, Jul. 23, 2025. [Online]. Available: https://fortune.com/2025/07/23/ai-coding-tool-replit-wiped-database-called-it-a-catastrophic-failure/

[60] BleepingComputer, “Amazon AI coding agent hacked to inject data wiping commands,” *BleepingComputer*, Jul. 2025. [Online]. Available: https://www.bleepingcomputer.com/news/security/amazon-ai-coding-agent-hacked-to-inject-data-wiping-commands/

[61] Invariant Labs, “GitHub MCP exploited: Accessing private repositories via MCP,” *Invariant Labs blog*, May 26, 2025. [Online]. Available: https://invariantlabs.ai/blog/mcp-github-vulnerability

[62] P. Reddy and A. S. Gujral, “EchoLeak: The first real-world zero-click prompt injection exploit in a production LLM system,” *arXiv:2509.10540*, 2025. [Online]. Available: https://arxiv.org/html/2509.10540v1

[63] Noma Security, “ForcedLeak: AI agent risks exposed in Salesforce Agentforce,” *Noma Security blog*, Sep. 25, 2025. [Online]. Available: https://noma.security/blog/forcedleak-agent-risks-exposed-in-salesforce-agentforce

[64] Legit Security, “CamoLeak: Critical GitHub Copilot vulnerability leaks private source code,” *Legit Security blog*, 2025. [Online]. Available: https://www.legitsecurity.com/blog/camoleak-critical-github-copilot-vulnerability-leaks-private-source-code

[65] Zenity Labs, “AgentFlayer: ChatGPT Connectors 0-click attack,” *Zenity Labs*, Aug. 2025. [Online]. Available: https://labs.zenity.io/p/agentflayer-chatgpt-connectors-0click-attack-5b41

[66] SafeBreach Labs, “Invitation is all you need: Hacking Gemini,” *SafeBreach blog*, 2025. [Online]. Available: https://www.safebreach.com/blog/invitation-is-all-you-need-hacking-gemini/

[67] Microsoft, “AI recommendation poisoning,” *Microsoft Security Blog*, Feb. 10, 2026. [Online]. Available: https://www.microsoft.com/en-us/security/blog/2026/02/10/ai-recommendation-poisoning/

[68] Varonis Threat Labs, “Reprompt: Single-click Copilot prompt injection,” *Varonis blog*, 2026. [Online]. Available: https://www.varonis.com/blog/reprompt

[69] Brave, “Unseeable prompt injections in screenshots: More vulnerabilities in Comet and other AI browsers,” *Brave blog*, Oct. 21, 2025. [Online]. Available: https://brave.com/blog/unseeable-prompt-injections/

[70] LayerX Security, “CometJacking: How one click can turn Perplexity's Comet AI browser against you,” *LayerX blog*, Oct. 2025. [Online]. Available: https://layerxsecurity.com/blog/cometjacking-how-one-click-can-turn-perplexitys-comet-ai-browser-against-you/

[71] Guardio Labs, “Scamlexity: We put agentic AI browsers to the test - they clicked, they paid, they failed,” *Guardio Labs*, 2025. [Online]. Available: https://guard.io/labs/scamlexity-we-put-agentic-ai-browsers-to-the-test-they-clicked-they-paid-they-failed

[72] Anthropic, “Piloting Claude for Chrome,” *Anthropic*, Aug. 2025. [Online]. Available: https://claude.com/blog/claude-for-chrome

[73] Cisco Talos, “Cybercriminals camouflaging threats as AI tool installers,” *Cisco Talos blog*, May 29, 2025. [Online]. Available: https://blog.talosintelligence.com/fake-ai-tool-installers/

[74] Trend Micro Research, “Claude.ai shared chat abused in malvertising campaign,” *Trend Micro*, Jun. 2026. [Online]. Available: https://www.trendmicro.com/en_us/research/26/f/claudeai-shared-chat-abused-in-malvertising.html

[75] Push Security, “LLMShare: A live malvertising campaign abusing shared AI chats,” *Push Security blog*, May 29, 2026. [Online]. Available: https://pushsecurity.com/blog/llmshare-malvertising-campaign

[76] Cato CTRL, “When trust becomes payload: Fake Codex ClickFix campaign,” *Cato Networks blog*, Aug. 24, 2026. [Online]. Available: https://www.catonetworks.com/blog/cato-ctrl-when-trust-becomes-payload-in-fake-codex-clickfix-campaign/

[77] Push Security, “What Push data reveals about the state of shadow AI,” *Push Security blog*, May 28, 2026. [Online]. Available: https://pushsecurity.com/blog/what-push-data-reveals-about-the-state-of-shadow-ai

[78] LayerX Security, “AiFrame: Fake AI assistant extensions targeting 260,000 Chrome users via injected iframes,” *LayerX blog*, 2025. [Online]. Available: https://layerxsecurity.com/blog/aiframe-fake-ai-assistant-extensions-targeting-260000-chrome-users-via-injected-iframes/

[79] Dark Reading, “Chrome extension harvests AI chatbot data,” *Dark Reading*, Dec. 2025. [Online]. Available: https://www.darkreading.com/endpoint-security/chrome-extension-harvests-ai-chatbot-data

[80] Endor Labs, “Malicious payload in ai-sdk-ollama npm package,” *Endor Labs*, Jun. 4, 2026. [Online]. Available: https://www.endorlabs.com/learn/malicious-payload-in-ai-sdk-ollama-npm-package

[81] ReversingLabs, “RL identifies malware ML model hosted on Hugging Face,” *ReversingLabs blog*, Jan. 2025. [Online]. Available: https://www.reversinglabs.com/blog/rl-identifies-malware-ml-model-hosted-on-hugging-face

[82] Pillar Security, “LLM backdoors at the inference level: The threat of poisoned templates,” *Pillar Security blog*, Jun. 2025. [Online]. Available: https://www.pillar.security/blog/llm-backdoors-at-the-inference-level-the-threat-of-poisoned-templates

[83] Pillar Security, “New vulnerability in GitHub Copilot and Cursor: How hackers can weaponize code agents,” *Pillar Security blog*, Mar. 18, 2025. [Online]. Available: https://www.pillar.security/blog/new-vulnerability-in-github-copilot-and-cursor-how-hackers-can-weaponize-code-agents

[84] The Hacker News, “Researchers find 175,000 publicly exposed Ollama AI servers,” *The Hacker News*, Jan. 2026. [Online]. Available: https://thehackernews.com/2026/01/researchers-find-175000-publicly.html

[85] J. Spracklen et al., “We have a package for you! A comprehensive analysis of package hallucinations by code generating LLMs,” *in Proc. 34th USENIX Security Symposium*, 2025. [Online]. Available: https://socket.dev/blog/slopsquatting-how-ai-hallucinations-are-fueling-a-new-class-of-supply-chain-attacks

[86] Veracode, “Spring 2026 GenAI code security update,” *Veracode blog*, Mar. 24, 2026. [Online]. Available: https://www.veracode.com/blog/spring-2026-genai-code-security/

[87] Netcraft, “Large language models are falling for phishing scams,” *Netcraft blog*, Jul. 1, 2025. [Online]. Available: https://www.netcraft.com/blog/large-language-models-are-falling-for-phishing-scams

[88] Hoxhunt, “AI-powered phishing vs. humans,” *Hoxhunt*, 2025. [Online]. Available: https://hoxhunt.com/blog/ai-powered-phishing-vs-humans

[89] Federal Bureau of Investigation, “Cryptocurrency and AI scams bilk Americans of billions,” *FBI press release (IC3 2025 report)*, Apr. 6, 2026. [Online]. Available: https://www.fbi.gov/news/press-releases/cryptocurrency-and-ai-scams-bilk-americans-of-billions

[90] IBM, “IBM 2026 X-Force Threat Index: AI-driven attacks are escalating as basic security gaps leave enterprises exposed,” *IBM Newsroom*, Feb. 25, 2026. [Online]. Available: https://newsroom.ibm.com/2026-02-25-ibm-2026-x-force-threat-index-ai-driven-attacks-are-escalating-as-basic-security-gaps-leave-enterprises-exposed

[91] The Hacker News, “Infostealer logs expose replayable AI session tokens,” *The Hacker News*, Sep. 2026. [Online]. Available: https://thehackernews.com/2026/09/infostealer-logs-expose-replayable-ai.html

[92] LayerX Security, “Browser security report 2025 (as reported by The Hacker News),” *The Hacker News*, Nov. 2025. [Online]. Available: https://thehackernews.com/2025/11/new-browser-security-report-reveals.html

[93] B. Krebs, “The ongoing fallout from a breach at AI chatbot maker Salesloft,” *Krebs on Security*, Sep. 2025. [Online]. Available: https://krebsonsecurity.com/2025/09/the-ongoing-fallout-from-a-breach-at-ai-chatbot-maker-salesloft/

[94] BleepingComputer / Push Security, “Learning from the Vercel breach: Shadow AI and OAuth sprawl,” *BleepingComputer*, 2026. [Online]. Available: https://www.bleepingcomputer.com/news/security/learning-from-the-vercel-breach-shadow-ai-and-oauth-sprawl/

[95] Autoriteit Persoonsgegevens, “Caution: Use of AI chatbot may lead to data breaches,” *Dutch Data Protection Authority*, Aug. 2024. [Online]. Available: https://www.autoriteitpersoonsgegevens.nl/en/current/caution-use-of-ai-chatbot-may-lead-to-data-breaches

[96] CNIL, “CNIL's Q&A on the use of generative AI systems,” *Commission nationale de l'informatique et des libertes*, 2024. [Online]. Available: https://www.cnil.fr/en/cnils-qa-use-generative-ai-systems

[97] Covington & Burling, “European Commission publishes Q&A on AI literacy,” *Inside Privacy*, 2025. [Online]. Available: https://www.insideprivacy.com/artificial-intelligence/european-commission-publishes-qa-on-ai-literacy/

[98] National Law Review, “AI meets HIPAA security: Understanding HHS's risk strategies and proposed changes,” *National Law Review*, 2025. [Online]. Available: https://natlawreview.com/article/ai-meets-hipaa-security-understanding-hhss-risk-strategies-and-proposed-changes

[99] HIPAA Journal, “HIPAA Security Rule update postponed,” *The HIPAA Journal*, 2026. [Online]. Available: https://www.hipaajournal.com/hipaa-security-rule-update-postponed/

[100] Emergo by UL, “Decoding MDCG 2025-6: Interplay between MDR/IVDR and the AI Act,” *Emergo by UL*, 2025. [Online]. Available: https://www.emergobyul.com/news/decoding-mdcg-2025-6-interplay-between-mdrivdr-and-aia

[101] Team-NB, “Position paper on the EU AI Act, version 2,” *Team-NB*, Apr. 9, 2025. [Online]. Available: https://www.team-nb.org/wp-content/uploads/2025/04/Team-NB-PositionPaper-EU-AI-Act-V2-20250409.pdf

[102] ITWeb, “InfoReg unpacks regulatory complexities posed by AI,” *ITWeb*, Sep. 17, 2026. [Online]. Available: https://www.itweb.co.za/article/inforeg-unpacks-regulatory-complexities-posed-by-ai/KA3WwMdzPryvrydZ

[103] American Bar Association, “ABA issues first ethics guidance on a lawyer's use of AI tools (Formal Opinion 512),” *ABA News*, Jul. 29, 2024. [Online]. Available: https://www.americanbar.org/news/abanews/aba-news-archives/2024/07/aba-issues-first-ethics-guidance-ai-tools/

[104] Fenwick & West, “The end of silent AI: Emerging AI exclusions, coverage fragmentation and practical implications,” *Fenwick insights*, 2026. [Online]. Available: https://www.fenwick.com/insights/publications/end-silent-ai-emerging-ai-exclusions-coverage-fragmentation-and-practical-implications

[105] Bird & Bird, “First ruling on works council rights in the use of artificial intelligence (ArbG Hamburg, 24 BVGa 1/24),” *Bird & Bird insights*, 2024. [Online]. Available: https://www.twobirds.com/en/insights/2024/germany/erstes-urteil-zu-rechten-des-betriebsrats-bei-einsatz-von-kuenstlicher-intelligenz

[106] Microsoft, “Use Microsoft Purview to manage data security and compliance for other AI apps,” *Microsoft Learn*, 2026. [Online]. Available: https://learn.microsoft.com/en-us/purview/ai-other-apps

[107] Gartner, “Gartner identifies critical GenAI blind spots that CIOs must urgently address,” *Gartner press release*, Nov. 19, 2025. [Online]. Available: https://www.gartner.com/en/newsroom/press-releases/2025-11-19-gartner-identifies-critical-genai-blind-spots-that-cios-must-urgently-address0

[108] ISACA, “AI use accelerates while governance and ROI lag, says new ISACA research,” *ISACA press release*, May 2026. [Online]. Available: https://www.isaca.org/about-us/newsroom/press-releases/2026/ai-use-accelerates-while-governance-and-roi-lag-says-new-isaca-research

[109] Kiteworks, “The AI policy gap: Shadow AI,” *Kiteworks*, 2026. [Online]. Available: https://www.kiteworks.com/cybersecurity-risk-management/ai-policy-gap-shadow-ai/

[110] University of Melbourne and KPMG, "Trust, attitudes and use of artificial intelligence: A global study 2025 (media release)," *University of Melbourne*, 2025. [Online]. Available: https://fbe.unimelb.edu.au/newsroom/media-release-global-study-reveals-trust-of-ai-remains-a-critical-challenge-reflecting-tension-between-benefits-and-risks

[111] Harmonic Security, "Employees use their personal AI accounts for work 64% of the time, potentially invisible to their employers, research finds," *Business Wire*, May 20, 2026. [Online]. Available: https://www.businesswire.com/news/home/20260520319293/en/Employees-Use-Their-Personal-AI-Accounts-for-Work-64-of-the-Time-Potentially-Invisible-to-Their-Employers-Research-Finds

[112] Gartner, "Gartner forecasts the market for securing AI will reach almost $5 billion in 2027," *Gartner press release*, Aug. 26, 2026. [Online]. Available: https://www.gartner.com/en/newsroom/press-releases/2026-08-26-gartner-forecasts-the-market-for-securing-ai-will-reach-almost-5-billion-in-2027

[113] KuppingerCole Analysts, "Leadership Compass: SaaS security and AI governance," *KuppingerCole*, 2026. [Online]. Available: https://www.kuppingercole.com/research/lc81081/saas-security-and-ai-governance

[114] SentinelOne, "SentinelOne to acquire Prompt Security to advance GenAI security," *SentinelOne press release*, Aug. 5, 2025. [Online]. Available: https://www.sentinelone.com/press/sentinelone-to-acquire-prompt-security-to-advance-genai-security/

[115] SecurityWeek, "Cato Networks acquires AI security firm Aim Security," *SecurityWeek*, Sep. 3, 2025. [Online]. Available: https://www.securityweek.com/cato-networks-acquires-ai-security-firm-aim-security/

[116] Check Point Software, "Check Point acquires Lakera to deliver end-to-end AI security for enterprises," *Check Point press release*, Sep. 16, 2025. [Online]. Available: https://www.checkpoint.com/press-releases/check-point-acquires-lakera-to-deliver-end-to-end-ai-security-for-enterprises/

[117] CrowdStrike, "CrowdStrike to acquire Pangea to secure every layer of enterprise AI," *CrowdStrike press release*, Sep. 16, 2025. [Online]. Available: https://www.crowdstrike.com/en-us/press-releases/crowdstrike-to-acquire-pangea-to-secure-every-layer-of-enterprise-ai/

[118] Zscaler, "Zscaler secures the enterprise AI lifecycle with acquisition of AI security pioneer SPLX," *Zscaler press release*, Nov. 3, 2025. [Online]. Available: https://www.zscaler.com/press/zscaler-secures-enterprise-ai-lifecycle-acquisition-innovative-ai-security-pioneer-splx

[119] The Hacker News, "ShadowLeak zero-click flaw leaks Gmail data via ChatGPT Deep Research agent," *The Hacker News*, Sep. 2025. [Online]. Available: https://thehackernews.com/2025/09/shadowleak-zero-click-flaw-leaks-gmail.html

© 2026 Lamplighter AI. Third-party figures are reproduced from the sources cited and have not been independently audited by Lamplighter AI. Product capabilities are described as of Lamplighter v0.18 (October 2026) and were audited against the product's code and automated tests: "Available" capabilities are in the product today; "In development", "Roadmap" and "Proposed" capabilities are being built, planned or under consideration, and are subject to change. Product and company names are trademarks of their respective owners. This paper is not legal advice.