Web site and AI agent security

What an automated scan sees, and what a human audit must see

Version 1.0 - August 2026

Horus Solutions - IT consulting and engineering firm

1. Why this paper

An automated test has become the first reflex: paste a URL, get a score, share a screenshot. That is useful. It is not governance.

The score describes a configuration visible from the Internet. It does not say whether the organisation has an owner, a mandate, an enforceable record, or whether a conversational agent can extract data or trigger a business action. Mixing the two exposes the company twice: first by believing it is covered, then on the day an incident requires a deliverable the scan never produced.

This paper is for the decision-maker - CIO, CISO, DPO, executive, integrator - who must circulate an argument internally. The technician already has the Horus Web Check and Horus AI Check Guides. The aim here is simpler: distinguish what an automated scan sees, what it does not see, and when a human audit under a written mandate becomes necessary.

Horus offers two public sister platforms, with no ERP link:

- Horus Web Check - pre-check of a website (TLS, DNS, headers, exposure).
- Horus AI Check - pre-check of an AI application (chatbot, RAG, agent, LLM API, MCP).

Neither is a penetration test. Neither is an audit. A Horus score is not a pentest.

2. Two surfaces, one governance

The digital company now has two exposed façades.

The first is the website: certificates, redirects, cookie policy, forgotten files at the root, reachable administration interfaces, mail authentication (SPF, DKIM, DMARC). This is the familiar ground of OWASP Top 10, OWASP Secure Headers, CWE and ISO/IEC 27001:2022 controls - indicative mappings, not a certification.

The second is the AI agent: a natural-language dialogue wired to tools, a document base (RAG), sometimes an MCP protocol, sometimes a metered API. Risks are no longer read only in a certificate. They are read in an overly permissive system prompt, retrieval that is too broad, a tool that writes into a file, a context leak from one customer to another. The vocabulary changes (OWASP LLM / GenAI); accountability does not.

Governing both as if there were only one is a common mistake. An "A+" TLS site does not stop a chatbot from returning an internal document. A "well prompted" agent does not excuse a public `.env`. Both surfaces need the same discipline - named owner, written scope, proof of ownership, logs - and different controls.

At Horus, an engagement covers a single track at a time: website or AI. A client with both a site and a chatbot follows two journeys. This is not a commercial device. It is the only way not to mix findings, mandates and deliverables.

Governance starts with prosaic questions, too often left to the model vendor or the web agency:

- Who is allowed to publish a new site version or a new system prompt?
- Where are the secrets (certificates, API keys, MCP tokens) and who revokes them?
- What happens if the AI subcontractor changes the model overnight?
- Which log proves that a user - or an agent - consulted a file?

Without written answers, the best score in the world only reassures a committee. With written answers, even an N1 report becomes actionable: you know whom to send it to, and you know what falls outside the scan.

3. The four Horus levels

Horus structures the work in four levels. They do not replace one another. They follow one another.

N1 - Free automated test on the platform

The user pastes a URL (site) or declares an AI target, within usage and ownership limits. The platform computes a report and a score. Deliverable: a visible configuration state, with no exploit payloads. Mandatory notice: this is not a Horus audit.

N2 - Deeper tests, still automated, still on the same platform

Paid options: sandbox and screenshot, multi-engine reputation, in-depth TLS grade, PDF retained for twelve months, exposure report (web); PRO pack (AI) for extra automated scenarios. Still non-destructive. Still on Web Check or AI Check - never a "pentest launched in one click".

N3 - In-depth human audit under a written mandate

A Horus operator works on a case, with a mandatory track (SITE_WEB or IA). The work uses the imported scan, a business checklist, and audit tools under contract. Deliverable: a dated audit report, enforceable in the client relationship, distinct from the automated PDF.

N4 - Remediation by Horus

After N3, if the client orders remediation: tracking of fixes (configuration, code, prompts, policies), verification, closure. This is not a scanner. It is engineering.

The sentence to keep: a score is not a pentest. N1 and N2 measure indicators. N3 records findings under a mandate. N4 fixes. Moving from N1 to N3 without saying so is selling an illusion.

4. Website chapter - what the scan sees

Horus Web Check observes what an HTTP client, a TLS handshake and a DNS resolution can observe from outside.

What N1 (free) sees

- DNS, IP address, redirect chain, final URL.
- HTTP status, response time.
- TLS certificate: HTTPS present, issuer, validity dates, obsolete protocols.
- Security headers: HSTS, CSP, clickjacking protection, cookies (Secure, HttpOnly, SameSite).
- Page: mixed HTTP/HTTPS content, minimal quality (language, title).
- Domain mail authentication: SPF, DKIM, DMARC - indicative, not a full mail audit.

- Sensitive files at the root when they answer (signals, not an internal map).
- Compression and a few performance signals - useful, less serious than a missing certificate.

The Horus score is weighted: a missing favicon does not weigh like an open `.git`. A DANGEROUS verdict is not triggered by a single cosmetic gap.

What N2 (credits) adds

- Isolated sandbox and screenshot: see the rendered page without visiting it yourself.
- Reputation (licensed specialised engines): a signal that a URL is already known as malicious - not a legal verdict.
- Laboratory-style in-depth TLS analysis: a grade that takes time, useful in a configuration audit.
- Horus PDF retained twelve months: dated proof, distinct from the screen that expires.
- Exposure report, after proof of ownership: limited probes (listing, banners, admin paths, upload hint). Still automated black-box, never exploitation.

Proof of ownership (`/.well-known/horus-webcheck.txt` or a DNS TXT record) is not a whim. Deep-scanning a third party's site without a mandate is a fault. N2 exposure and N3 require it.

What the scan does not see

- Real authentication (passwords, SSO, MFA, session).
- Business logic: a payment form that looks "green" on headers may still be weak in the application.
- The internal network, unpublished APIs, back offices off the Internet.
- Processes: who deploys, who revokes a certificate, who reads the logs.
- Controlled exploitation of a flaw (that is N3 under mandate, or a contractual pentest).

A site can therefore show a high score and remain unacceptable for a DPO: tracking cookies without a legal basis, a health form in clear text, an admin panel on a default password the scan does not try. Conversely, a medium score may hide a healthy site whose only gap is an incomplete CSP. The report is a starting point, not a sentence.

Frequent readings, to avoid false debates in committee:

- Missing HTTPS or expired certificate: not an infrastructure detail. It is channel confidentiality. High priority, often a fast fix.
- Missing HSTS or CSP: a maturity signal. Serious when stacked with other gaps; rarely a crisis on its own.
- Incomplete SPF / DMARC: the site may be "clean" while the domain is used for mail spoofing. The scan flags it; the fix is DNS and sending policy, not a cookie banner.
- Sensitive file or listing: treat as a disclosure incident until proven otherwise, then harden the server.
- Poor performance score: comfort and SEO, not a vulnerability. Do not confuse it with a security verdict.

The exposure report (N2, proof of ownership) adds a layer: it looks for hints of a wider surface (admin, repository, upload). It does not replace an asset inventory. If a probe finds something, N3 is there to confirm without improvising from an unmandated workstation.

5. AI agents chapter - what the scan sees

Horus AI Check targets a reachable AI application: chatbot, RAG, tool-using agent, model API, MCP server. The automated scanner (Promptfoo-style indicators, isolated) looks for vulnerability signals. It does not "hack" the bot.

What N1 (FREE) sees

About twenty-five tests depending on the target type, within a time budget. The report gives a score out of 100, a grade, domains (leakage, instruction injection, robustness, and so on) and findings. Detailed evidence and payloads stay masked in the public deliverable: AI Check is not an attack kit.

Before the scan: identity, timestamped authorisation, e-mail confirmation, proof of ownership (DNS, well-known file, or manual validation). One does not run a test suite against a third party's chatbot "just to see".

What N2 (PRO) adds

Broader automated scenarios - bot authentication, stricter retrieval, tools - still non-destructive. The report PDF, if purchased, keeps a dated history, without payloads.

AI Check is not an AI Pentest

This sentence must appear everywhere: UI, terms, e-mails, reports. An AI pentest is a human, contractual, authenticated mission, with business scenarios, prompt review, RAG chain review, and often a specialised partner. AI Check prepares the ground and qualifies a need. It does not replace it.

What the scan does not see

- Internal accounts, roles, API secrets, real token billing.
- A RAG wired to a document share the public bot never cites in FREE tests.
- A business tool that writes (creating a file, sending mail, moving stock) once the agent is authenticated.
- Compliance (AI Act, GDPR, professional secrecy): a "possible leak" finding is not a data-protection impact assessment.
- Exploitation: actually extracting a database, bypassing a guardrail, chaining tools until harm.

An agent can therefore "pass" N1 and remain unacceptable: it has write access to an ERP, it mixes two customers' contexts, it has no human owner. The score measures conversational-surface indicators. Governance measures the agent's real power.

Four families of questions help a committee read an AI Check report without over-interpreting it.

First family: the channel. Is the tested URL production? A staging bot open on the Internet is not the same stake as an authenticated assistant behind a VPN. The scan does not guess; the N3 mandate does.

Second family: the data. Does the agent see only public documents, or the customer file, payroll, health, litigation? FREE looks for instruction and context leaks. It does not classify your processing under GDPR.

Third family: the tool. A chatbot that only answers is not an agent that creates a record, sends mail or triggers a payment. As soon as a write tool exists, N1 is not enough as the only net. N2 can widen automated tests. N3 examines the real chain.

Fourth family: identity. Who talks to the bot - anonymous, customer, employee? Which logs? An agent with no identity in the information system is the equivalent of a shared account: convenient, indefensible on audit day.

Instruction injection (prompt injection) and RAG contamination are real. They are not "patched" like an HSTS header. They require design: separating instructions from documents, limiting tools, human supervision of acts, regular tests. The automated scan detects symptoms. The human audit reconstructs the design.

Tatiana - we use our own tools on our own infrastructure.
Horus runs an internal assistant, Tatiana, on its infrastructure. She has a product identity distinct from a human login: no shared HTTP session, explicit rights, traces, read-only business access when she consults data. When a colleague speaks to her, the colleague is accountable for the question; when she acts in the background, she is the one who is traced.
This box is not a commercial offer. It states a requirement we apply to ourselves before we ask it of a client: an AI agent is an actor in the information system. It has a name, a scope, logs. It is not an anonymous input field wired to the Internet.

6. When to move to N3

N3 is not "the paid level of the scan". It is the moment the organisation needs a human, dated, mandated, enforceable record.

Frequent signals, not exhaustive:

- Low score with at least one critical gap (web) or high / critical findings (AI).
- Personal data, health data, financial data, trade secrets in the target scope.
- An agent with tools that write or call a business API.
- Administration interface, repository, backup or environment file exposed.
- Missing HTTPS, expired certificate, missing HSTS on a site that authenticates users.
- Contractual or regulatory duty to produce an audit deliverable (client, insurer, authority, ISO).
- Internal disagreement: engineering says "it is nothing", the business says "we cannot stay like this".
- Need for tracked remediation (N4): one does not order a fix on the faith of a screen that expires in twenty-four hours.

The right reflex is not to buy credits "until the score looks nice". The right reflex is to decide whether the stake justifies a mandate. N2 credits deepen the measurement. N3 establishes the record.

Three concrete situations help decide.

A brochure site with no user accounts, a valid certificate, correct headers, no sensitive file: N1 is often enough as a yearly snapshot, optionally N2 PDF for the archive. N3 is justified only if a client contract requires it or if the activity changes (customer area, payment).

A site with authentication, customer data, exposed admin, missing DMARC: N2 (exposure + PDF) documents; N3 becomes reasonable before high season or an ISO renewal. Waiting for the incident to mandate costs more than the audit.

An agent connected to business tools, even with a good FREE grade: treat it as a new actor in the information system. N3 AI track, tool scope, prompt and retrieval review. N4 if safeguards are missing (read-only, traces, identity). The scan served as an alert, not a discharge.

In every case, the decision-maker must be able to explain to a board, in one sentence, what was measured automatically and what was not audited. If that sentence is impossible, you are no longer in N1: you are in the fog. Fog is not a Horus level.

7. Decision-maker checklist

Before closing the topic - or before writing to Horus for an audit:

- A named owner (business + security) for the site, and a distinct owner or the same person explicitly for each AI agent.

- Proof of ownership of the domain or endpoint is possible (DNS or well-known file).
- The scope is written: URL, environments, exclusions, data in transit.
- N1 (indicator) and N3 (audit) are distinguished. Nobody in the chain calls the free scan a pentest.
- Data: categories, legal basis, retention, AI subprocessors where relevant.
- Agent tools: read-only or write; who validates an action.
- Logs: who reads them, for how long, what is done with a finding.
- A written mandate is ready if N3 is decided (SITE_WEB track or IA track, not both mixed).
- An N4 budget is discussed only after the N3 record, not as a scan option.
- This paper and the scan report are classified: they are not audit documents.

8. Notices

Horus Solutions document, version 1.0, August 2026. Public use, free download.

The mappings cited (OWASP Top 10:2021, OWASP Secure Headers, OWASP LLM / GenAI, CWE, ISO/IEC 27001:2022) are indicative. Horus Web Check is not an OWASP ASVS audit. Horus AI Check is not an AI Pentest. This white paper is not a certification, not legal advice, not an audit report.

No exploit payload is published here. No attack recipe. The platforms do not expose offensive tests in the clear.

Horus Web Check and Horus AI Check are standalone sites. They share no data with an ERP. A scan does not open a company account at Horus.

For an N3 audit or N4 remediation: request from the platform, under a written mandate, one track at a time.

Horus Solutions - Everywhere you are, you can work.