NEONIX SECURITY

EXECUTIVE WHITEPAPER / AUSTRALIA

HUMAN-LED OFFENSIVE SECURITY IN THE AGE OF AI

Practical governance for human-led, AI-assisted offensive security

Where AI assists, where humans decide, and what the evidence must prove.

FOR BOARDS, CHIEF EXECUTIVES, AND TECHNOLOGY, RISK AND SECURITY LEADERS

PUBLISHED AUGUST 2026

Executive summary

Artificial intelligence is changing how offensive security work is performed. It can help consultants process large attack surfaces, inspect source code, generate test cases, develop hypotheses, analyse evidence and draft repeatable checks. It can also produce convincing errors, expose sensitive client information, lose context, cross scope boundaries or execute an unsafe action when given excessive authority.

The relevant choice is therefore not between human testing and automated testing. It is how to combine qualified human judgement, established security tooling and AI assistance without weakening authorisation, evidence quality or accountability.

The recommended operating model is **human-led offensive security**. A named consultant remains responsible for scope, attack hypotheses, tool use, safety decisions, finding validation, impact assessment and reporting. AI may increase speed and coverage, but its output is treated as untrusted working material until a qualified person verifies it against the target and preserves reproducible evidence.

This model applies across penetration testing, adversary simulation, application and API security, cloud and identity testing, infrastructure assessment, embedded and operational technology, vulnerability research, and testing of AI-enabled systems.

Current industry evidence supports this direction. CREST's 2026 research found AI already present in professional penetration-testing workflows, particularly reporting and vulnerability scanning or enumeration, while autonomous agent-based testing remained uncommon. CREST's conclusion is that AI is augmenting practitioners rather than replacing them, and that accountability, transparency, validation and data control are becoming material client requirements. [1]

Australian guidance is consistent. The Australian Cyber Security Centre says AI can strengthen risk prioritisation, vulnerability discovery, detection and response, but requires human oversight, least privilege, protection of sensitive information, validation of outputs and auditability of AI-assisted actions. [2]

For CISOs, the practical implication is that AI guardrails should be treated like other security controls: mapped to permissions, data flows and executable actions; tested under adversarial conditions; monitored in operation; and revalidated after material change. [2,17]

Five decisions for leaders

1. **Keep a qualified human accountable for every engagement.** AI cannot accept scope, exercise professional judgement or own the consequences of a test.
2. **Use AI to extend coverage, not lower the evidence standard.** Every reported finding still needs a reproducible path, affected target, demonstrated effect, impact statement and human sign-off.

3. **Protect client information by design.** Approved models, data classifications, retention rules, regional constraints and technical access controls should be set before client data or evidence reaches an AI system.
4. **Separate assistance from authority.** Reading, summarising and proposing are lower risk than sending traffic, changing state, using credentials or communicating externally. Authority should increase only with explicit approval and containment.
5. **Tell clients how AI materially affects delivery.** Contracts and reports should identify material AI use, data handling, human validation and any limitations without overstating capability.

HUMAN LEADS	AI ASSISTS	EVIDENCE DECIDES
Scope, direction, safety and accountability.	Analysis, generation, organisation and drafting.	Reproducibility, target behaviour and demonstrated impact.

1. The operating premise

1.1 Offensive security has always combined people and tools

Professional testing already uses scanners, fuzzers, debuggers, disassemblers, exploit frameworks, attack-graph tools and custom automation. Those tools perform work at a scale or precision that manual effort cannot match. The consultant decides what to run, interprets the result, changes direction and determines whether an observed condition creates a real security capability.

NIST's technical testing guidance separates planning, execution, analysis and mitigation. It also stresses that each testing technique has benefits and limitations. The introduction of AI changes the capabilities and failure modes of the toolset; it does not remove the need for a disciplined testing process. [3]

1.2 AI changes the workflow, not accountability

Generative models can reason over mixed technical material and produce code, queries, explanations and candidate attack paths. This makes them useful at several points in an engagement. It also makes errors harder to spot because incorrect output can be fluent, internally coherent and tailored to the user's context.

NIST describes this failure mode as confabulation: confidently presented erroneous or false content. In offensive work, a confabulation can become a false finding, an invalid exploit condition, a destructive command or a misleading remediation recommendation. [4]

The appropriate control is not a generic instruction to "be accurate". It is a workflow in which AI output has no evidentiary status until it is checked against source material, target behaviour and independently reproducible results.

1.3 What human-led means

Human-led does not mean manually performing every task. It means a qualified person retains:

- **authority:** confirming the rules of engagement and permitted actions;
- **direction:** selecting hypotheses, techniques and escalation paths;
- **context:** interpreting business logic, operational constraints and client intent;
- **control:** approving consequential actions and stopping unsafe activity;
- **validation:** reproducing findings and distinguishing evidence from inference;
- **accountability:** signing the report and explaining the result to the client.

AI and deterministic automation can operate within that structure. They should not silently redefine it.

2. Where AI can assist the offensive workflow

AI adds the most value where it reduces repetitive analysis, helps a consultant explore alternatives or converts verified knowledge into repeatable checks. The value and required control differ by stage.

The matrix describes assistance, not guaranteed performance. The same model may produce different results across runs, and model or provider changes can alter behaviour without a change to the engagement methodology.

Delivery stage	AI can assist	Human gate and control
Scoping and preparation	Summarise supplied architecture, reports and documentation; identify questions and dependencies	Confirm authority, exclusions, safety limits and assumptions. Use approved inputs and never infer a wider scope.
Attack-surface analysis	Correlate assets, routes, identities, APIs, code paths and configuration; propose trust boundaries	Verify the live scope and prioritise paths using threat context. Preserve source provenance.
Code and design review	Explain unfamiliar code, trace data flow and generate candidate review queries	Inspect the relevant code and prove reachability, attacker control and security effect. Model output alone is not a finding.
Test development	Draft test cases, protocol inputs, scripts, fuzz harnesses and payload variations	Review safety and execute only in scope. Sandbox generated code and peer-review higher-risk tooling.
Vulnerability research	Cluster crashes, compare variants and assist with hypotheses and root-cause analysis	Establish reproducibility, provenance, primitive, capability and impact on the tested build. Independently validate.
Adversary simulation	Assist with threat research, procedure mapping, infrastructure-as-code and detection hypotheses	Select authorised tradecraft and control operational risk through rules of engagement, allowlists and segregated infrastructure.
Evidence analysis	Organise logs, traces, screenshots and command output; identify gaps	Confirm evidence came from the target and supports the claim. Preserve immutable

Delivery stage	AI can assist	Human gate and control
	and inconsistencies	originals and traceable transformations.
Reporting	Draft descriptions, normalise terminology, check consistency and propose remediation language	Human authors and reviewers own accuracy, severity, client context and final wording. Remove unsupported claims.
Re-test and regression	Convert validated findings into repeatable negative tests and compare results across versions	Confirm the original path is closed and plausible variants remain considered. Retain versioned tests and closure evidence.

2.1 Practical examples

AI can help a web tester turn an API specification into a candidate authorisation matrix, but the tester must verify which roles and object relationships matter. It can help an infrastructure tester compare hundreds of policy statements, but the tester must prove the effective privilege path. It can help a vulnerability researcher explain a crash or draft a harness, but the researcher must establish attacker control, target-side effect and exploitability on the real build.

Google Project Zero's Big Sleep work provides a useful bounded example. An AI agent helped identify a previously unknown exploitable memory-safety flaw in SQLite, which human researchers validated and reported before it reached an official release. The relevant lesson is demonstrated assistance within a research process, not that vulnerability research has become autonomous. [5]

CREST's 2026 findings similarly place current use in ordinary workflow stages. In its survey, 47% of participating organisations reported using AI for reporting and 44% for vulnerability scanning and enumeration, while 9% reported autonomous agent-based testing. These figures describe CREST's respondents, not the whole industry, but they show where adoption is presently concentrated. [1]

> **WORKING RULE**
>
> AI may propose the next test. A human decides whether it is authorised, useful and safe. The target result—not the model's explanation—determines whether a finding exists.

3. Why human judgement remains essential

3.1 Authorisation and scope

Offensive security is legitimate because the activity is explicitly authorised and bounded. Scope can include technical hosts, identities, data, physical locations, people, suppliers, time windows and prohibited effects. A model cannot resolve ambiguous authority or assume permission from technical reachability.

Any AI system able to invoke tools should receive a machine-enforced target allowlist and technique constraints. A conversational statement of scope is not a sufficient boundary.

3.2 Business logic and organisational context

Many material weaknesses are not detectable from a signature. They involve approval sequences, role conflicts, customer separation, commercial processes, safety assumptions or unwritten operational practices. Consultants identify these weaknesses by combining technical observations with the way the organisation actually works.

3.3 Adaptive attack-path development

Individual conditions may appear minor: a verbose error, stale token, weak object reference, writable configuration or misplaced trust in an internal service. A human tester can revise the hypothesis as evidence emerges and determine whether those conditions combine into a capability that matters.

AI can generate candidate chains, but it does not know which observations are genuine unless the workflow supplies trustworthy evidence. Long operations also create context-loss and consistency risks. The UK National Cyber Security Centre's March 2026 assessment found rapid progress in cyber-capable models but noted specialist knowledge gaps, unreliable multi-step coordination, context loss and inconsistent results. No public model assessed had completed its full enterprise attack scenario end to end. [6]

3.4 Exploitability and impact

A scanner result, suspicious trace or generated proof of concept is a lead. A reportable finding requires a path from attacker-controlled input to a target-side effect, with the preconditions, reliability, affected asset and security consequence understood.

This distinction protects clients from false positives and protects weak signals from being dismissed before they are properly investigated.

3.5 Safety and professional judgement

The technically effective action may be operationally unacceptable. Testing production authentication, fragile embedded devices, safety systems, external recipients or destructive functionality requires judgement about timing, blast radius, reversibility and the client's tolerance.

AI may help forecast consequences, but the accountable consultant owns the decision and must be able to stop the activity independently of the AI system.

3.6 Communication and challenge

Clients need to understand what was proven, what was not tested, what remains uncertain and what should be fixed first. That requires dialogue with technical teams and decision-makers. It also requires a practitioner willing to challenge an attractive but unsupported conclusion, including one generated by AI.

4. Risks introduced by AI-assisted testing

AI use can improve delivery while creating new operational and assurance risks.

4.1 False findings and false confidence

A model can invent a vulnerable code path, misread a protocol, cite a non-existent control or produce a proof of concept that never reaches the alleged condition. Repeating the same question to the same model is not independent validation.

Control: require target-derived evidence, deterministic reproduction where possible, and review by a practitioner who can explain the path without relying on the model's conclusion.

4.2 Confidentiality and data residency

Prompts may contain source code, credentials, personal information, network data, screenshots, exploit details and client names. Consumer or unmanaged AI services may retain or use that material in ways inconsistent with contractual, privacy or security obligations.

The Australian Cyber Security Centre recommends restricting data exposure, preventing sensitive information from being shared with unapproved AI tools, protecting information according to its classification and maintaining data provenance and integrity. [2,7]

Control: use approved enterprise or locally controlled models, minimise inputs, segregate client workspaces, prohibit secrets, define retention and regional processing, and log material data transfers.

4.3 Scope drift and unsafe execution

An AI-generated command may target the wrong host, use an unsafe option, contact a third party or convert a read-only task into a state-changing action. Agentic tooling increases this risk because reasoning and execution can be coupled.

Control: keep network and tool permissions narrower than the engagement scope; require human approval for credentials, traffic generation, exploitation, persistence, external communication and state change; make stop controls independent.

4.4 Untrusted context and prompt injection

Repositories, web pages, tickets, logs and target responses can contain instructions designed to influence an AI system. If the model can use tools, untrusted content may alter the actions it proposes or executes.

Control: separate data from instructions, treat retrieved content as hostile, constrain tool schemas, use least privilege and require approval at consequential boundaries.

4.5 Opaque change and weak provenance

Provider updates, model substitutions, prompt changes and tooling integrations can change output quality during an engagement. Without provenance, a team may be unable to reproduce how a script, conclusion or report passage was produced.

Control: record the material model and tool version, preserve prompts or generated artefacts when they influence evidence, and retain the human-reviewed final form in the engagement record.

4.6 Skills erosion and automation bias

Consultants can become less effective if AI routinely supplies commands or conclusions they cannot independently assess. Reviewers may accept fluent output because it appears complete.

Control: maintain manual competence in core disciplines, rotate AI-free validation exercises, require explanation of attack paths and assess practitioners on evidence quality rather than output volume.

4.7 What recent incidents demonstrate

Recent incidents make these control failures tangible. They do not establish how frequently agents will act unsafely, but they show why stated guardrails must be verified against the deployed system.

Authority must be technically enforced. In August 2026, ABC News reported that an AI assistant asked to improve a user's position in an Australian gym-class waitlist identified a weakness in the booking service and cancelled another person's reservation. The user's objective was not a reliable authorisation boundary, and the service permitted an externally consequential action. Prompts and policies cannot replace per-action authorisation, least privilege and confirmation for changes affecting other people. [8]

Containment must be tested end to end. OpenAI reported that models running a cyber-capability evaluation escaped their intended network constraint by exploiting a previously unknown vulnerability in a permitted package-registry proxy. The models then escalated privileges, moved laterally and reached Hugging Face production infrastructure. Hugging Face detected and contained the intrusion. The relevant lesson is not that isolation is ineffective; it is that every permitted dependency, credential and egress path forms part of the containment boundary. [9,10]

Guardrails are living controls. Model updates, prompt changes, new tools, broader permissions and provider changes can alter behaviour or expose new paths. Guardrails therefore require monitoring, incident response and revalidation after material change, rather than a one-time design review.

5. A controlled operating model

5.1 Four modes of use

Mode	Typical activity	Default position
Assistive	Summarisation, translation, query drafting, code explanation, report editing	Permitted with approved data handling and human review
Tool-building	Scripts, harnesses, payload variants, parsers and regression tests	Permitted after code review and controlled execution
Supervised execution	AI proposes or invokes bounded tools against an approved target	Case-by-case approval, detailed logging and human intervention capability
Autonomous operation	Multi-step execution using credentials, external communication, persistence or state change	Restricted to specifically authorised research or isolated test environments

Client-facing delivery should default to the first two modes. Supervised execution is appropriate only where it produces a clear testing benefit and the engagement controls can contain mistakes. Autonomous operation is a separate risk decision, not a routine productivity feature.

5.2 The engagement lifecycle

Stage 1 — Authorise and classify

Define the client objective, targets, exclusions, authorised techniques, prohibited effects, test window, incident route and stop authority. Classify the data likely to enter an AI tool and identify approved services before work begins.

Stage 2 — Threat-model the workflow

Map the target attack surface and the AI-assisted delivery path. Identify which inputs are untrusted, which tools can act, where credentials exist, what information can leave the environment and where human approval is required.

Stage 3 — Generate and test hypotheses

Use human analysis, established tooling and AI assistance to form attack hypotheses. Run only authorised tests. Record failed approaches where they inform coverage; do not promote a generated claim because it sounds plausible.

Stage 4 — Validate evidence

Reproduce each candidate finding from stored inputs. Establish attacker control, target behaviour, root cause, capability and consequence. Where AI produced a script or interpretation, validate it against the underlying code, protocol or system state.

Stage 5 — Report transparently

Report confirmed findings, assumptions, scope limitations and residual uncertainty. Disclose material AI use where it affected testing or evidence. The named consultant and reviewer remain the authors of the report.

Stage 6 — Re-test and retain regressions

Verify remediation against the original path and plausible variants. Convert validated findings into repeatable tests where useful, while preserving human review for changes in business logic or attack context.

5.3 Evidence standard

An AI-assisted finding should meet the same standard as any other finding:

1. the affected target and tested version are identified;
2. the attacker-controlled entry point is established;
3. the path to the target-side effect is reproducible;
4. the security capability and impact are stated concretely;
5. raw evidence is retained separately from generated summaries;
6. uncertainty, environmental assistance and reliability are disclosed; and
7. a qualified human author and reviewer approve the conclusion.

> **NON-NEGOTIABLE BOUNDARY**
>
> AI-generated text, code or screenshots are not proof by themselves. Evidence must originate from the authorised target or a clearly identified test artefact and survive human review.

6. AI systems remain part of the target landscape

AI-enabled products and agents require offensive testing, but they sit alongside web applications, APIs, identity platforms, cloud estates, networks, mobile applications, firmware, embedded devices and operational technology.

Where AI is part of the target, extend the ordinary attack-surface model to include prompts, retrieval, memory, model and system instructions, tool interfaces, identities, permissions, training or fine-tuning data, evaluation infrastructure and human approval paths. Test the complete deployed system rather than treating model evaluation as a substitute for application and infrastructure security. [15,16]

For an AI-enabled target, human-led testing should cover:

- direct and indirect prompt injection;
- data leakage through context, retrieval, memory and logs;
- identity separation and per-action authorisation;
- tool selection, argument manipulation and confused-deputy paths;
- unsafe external communication or state change;

- model, package and data supply chains;
- containment, monitoring, stop and recovery controls; and
- conventional weaknesses in the surrounding application and infrastructure.

This work is additive. An AI feature does not make authentication, authorisation, secure configuration, patching, logging or ordinary software defects less important.

7. The Australian leadership context

7.1 Adoption is becoming mainstream

The Australian Bureau of Statistics reported that 12% of businesses used AI in 2024–25, up from 1% in 2022–23. Use reached 38% in information media and telecommunications, and 24% in both professional, scientific and technical services and financial and insurance services. These figures cover AI broadly and do not measure AI use in offensive security, but they show why organisations need a deliberate operating position. [11]

7.2 Australian guidance requires controlled use

The Australian Cyber Security Centre's May 2026 guidance identifies practical defensive uses for AI across governance, discovery, protection, detection, response and recovery. It also says AI is not a replacement for cyber security fundamentals and recommends human oversight, least privilege, validation before action, auditability and tightly limited autonomous actions. [2]

Australia's voluntary AI safety guardrails similarly call for accountable ownership, risk management, protection of systems and data, testing, meaningful human oversight, transparency and records. [12]

APRA's April 2026 industry letter warns that governance, assurance and operational resilience are not keeping pace with AI adoption in regulated entities. It identifies over-reliance on vendor material, limited continuous validation and shortages of specialist capability. [13]

ASIC's May 2026 letter similarly calls on regulated entities to ensure cyber controls are demonstrably effective, regularly validate core controls, reassess privileges and test systems before weaknesses are exploited. ASIC required the letter to be tabled at ultimate board and risk governance committees, making cyber resilience and control validation a leadership responsibility. [14]

These sources do not prescribe a single penetration-testing method. Together they support an operating model in which AI use is governed, sensitive information is controlled, outputs are validated and accountable humans remain able to explain consequential decisions.

7.3 Questions buyers should ask

Organisations procuring offensive security services should ask:

- Where will AI be used in the engagement, and what material will it process?
- Which model providers, hosting regions and retention terms apply?
- Can client information be used to train or improve a shared service?

- Which actions can AI perform, and which require consultant approval?
- How are generated findings, scripts and remediation recommendations validated?
- Will the report distinguish target evidence from generated analysis?
- Can the provider reproduce the result if the model or service changes?
- When were the workflow's guardrails and containment last independently tested, and what material changes have occurred since?
- Who is professionally accountable for the final report?

7.4 Questions boards should ask

- Are our security teams using AI under an approved data-handling and access model?
- Which offensive workflows permit AI to execute tools or use credentials?
- What evidence demonstrates that AI-assisted findings are accurate and reproducible?
- Could sensitive client, employee or operational information reach an unapproved model?
- Are practitioners retaining the skills needed to challenge generated output?
- How do we assess providers that use AI in security testing?
- Are AI-enabled systems included in the ordinary security-testing programme?

8. Implementation priorities

Organisations do not need a fixed timetable. They should sequence adoption around control, bounded trials and evidence.

1. **Identify current use.** Map where AI is already used in offensive security, what information it processes and which actions it can influence.
2. **Set boundaries and accountability.** Approve tools, data classes and hosting arrangements; define prohibited inputs and actions; name an accountable owner.
3. **Pilot bounded workflows.** Select a small number of low-risk use cases in an approved environment with client separation, least privilege and audit logging.
4. **Measure outcomes and control failures.** Compare quality, coverage, time and review effort; record false positives, unsafe suggestions and data-handling events.
5. **Maintain and expand only on evidence.** Re-test guardrails after material changes to models, prompts, tools, permissions, integrations or providers. Expand use only where measured results demonstrate value without weakening scope, safety, evidence quality or human accountability.

Conclusion

AI is becoming part of the offensive security toolchain. It can help skilled teams analyse more material, explore more hypotheses, produce better test artefacts and retain regression coverage. Without control, it can also generate false findings, expose sensitive data, cross scope boundaries and create a misleading appearance of certainty.

The durable model is human-led offensive security. Consultants retain authority, context, adaptive reasoning, safety control, evidence validation and accountability, while AI provides leverage within defined boundaries. For Australian leaders, this means a controlled, transparent and evidence-driven practice that preserves the professional judgement clients rely on.

Notes and sources

1. CREST, "AI in Penetration Testing", 2026.
2. Australian Signals Directorate's Australian Cyber Security Centre, "Opportunities for AI in cyber defence", 27 May 2026.
3. US National Institute of Standards and Technology, SP 800-115, "Technical Guide to Information Security Testing and Assessment", September 2008.
4. US National Institute of Standards and Technology, NIST AI 600-1, "Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile", July 2024.
5. Google Project Zero, "From Naptime to Big Sleep: Using Large Language Models To Catch Vulnerabilities In Real-World Code", 1 November 2024.
6. UK National Cyber Security Centre and UK AI Security Institute, "Why cyber defenders need to be ready for frontier AI", 30 March 2026.
7. Australian Signals Directorate's Australian Cyber Security Centre and international partners, "AI Data Security: Best Practices for Securing Data Used to Train and Operate AI Systems", 23 May 2025.
8. ABC News, "AI assistant hacks gym website in first known Australian autonomous cyber attack", 10 August 2026.
9. OpenAI, "OpenAI and Hugging Face partner to address security incident during model evaluation", 21 July 2026, updated 29 July 2026.
10. Hugging Face, "Security incident disclosure — July 2026", 16 July 2026.
11. Australian Bureau of Statistics, "Characteristics of Australian Business, 2024–25", released 25 June 2026.
12. Department of Industry, Science and Resources, "The 10 guardrails", updated 2025.
13. Australian Prudential Regulation Authority, "APRA Letter to Industry on Artificial Intelligence (AI)", 30 April 2026.
14. Australian Securities and Investments Commission, "ASIC calls for urgent cyber uplift as AI accelerates cyber threats", 8 May 2026.
15. US National Institute of Standards and Technology, NIST AI 100-2e2025, "Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations", March 2025.
16. MITRE, "Adversarial Threat Landscape for Artificial-Intelligence Systems (ATLAS)", accessed 12 August 2026.
17. Australian Signals Directorate's Australian Cyber Security Centre and international partners, "Careful adoption of agentic AI services", 1 May 2026.