WHITE PAPER

HUMAN-LED PENETRATION TESTING IN AN AI-DRIVEN THREAT LANDSCAPE

EXPOSING REAL-WORLD VULNERABILITIES THAT AI TOOLS MISS

Abstract

Generative AI is shifting the balance in cybersecurity. Attackers now use it to automate reconnaissance, generate convincing phishing, and chain misconfigurations that span cloud, identity, and SaaS environments. Defenders are also adopting AI, but most tools still struggle to understand business context, intent, and operational impact.

This whitepaper explores why human-led, AI-supported penetration testing is still the most effective way to uncover what's truly at risk. This whitepaper demonstrates how AI-driven threats behave in the real world, where automated tools still fall short, and why expert testers remain essential in an increasingly autonomous threat landscape.

Red teaming (simulated attacks) and kill chain analysis (mapping an attacker's path to impact) are among the methods examined, showing how AI changes both offensive and defensive strategies.

themissinglink®
An Infosys company

Infosys®
Navigate your next

Overview and industry problem

1. The rise of AI-powered adversaries

AI has become instrumental in offensive cyber operations, moving beyond support roles and acting as a primary operator. Threat intelligence indicates that malicious actors now embed AI tools into reconnaissance, exploitation, lateral movement, credential harvesting, and extortion[1]. In a major 2025 cyber espionage incident, a state-sponsored group used AI agents to autonomously execute approximately 80 to 90 percent of its intrusion activity[2]. This marked a turning point where AI shifted from a support tool to a primary operator.

Criminal groups have adopted AI to write malware, automate fraud, analyse stolen data, and craft highly targeted phishing. Reports show that AI significantly lowers the skill threshold required to conduct advanced cyberattacks. Individuals who previously lacked the expertise needed for complex intrusions can now achieve them using AI-driven tooling[3]. This shift is broadening the pool of threat actors and accelerating the development of more advanced and accessible offensive toolkits.

2. How attackers exploit AI capabilities

AI enables broad reconnaissance by analysing exposed assets, cloud resources, and identity footprints at speed[3]. Machine learning techniques generate constantly evolving malware that can avoid signature-based detection and be quickly written and adapted.. Attackers also use AI to chain misconfigurations across identity systems, cloud platforms, and SaaS applications. These chains create attack paths that traditional scanners are unable to identify.

AI has also changed the nature of social engineering. Large language models (LLMs) produce highly contextualised messages that mirror an organisation's internal tone. Deepfake technologies allow attackers to convincingly impersonate executives or staff members. Threat actors are also manipulating LLMs directly, using prompt injection, interface abuse, and API exploitation to extract sensitive information or trigger unintended behaviours[1].

Anthropic identifies these tactics as growing concerns in publicly accessible systems, particularly where model alignment controls are weak or absent[1].

At the same time, defensive tools powered by AI are improving visibility, but often lack contextual understanding. They struggle to interpret cross-system relationships, business workflows, and human behaviour. ISACA notes that malicious AI-enabled behaviour increasingly resembles legitimate activity, making it more difficult to distinguish threat signals[4].

3. Escalating risk and business impact

Global cybercrime is projected to exceed 10.5 trillion US dollars annually by 2025[5]. AI amplifies this trend by enabling attackers to move faster and scale operations with greater precision. Nation-state groups and organised cybercriminal operations now combine AI with traditional tradecraft to form hybrid threats that adapt quickly and target organisations more effectively. Smaller organisations face increased exposure due to limited resources and slower response capabilities.

As a result, defenders must combine AI-enabled tools with human expertise to understand real-world exploitability and maintain resilience.

Recommended solution: Human-led, AI-supported penetration testing

Traditional testing methods struggle to keep pace with AI-enabled threats. AI helps process large datasets and identify surface-level patterns, speeding up early-stage testing. Human testers bring the context. They determine exploitability, chain weaknesses across systems, and assess business impact.

In early-stage testing, AI can summarise large datasets, highlight anomalies, and propose exploratory paths. These outputs help human testers focus where it matters most. But AI cannot interpret how misconfigurations interact or assess workflows in context or how to chain multiple low risk issues in complex ways. Human expertise is still essential to confirm findings, simulate adversarial intent, and reveal meaningful chains of attack.

Red and purple team exercises increasingly incorporate AI. Red teams use it to generate phishing content or identify escalation paths. Purple teams apply AI-derived insights to refine detection and response. In both cases, human oversight ensures testing remains realistic, safe, and relevant.

Human-led testing is indispensable because it examines vulnerabilities within a real-world operational context. Automation may reveal isolated issues, but only skilled testers can show how those issues combine and what an attacker could achieve. This insight turns technical findings into strategic risk awareness.

This approach helps organisations understand how attackers could move through their environment and how that movement translates into real business risk.

AI-enabled attack scenarios

Understanding AI threats means going beyond theory. The following scenarios show how AI-enabled attackers exploit security gaps across cloud, identity, SaaS, and human workflows, often bypassing even mature controls. Each example is drawn from simulated or observed activity and demonstrates how even mature security controls can be bypassed when AI is used to scale, automate or deceive. These are the kinds of attack paths that only human-led testing, with AI support, can fully uncover.

1. Deepfake voice leading to MFA bypass

Context:

A global financial services organisation with strict multi-factor authentication requirements was targeted during reconnaissance. When initial password spraying and phishing attempts failed, the attacker pivoted to social engineering using AI-generated voice impersonation.

Approach:

The attacker created a deepfake of a senior manager's voice using public recordings[6 7]. Guided by an AI-generated script, the attacker phoned the helpdesk and described a fabricated token synchronisation issue. Lacking rigorous verification procedures, the helpdesk approved the MFA reset. The attacker also carried out an SMS interception tactic by spoofing a telecom provider and coaching the victim through the process.

Outcome:

The attacker enrolled a new MFA device and accessed internal systems. This foothold enabled further reconnaissance, credential collection, and access to sensitive repositories. The incident demonstrated that strong MFA can be bypassed when voice verification processes are not hardened.

2. Chained misconfiguration across Azure AD and Microsoft 365

Context:

A regional utilities provider managed an Azure Active Directory environment with legacy Conditional Access policies and weak role separation. Several stale roles included Global Reader access combined with Exchange administrative privileges.

Approach:

Attackers used AI-enhanced tools to map group memberships and identify a service account with excessive privileges[8 9]. They exploited a misconfigured refresh token policy that failed to enforce MFA for static IP addresses. The attacker then abused OAuth consent flows, tricking the tenant into granting access to an application that forwarded tokens back to the attacker.

Outcome:

The attacker exported email and SharePoint data while evading DLP controls. Detection was delayed because the activity resembled legitimate administrator behaviour. The incident showed how minor misconfigurations can be chained into high-impact compromises without phishing or brute force[8 9].

3. LLM prompt injection leading to sensitive data exposure

Context:

An insurer deployed an internal LLM-powered claims assistant connected to backend data systems containing customer and financial information.

Approach:

Testers crafted prompts containing hidden instructions to bypass filtering[10 11]. These instructions caused the model to query unauthorised backend APIs. For example, a benign request for policy information included concealed instructions that retrieved partial customer identifiers.

Outcome:

The system returned sensitive data, highlighting significant weaknesses in prompt filtering, backend permissions, and LLM integration controls.

Comparing AI-enabled attack and defence models

Attackers and defenders may both use AI, but their methods and outcomes look very different. The following diagrams show the two contrasting approaches. The first outlines how threat actors are using AI to automate reconnaissance, exploitation, and decision-making at scale. The second shows how human-led penetration testing integrates AI tools while keeping strategy, context, and business impact at the core. Together, they provide a clear view of how cyber operations are evolving and why human oversight is still essential.

4.1 Simplified autonomous attacker model

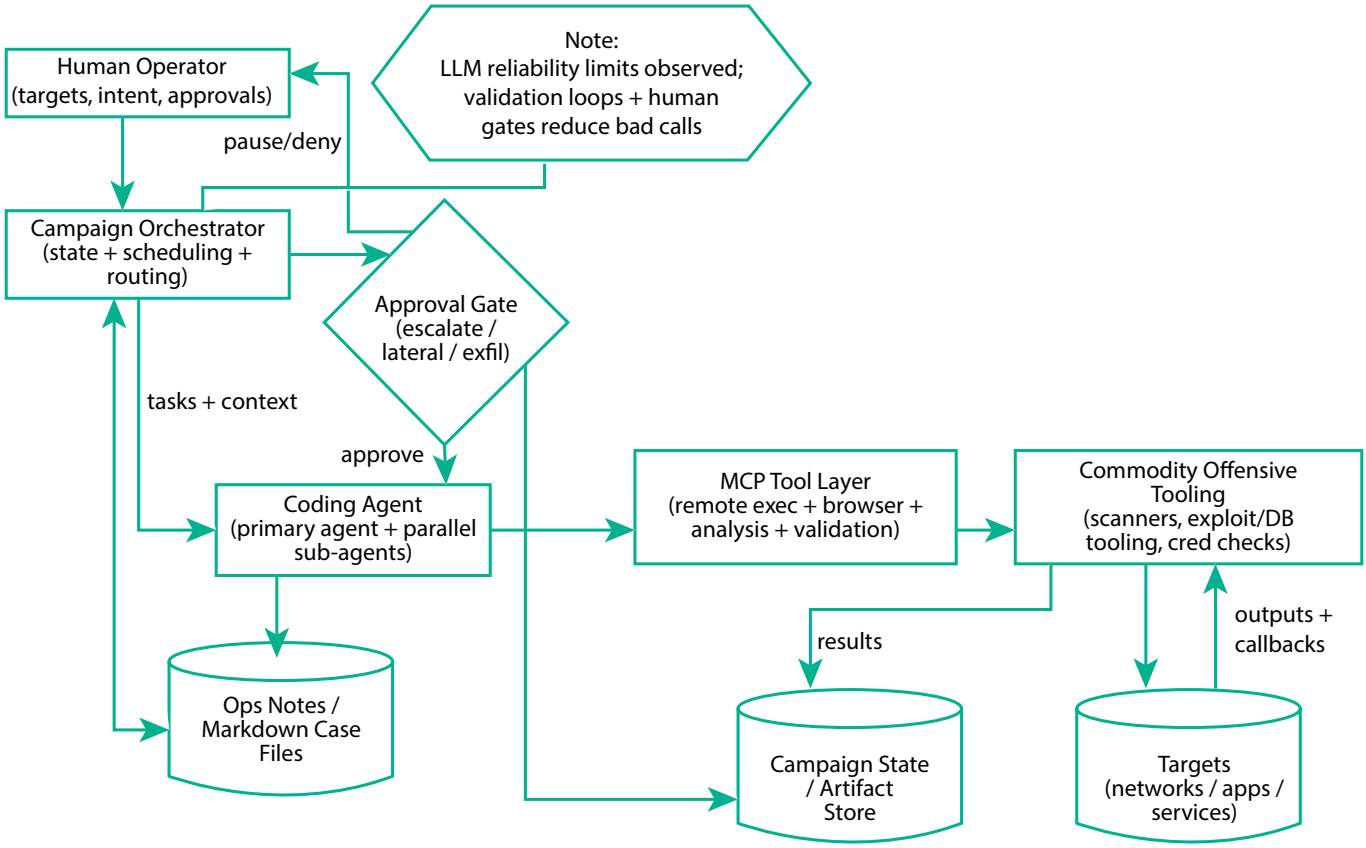


Figure 1: *A conceptual view highlighting orchestration, approval gates, and execution layers within an AI-assisted attack workflow.*

This simplified model illustrates how human intent, initial targeting, or campaign objectives flow into an AI-led orchestration engine. While the AI performs most tactical activities, an approval gate shows how attackers may occasionally supervise, escalate, or adjust AI-driven actions. This hybrid structure blends human decision making with AI efficiency.

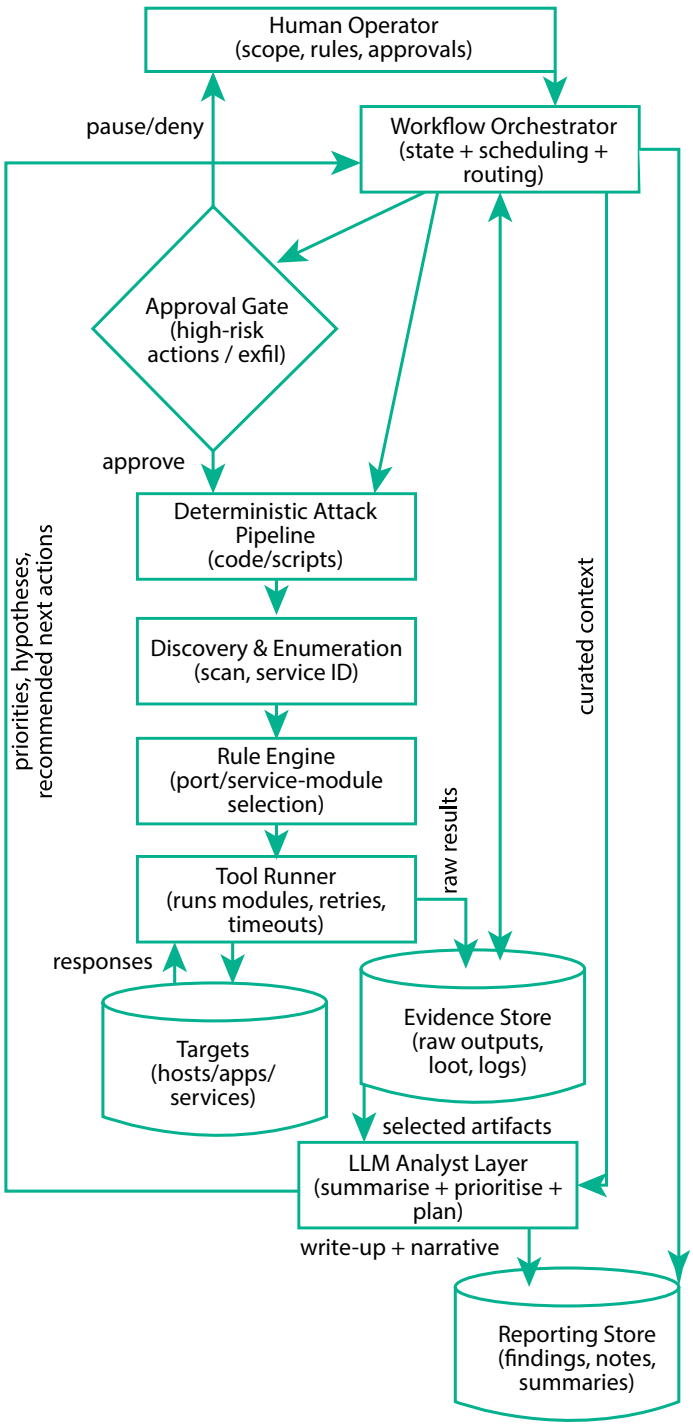
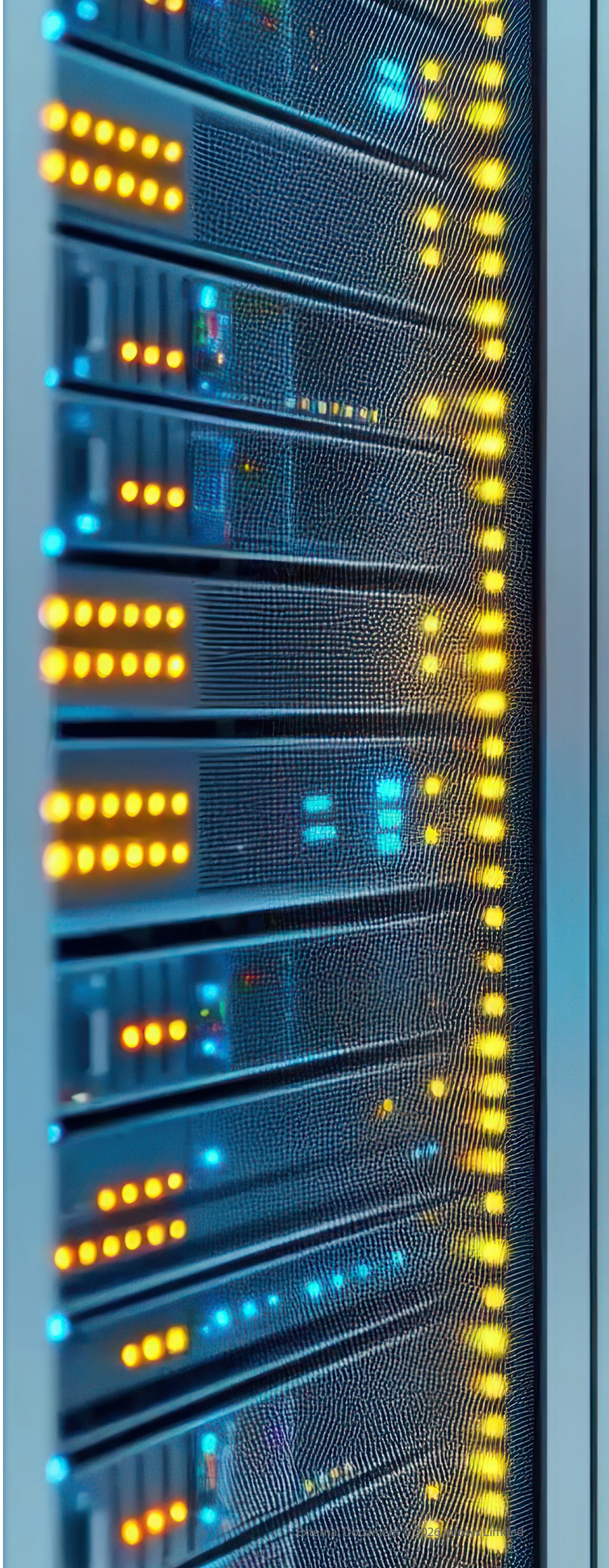


Figure 2: *A defensive architecture showing how human testers leverage automation and AI-assisted workflows to simulate attacker behaviour and evaluate complex attack surfaces.*

This model represents the defensive counterpart to autonomous attacker workflows. Human testers drive strategy, oversight, and interpretation. An orchestrator assists with scheduling and routing. Automated modules accelerate reconnaissance and evidence collection. Rule engines and analysis layers help process findings. This architecture ensures human judgement remains central while AI increases efficiency and coverage.



Best practices for AI-aware penetration testing

AI-driven threats are not just faster. They are more layered, more convincing, and can be harder to detect. These attacks combine realistic social engineering with automated decision-making, dynamic payloads, and behaviours that closely resemble legitimate activity. As a result, traditional detection tools that depend on static rules, known signatures, or predictable patterns often struggle to identify and stop them effectively[12]. Traditional penetration testing is effective at identifying complex attack paths; however, AI-aware testing enhances this by rapidly uncovering multi-layered attack vectors across cloud, identity, and AI-enabled systems, while reducing both the time and cost required to achieve comparable depth. It also addresses emerging threats like prompt injection and AI misuse that conventional methods often miss. To remain effective, testing practices need to evolve. This section outlines how to adapt your approach to reflect current attack behaviours, from simulating deepfake-enabled social engineering to assessing prompt injection risks, and aligning outcomes with frameworks such as MITRE ATT&CK, informed by controls like OWASP's Top 10 for LLM Applications.

Evolving your testing methodology

Each of the following recommendations represents a key practice in adapting penetration testing to the AI-driven threat landscape:

1. Use AI for speed, not strategy

AI accelerates tasks such as enumeration and log analysis, but lacks contextual understanding. Human testers must validate AI findings, determine exploitability, and interpret meaning within the operational environment[1].

Table 1. Comparison of automated AI tools and human-led penetration testing

Capability	Automated AI tools	Human-led penetration testing
Analysis quality	Fast pattern recognition	Contextual and creative reasoning
Vulnerability discovery	Identifies known weaknesses	Reveals chained and business logic issues
Social engineering	Limited simulation capability	Realistic interaction and impersonation
Misconfiguration chaining	Limited cross-system insight	End-to-end chaining across identity, cloud, SaaS
Business impact	Not supported	Clear operational and strategic risk interpretation

2. Assess AI-driven attack vectors directly

LLM-integrated systems require specialised adversarial testing. Assessments must examine prompt injection, data leakage, model manipulation, and excessive backend permissions.[10]

3. Simulate AI-enabled social engineering

Deepfake impersonation, contextual phishing, and adaptive conversational fraud should be replicated to test human and procedural resilience[6].

4. Evaluate full kill-chain scenarios in business context

AI-enabled attackers often exploit subtle misalignments across identity, cloud, and SaaS systems. Testing should simulate full attack sequences that reflect how these gaps are linked in real-world breaches[3].

5. Align findings with recognised frameworks

Mapping results to MITRE ATT&CK, NIST CSF, OWASP LLM Top 10, and ISO 27001 provides clarity and establishes governance.

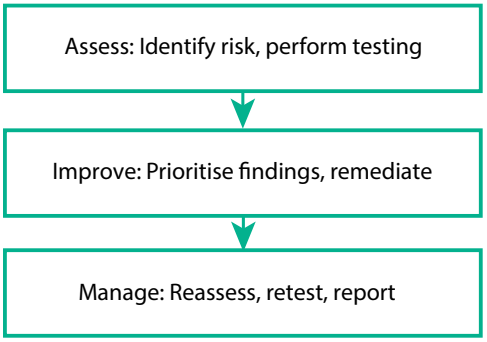
The table below summarises how AI-aware penetration testing aligns with key frameworks, helping to support governance, compliance, and risk visibility.

Table 2. Framework alignment summary

Framework/ Guidance	Focus area	Penetration testing insight
MITRE ATT&CK	Adversary behaviours	Coverage of tested attack techniques
OWASP LLM Top 10	AI and LLM-specific risks	Identification of prompt injection and misuse
NIST CSF	Control maturity	Evaluation of PR.AC and detection capabilities
ISO 27001	Governance and controls	Evidence supporting compliance and risk management

6. Integrate results into continuous improvement

AI-aware penetration testing must support a continuous cycle of Assess, Improve, and Manage. Scenario-based exercises and periodic retesting strengthen resilience.



Strategic takeaways

Security leaders now face adversaries who operate with speed and adaptability that were once thought impossible. AI enables attackers to automate reconnaissance, chain misconfigurations, and perform social engineering at scale. Defending against these threats requires a strategic shift in how resilience is assessed.

Human-led, AI-supported penetration testing shows how adversaries exploit identity drift, cloud weaknesses, procedural gaps, and LLM vulnerabilities. This contextual view enables leaders to prioritise investments, strengthen governance, and align remediation with operational risk.

Automation increases efficiency but cannot replace human judgement. Organisations that adopt AI-aware penetration testing and embed it into continuous improvement frameworks will be best positioned to detect, contain, and recover from AI-enabled attacks.

Conclusion

AI is advancing both offensive and defensive cybersecurity. Attackers use it to scale reconnaissance, automate exploitation, and craft tailored social engineering campaigns. Defenders are deploying AI tools to increase speed and visibility, but automation alone cannot assess how threats unfold across complex systems.

Human-led, AI-supported penetration testing remains essential. It brings the context, intent, and adversarial perspective that automated tools cannot deliver. This approach helps uncover meaningful vulnerabilities, test defensive coverage, and evaluate the real impact of an attack path.

Automation can highlight what is exposed. Human expertise determines what can be compromised.

Organisations that integrate AI-aware penetration testing into their security strategy are better positioned to understand how threats evolve, and how to respond before real damage is done.

References

1. Anthropic. (2025, August). Detecting and countering misuse of AI. https://www.anthropic.com/news/detecting-and-countering-misuse-of-ai
2. Anthropic. (2025). Disrupting the first reported AI-orchestrated cyber espionage campaign. https://assets.anthropic.com/m/ec212e6566a0d47/original/Disrupting-the-first-reported-AI-orchestrated-cyber-espionage-campaign.pdf
3. ISACA. (2025). State of Cybersecurity 2025. https://www.isaca.org/resources/state-of-cybersecurity-2025
4. Cybersecurity Ventures. (2025). 2025 cybercrime report. https://cybersecurityventures.com/2025-cybercrime-report
5. Delinea. (2023, November 28). AI-driven attacks to become faster, more sophisticated and harder to detect by 2026. Australian Cyber Security Magazine. https://australiancybersecuritymagazine.com.au/ai-driven-attacks-to-become-faster-more-sophisticated-and-harder-to-detect-by-2026
6. Group-IB. (2023). The anatomy of a deepfake voice phishing attack. https://www.group-ib.com/blog/voice-deepfake-scams
7. Goodin, D. (2023, August 4). Here's how deepfake vishing attacks work, and why they can be hard to detect. Ars Technica. https://arstechnica.com/security/2023/08/heres-how-deepfake-vishing-attacks-work-and-why-they-can-be-hard-to-detect/
8. Microsoft Security Blog. (2023, July 14). Analysis of Storm-0558 techniques for unauthorized email access. Retrieved from https://www.microsoft.com/en-us/security/blog/2023/07/14/analysis-of-storm-0558-techniques-for-unauthorized-email-access/
9. Wiz Research. (2023). Storm-0558 Update: Takeaways from Microsoft's Latest Report. Retrieved from https://www.wiz.io/blog/key-takeaways-from-microsofts-latest-storm-0558-report
10. Gulyamov, S., Rodionov, A., Khursanov, R., & Zhou, Y. (2026). Prompt injection attacks in large language models and AI agent systems: A comprehensive review of vulnerabilities, attack vectors, and defense mechanisms. Information, 17(1), Article 54. https://doi.org/10.3390/info17010054
11. OWASP Foundation. (2024). OWASP Top 10 for LLM applications. https://genai.owasp.org
12. Uddin, M., Islam, M. R., & Rahman, M. M. (2025). The generative AI revolution in cybersecurity. Springer. https://link.springer.com/book/10.1007/978-981-99-1350-1
13. MITRE. (2023). ATT&CK framework (Version 12). https://attack.mitre.org

Further Reading

Marchal, N., Lee, J., & Yu, C. (2024). Generative AI misuse: A taxonomy of observed threat activity. arXiv. https://arxiv.org/abs/2405.10340

Guembea, B., Alvarez, R., & Lam, S. (2022). AI-driven cyber-attacks: Emerging threat patterns. Cybersecurity Journal, 4(3), 122–136.

Adewale, T., Hassan, M., & Patel, A. (2024). AI-driven cyber-attacks and defense tactics. ResearchGate. https://www.researchgate.net/publication/AI-Cyber-Defense

MITRE. (2024). ATLAS: Adversarial Threat Landscape for Artificial Intelligence Systems. https://atlas.mitre.org

NIST. (2024). Secure deployment of AI systems: Guidance for managing AI-specific risks (NIST AI 600-1). National Institute of Standards and Technology. https://www.nist.gov/publications/nist-ai-600-1

NIST. (2024). Adversarial machine learning: Taxonomy and terminology (NISTIR 8332). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.IR.8332

Google DeepMind. (2024). Frontier AI threat modelling report. https://www.deepmind.com/research/frontier-ai-threat-modelling

UK National Cyber Security Centre. (2024). Safety and security risks of large language models. https://www.ncsc.gov.uk/whitepaper/large-language-models-guidance

SANS Institute. (2024). Adversarial AI and red teaming for LLM systems. https://www.sans.org/white-papers/adversarial-ai-llm-red-teaming

ENISA. (2023). Threat landscape for artificial intelligence. https://www.enisa.europa.eu/publications/artificial-intelligence-threat-landscape

Microsoft. (2024). Generative AI security framework. https://www.microsoft.com/security/blog/2024/03/21/generative-ai-security-framework

Valencia, L. J. (2024). Artificial intelligence as the new hacker: Offensive agents for cyber operations. arXiv. https://arxiv.org/abs/2404.09512

ScienceDirect. (2023). Artificial intelligence for cybersecurity: Literature review. https://www.sciencedirect.com/science/article/pii/S1877050923000864

About the Author

Matt Dobinson

Head of Security Consulting, The Missing Link

Matt leads The Missing Link's penetration testing, red teaming, and adversary simulation services. His experience covers AI-enabled offensive tooling, identity exploitation, cloud compromise simulation, and LLM security assessments. He is a CREST Australasia Advisory Board member and a frequent presenter on emerging attack techniques and modern testing methodologies. He holds industry-recognized certifications such as OSCP, OSCE, OSWP, SANS SEC564, CREST CRPT, and CPSA.

Contact The Missing Link

E-mail: contactus@themissinglink.com.au

Website: www.themissinglink.com.au

For more information, contact askus@infosys.com

Infosys®
Navigate your next