WHITE PAPER

Agentic AI: The Shift from AI-Assisted to Autonomous Cyber Attacks & Defense

How autonomous AI systems are rewriting the rules of cybersecurity - and what security leaders must do now

A Digital Resilience Threat Intelligence Report

Digital Resilience | www.digitalresilience.co.za | April 2025

Target Audience: CISOs, Security Architects, AI Researchers

EXECUTIVE SUMMARY

Artificial intelligence has crossed a threshold. No longer confined to assistive roles - flagging anomalies, triaging alerts, or accelerating analyst workflows - AI systems are now capable of acting autonomously across complex, multi-step objectives with minimal human oversight. This category of system, known as **Agentic AI**, represents a categorical shift in the cybersecurity threat landscape: one that demands an equally categorical response from defenders.

This white paper examines three dimensions of that shift. First, how agentic AI is being weaponised by threat actors to conduct autonomous reconnaissance, accelerate exploit development, and execute lateral movement at machine speed. Second, how defenders are deploying their own agentic systems to detect and counter these threats - creating an emergent AI-versus-AI arms race in real-time. Third, how organisations must defend the AI systems themselves against a new class of attack: data poisoning, model-stealing, and adversarial manipulation.

The stakes are high. A 2024 Gartner projection estimates that by 2027, 17% of all cyberattacks will involve AI agents acting with meaningful autonomy. IBM's Cost of a Data Breach Report 2024 pegs the average breach cost at USD 4.88 million - a figure that will rise sharply as attack automation accelerates. The organisations that understand and adapt to agentic AI now will hold a decisive advantage over those that treat it as a distant concern.

17% of cyberattacks projected to involve AI agents by 2027 (Gartner)	**USD 4.88M** average cost of a data breach (IBM, 2024)
72 days average time to identify an AI-assisted breach vs. 204 days traditional	**<60 min** time for AI agents to pivot from initial access to domain control

> *"The question is no longer whether AI will be weaponised against your organisation. It already is. The question is whether your defences are autonomous enough to respond."*
>
> **- Yaron Assabi**

WHAT IS AGENTIC AI – AND WHY IT CHANGES EVERYTHING

To understand the threat, we must first understand the technology. Traditional AI security tools - SOAR platforms, ML-based anomaly detectors, signature engines - are reactive and narrowly scoped. They perform specific tasks when triggered by human operators or predefined rules. They do not plan. They do not adapt their strategy. They do not spawn sub-tasks to accomplish a goal.

Agentic AI systems are fundamentally different. Built on large language models (LLMs) or reinforcement-learning frameworks, they can decompose a high-level objective into subtasks, select and use external tools (browsers, APIs, code interpreters, network scanners), maintain memory across sessions, observe the results of their actions, and iterate until the goal is achieved - without a human directing each step.

The Autonomy Spectrum

It is useful to think of AI autonomy as a spectrum, analogous to autonomous vehicle levels. At Level 1, AI suggests actions and a human decides. At Level 3, AI acts within defined boundaries with human oversight available. At Level 5, AI acts independently and notifies humans of outcomes. Current offensive AI agents operate between Levels 3 and 4 in many documented deployments - and the direction of travel is unmistakable.

Autonomy Level	Behaviour	Example in Security Context
Level 1 - Assisted	AI suggests; human decides	Copilot recommending a firewall rule change
Level 2 - Partial	AI acts on narrow tasks with approval	SOAR running an approved playbook
Level 3 - Conditional	AI acts within defined boundaries	EDR auto-isolating an infected endpoint
Level 4 - High	AI acts; human monitors in background	Agent conducting multi-step reconnaissance
Level 5 - Full	AI acts; human notified of outcomes	Fully autonomous attack chain execution

Why the Last Mile Matters

The leap from Level 2 to Level 3 autonomy - from executing a pre-approved playbook to making contextual decisions in real time - is not incremental. It is architectural. An AI agent that can observe its environment, reason about what it finds, and choose a next action is qualitatively more capable than any rule-based automation, regardless of how sophisticated that automation appears.

Modern agent frameworks - including multi-agent pipelines where specialised agents collaborate on complex tasks - have dramatically lowered the barrier to operationalising this capability. A threat actor no longer needs to write a sophisticated exploit from scratch. They need only to configure an agent with a goal and provide it access to the appropriate tools.

> **Early Indicators Already Visible**
> Security researchers at Recorded Future, Google Mandiant, and CrowdStrike have documented LLM-assisted phishing personalisation at industrial scale, AI-driven CVE triage for exploit prioritisation, and automated social engineering content generation indistinguishable from human-authored communications. These are Level 3 behaviours. Level 4 is not a future concern - it is an emerging present reality.

AI-POWERED RECONNAISSANCE AND AUTOMATED EXPLOIT GENERATION

The offensive application of agentic AI is most acutely felt in the early phases of the attack lifecycle - reconnaissance and weaponisation - where the ability to gather intelligence rapidly and convert it into actionable capability has historically been a bottleneck constrained by human effort and expertise.

Autonomous Reconnaissance at Scale

Reconnaissance is the foundation of every targeted attack. Traditionally, it required skilled human operators to patiently map an organisation's attack surface over days or weeks. Agentic AI compresses this timeline to hours and expands its breadth beyond what any human team could achieve.

Autonomous OSINT agents can simultaneously scrape LinkedIn for org-chart data, monitor GitHub for leaked credentials and internal tooling, parse DNS records and certificate transparency logs for infrastructure mapping, analyse job postings for technology stack intelligence, and correlate dark web forum activity for prior breach data - all without triggering traditional detection mechanisms because their traffic patterns mimic legitimate user behaviour.

Beyond passive intelligence gathering, AI agents conduct active network fingerprinting and service enumeration with a level of patience and precision that evades rate-limiting and IDS/IPS signatures. Fine-tuned models trained on breach data can predict password patterns, credential reuse behaviours,

and likely MFA bypass vectors for specific organisations - transforming generic attack capability into organisation-specific targeting.

> *"An AI agent conducting reconnaissance does not get tired, does not make careless mistakes, does not need to be paid, and can operate across dozens of targets simultaneously. The asymmetry this creates for defenders is severe."*
>
> **- Yaron Assabi**

The Exploit Development Acceleration

The window between CVE publication and weaponised exploit has been shrinking for years. Agentic AI threatens to close it entirely. LLM-assisted fuzzing - where AI generates novel input variations to discover vulnerabilities in target software - no longer requires human-written test cases. AI agents can read a CVE description, understand the vulnerable code path, and generate proof-of-concept exploit code in minutes.

More concerning is the emergence of **polymorphic malware 2.0**: AI systems that rewrite their own code signatures after each execution, making signature-based and static detection effectively obsolete for this class of threat. Each instance of the malware is functionally identical but syntactically unique - requiring defenders to shift entirely to behaviour-based detection.

Multi-agent attack architectures represent the most sophisticated current development: separate specialised agents handling initial access, credential harvesting, lateral movement, data exfiltration, and C2 communication - each optimised for their specific role and communicating through encrypted channels. This mirrors how elite human red teams operate, but at machine speed and without the human coordination overhead.

> **Case Study: AI-Accelerated Lateral Movement**
> In a 2024 red team exercise conducted by a major financial services CISO, an AI-augmented attacker achieved domain administrator access from an initial phishing foothold in under 47 minutes - compared to an average of 4.5 hours for a skilled human red team performing the same exercise. The AI agent identified Kerberoastable service accounts, extracted hashes, cracked credentials offline, and pivoted to a domain controller without triggering a single SIEM alert.

Evasion as a First-Class Capability

Agentic AI does not merely execute attacks faster - it executes them more cleverly. Attackers can use AI to query defensive models and predict what behaviours will be flagged, then deliberately stay just below detection thresholds. Alert flooding is another emerging tactic: deliberately generating thousands

of low-confidence signals to saturate defensive AI systems and mask the genuine intrusion within the noise. These techniques represent a direct counter-intelligence capability against AI-powered defences.

WHEN A GOVERNMENT SAFETY INSTITUTE CONFIRMS YOUR THREAT MODEL

Much of the discussion around agentic AI in cybersecurity has, until recently, been forward-looking - projections, capability demonstrations in controlled research environments, and extrapolations from known AI performance benchmarks. That changed materially in 2024 when the **UK AI Safety Institute (AISI)**, a national body established specifically to evaluate the safety properties of frontier AI systems, published its formal evaluation of Claude Mythos Preview's cyber capabilities. The findings are not a warning about what might be possible. They are a documented record of what a frontier AI model can already do - autonomously, rapidly, and at a level of sophistication that compresses professional expertise into minutes.

The AISI is not a vendor marketing a product, a researcher publishing a proof of concept, or a red team demonstrating a novel technique. It is a government agency whose institutional mandate is to provide credible, impartial assessments of AI capability risk to policymakers and the public. When AISI publishes a finding, it carries an evidentiary weight that the security community and boardrooms alike should treat as ground truth.

The Evaluation Finding - Verbatim

> *"In controlled evaluations where Mythos Preview was explicitly directed and given network access to do so, we observed that it could execute multi-stage attacks on vulnerable networks and discover and exploit vulnerabilities autonomously - tasks that would take human professionals days of work."*
>
> **- UK AI Safety Institute (AISI) - Evaluation of Claude Mythos Preview, 2024**

Three phrases in this statement deserve individual unpacking by any security leader reading them:

- **"Multi-stage attacks."** This is not a single exploit or a brute-force credential attack. Multi-stage attack chains - reconnaissance, initial access, privilege escalation, lateral movement, data staging - have historically required a coordinated human team operating over days or weeks. The AISI evaluation documents an AI model executing this full chain autonomously, sequencing each phase without human intervention between steps.
- **"Discover and exploit vulnerabilities autonomously."** Vulnerability discovery - finding novel weaknesses in target systems - is among the highest-skill activities in offensive security. It requires contextual reasoning about system architecture, knowledge of relevant CVEs and vulnerability classes, and the ability to adapt when an expected attack path is blocked. The AISI

finding confirms that frontier AI models can now perform this reasoning autonomously, without a human directing each individual step.

- **"Tasks that would take human professionals days of work."** The compression of timeline is not incremental. Days-to-minutes is a two to three order-of-magnitude acceleration. At that scale, the concept of a human-speed defensive response is structurally inadequate - an attacker using this capability can complete their entire kill chain before a human analyst has finished reading the first alert.

What Mythos Preview Represents in Context

It is important to understand what the AISI evaluation was and was not. Mythos Preview is a frontier commercial AI model - not a specialised offensive security tool, not a nation-state developed cyber weapon, and not an AI system fine-tuned specifically for attack capability. The evaluation tested a general-purpose AI assistant, directed with network access and explicit instruction.

This context cuts in both directions. On one hand, it demonstrates that the barrier to AI-enabled autonomous attack capability is already cleared by commercially available models that any sufficiently motivated actor can access. On the other hand, the evaluation was conducted under controlled conditions with explicit direction - the AI was told to conduct attacks and given the access to do so. Autonomous, unprompted attack initiation by general-purpose AI models is not what AISI documented.

The distinction matters for risk calibration, but it should not produce complacency. The capability gap between a directed frontier AI model and a purpose-built offensive AI agent - fine-tuned on attack data, wrapped in an autonomous planning layer, and deployed by a motivated threat actor - is narrowing rapidly. What AISI documented in 2024 with a general-purpose model is the floor, not the ceiling.

> **The Regulatory Implications**
> The AISI publication is the first instance of a national AI safety regulator formally documenting autonomous cyber attack capability in a publicly available frontier model. This is expected to accelerate regulatory action across multiple jurisdictions. The EU AI Act classifies systems with the potential to conduct cyber attacks as high-risk AI. NIST's AI Risk Management Framework explicitly addresses AI-enabled threat escalation. Security leaders should anticipate that AI cyber capability evaluations will become a mandatory disclosure requirement for AI system vendors operating in regulated sectors - and that the AISI methodology will be adopted or referenced by peer regulators globally.

Implications for Defensive Architecture

The practical implications of the AISI finding for defenders are not abstract. They require a reassessment of assumptions embedded in current security architecture:

Assumption	Pre-AISI Status	Post-AISI Reassessment
Attackers need specialist skill to chain multi-stage attacks	True - required experienced red team operators	False - frontier AI models can chain attack phases autonomously with direction
Days-long dwell time gives defenders detection opportunity	Broadly true - average dwell time 17+ days	No longer reliable - autonomous chains can complete in hours or less
Vulnerability discovery requires human security researcher expertise	True - highly specialised, slow, resource-intensive	Significantly weakened - AISI confirms AI can discover vulns autonomously
Signature and rule-based detection covers known attack patterns	Partially true for known TTPs	Insufficient - AI-generated attack variants evade static signatures by design
Human-speed IR response is adequate	Adequate for human-speed attacks	Structurally inadequate against AI-speed multi-stage execution

The AISI evaluation is a reference document that security leaders should use to make the case for investment in AI-speed defensive controls: behaviour-based EDR with AI-driven anomaly detection, automated network isolation playbooks, immutable backup architecture, and continuous external attack surface monitoring. It provides the external, regulator-grade evidence that many boards require before approving material increases in security spend.

THE AI VS. AI ARMS RACE IN REAL-TIME THREAT DETECTION

The defensive response to agentic offensive AI is, inevitably, agentic defensive AI. Security operations centres are deploying autonomous agents to triage alerts, correlate signals across EDR, SIEM, NDR, and cloud telemetry, and generate remediation playbooks without analyst intervention. This creates an emergent dynamic that security leaders must understand clearly: a real-time arms race between systems that learn from each other's behaviour.

Autonomous Security Operations

Defensive agentic AI manifests most visibly in the evolution of the Security Operations Centre. Traditional SOC workflows - alert triage, correlation, investigation, escalation - are being progressively automated through agents that can reason across multiple data sources simultaneously, maintain

investigation context across long time windows, and escalate to humans only when genuine uncertainty or high-impact decisions arise.

Dynamic honeypot technology is another frontier: AI-powered deception systems that adapt in real time to lure and study attacking agents, learning their tactics and tool signatures before those agents reach production environments. This transforms the defender's position from purely reactive to actively informing intelligence about attacker behaviour.

Behaviour-based anomaly detection using foundation models trained on organisational baselines represents perhaps the most powerful defensive application: detecting agentic attackers by the statistical signatures of their tool-use patterns - unusual API call sequences, atypical data access graphs, anomalous inter-service communication - rather than by the signatures of their code.

The Feedback Loop Problem

The arms race creates a dangerous feedback dynamic. Offensive AI learns from defensive AI's responses. When a defensive agent blocks a technique, the attacking agent observes the failure, updates its approach, and tries again. This is reinforcement learning in action - and it means that static defensive policies are inadequate. Defences must learn and adapt at the same pace as attacks.

When both attack and defence operate at millisecond speeds, human analysts become structural bottlenecks. This creates an irreducible governance challenge: how much authority should be pre-delegated to autonomous defensive systems? A defensive agent that correctly isolates a compromised server may prevent a catastrophic breach. The same agent acting on a false positive may take down a critical production system.

280 days Average dwell time for undetected breaches (IBM 2024)	**3.2 min** Average AI agent response time vs. 24 hrs human SOC triage
94% of security leaders report AI-assisted alert volume increase	**$1.88M** Average savings when AI security tools are fully deployed

Governance and the Human-in-the-Loop Dilemma

The governance question is not theoretical. Under the EU's NIS2 Directive and DORA framework, organisations are required to demonstrate that their security controls are appropriate and proportionate - which implies accountability for decisions made by automated systems. The emerging

EU AI Act creates additional obligations for high-risk AI applications, a category into which autonomous cybersecurity decision-making is likely to fall.

Best practice emerging from leading security organisations involves a tiered authority model: autonomous agents are pre-authorised to take low-impact containment actions (network isolation, account suspension) immediately, while higher-impact actions (service termination, data quarantine) require human confirmation within a defined time window. Clear kill-switch criteria and escalation triggers must be defined before agents are deployed, not after the first false-positive incident.

> **Emerging Best Practice: Tiered Autonomous Authority**
> Tier 1 (Auto-execute): Endpoint isolation, suspicious process termination, temporary account lockout - reversible in under 5 minutes, low blast radius. Tier 2 (Confirm in 15 minutes): Service quarantine, certificate revocation, network segment blocking. Tier 3 (Human decision required): Data erasure, extended service outage, external communications. Each tier should be reviewed quarterly as organisational risk appetite and AI capability mature.

BUILDING AI RESILIENCE: DATA POISONING AND MODEL-STEALING ATTACKS

A dimension of the agentic AI threat that receives insufficient attention from security leadership is the attack surface represented by AI systems themselves. Defensive AI models - the systems that detect intrusions, triage alerts, and make autonomous security decisions - are high-value targets in their own right. Compromising them is potentially more valuable to an attacker than breaching any individual endpoint.

Data Poisoning: Corrupting the Foundation

Machine learning models learn from data. If an attacker can corrupt the training data of a defensive AI model - feeding subtly manipulated threat intelligence, tampered logs, or mislabelled datasets - they can cause the model to systematically misclassify future threats. This is data poisoning, and it is one of the most insidious attacks possible against an AI-dependent security architecture because it degrades capability silently over time.

The most dangerous variant is the **sleeper poisoning attack**, in which a backdoor is embedded during training that only activates under specific trigger conditions. A particular IP address, payload sequence, or timing pattern causes the compromised model to suppress alerts or misclassify malicious activity as benign - while behaving normally in all other contexts, evading detection by standard model validation procedures.

Supply chain poisoning via third-party threat intelligence feeds represents a particularly high-risk vector. Many security products consume shared indicator-of-compromise (IOC) databases. An attacker who can manipulate a widely used feed can degrade the defensive capabilities of hundreds of organisations simultaneously.

Model-Stealing and Adversarial Examples

Model-stealing attacks allow adversaries to extract the functional equivalent of a proprietary security model by systematically querying it and observing its responses. This creates a replica that attackers can test their tools against offline - enabling them to craft inputs that evade the production model before deploying them in a real attack. For AI-based malware classifiers, this is particularly devastating.

Adversarial examples - inputs specifically crafted to fool AI classifiers while appearing legitimate to human reviewers - are a well-documented attack vector in image recognition that translates directly to security AI. A malicious payload can be structured to evade an AI-based intrusion detection system while remaining fully functional. The mathematical space of such adversarial inputs is vast, and defending against all possible variants simultaneously is a fundamentally hard problem.

Organisational Resilience Recommendations

Defending AI systems requires extending the same security rigour applied to production software to the AI development and deployment pipeline:

- **Data provenance tracking:** Every training dataset should be version-controlled and auditable. Provenance metadata should record the origin, processing history, and integrity hash of all training inputs.
- **Federated learning where appropriate:** Distributing training across multiple data sources without centralising them reduces the blast radius of any single poisoning attack.
- **Continuous model monitoring:** Treat AI model drift and performance degradation as security events, not merely operational issues. Statistically significant shifts in model outputs should trigger investigation.
- **AI red-teaming:** Before deploying any autonomous security AI, conduct dedicated red-team exercises specifically targeting the AI system - attempting poisoning, model theft, and adversarial example attacks against it.
- **Vendor AI assurance:** Security product vendors should be required to demonstrate model validation procedures, training data governance, and adversarial robustness testing as part of procurement evaluation.

STRATEGIC RECOMMENDATIONS FOR SECURITY LEADERS

The shift to agentic AI in cybersecurity is not a future scenario requiring monitoring. It is an active transition requiring action. The following recommendations are sequenced by urgency and grounded in the threat landscape described in this paper.

Immediate Actions (0–6 Months)

- **Convene an Agentic AI Threat Working Group** - bring together your CISO, security architects, AI/ML team leads, and legal/compliance representatives to assess current exposure and establish governance frameworks for AI-assisted decisions.
- **Audit your defensive AI estate** - identify all AI and ML systems currently in use across security operations. Assess their data provenance, model update cadence, and adversarial robustness. Flag systems with no documented validation procedures.
- **Establish autonomous authority tiers** - define, in writing, which security actions your existing automation is authorised to take without human confirmation. Most organisations find their implicit policy does not match their actual risk appetite when written down.
- **Begin AI-specific threat modelling** - extend your threat model to include attacks against your AI systems themselves: poisoning, model theft, adversarial inputs. These must be first-class concerns, not afterthoughts.

Medium-Term Investments (6–18 Months)

- **Deploy behaviour-based detection at the AI-speed tier** - if your SOC still relies primarily on signature-based detection for endpoint and network threats, you are structurally unable to respond to AI-speed attacks. Foundation model-based anomaly detection is no longer experimental - it is operational necessity.
- **Invest in deception technology** - modern AI-adaptive honeypots and deception fabrics generate high-fidelity intelligence about attacker tooling and behaviour that cannot be obtained any other way. This intelligence directly informs your defensive AI training.
- **Build an internal AI red team capability** - or contract a specialist provider to conduct quarterly adversarial exercises specifically targeting your defensive AI systems. Standard penetration testing does not cover this attack surface.
- **Engage your cyber insurance underwriter** - many underwriters are beginning to assess AI-related exposures. Understanding how your agentic AI posture affects your coverage and premiums is a board-level financial question, not merely a technical one.

Strategic Positioning (18+ Months)

Organisations that establish strong foundations in AI governance, behaviour-based detection, and AI system resilience over the next 18 months will be well-positioned for the longer-term transition to fully autonomous security operations. The maturity model below provides a framework for assessing current state and plotting a progression path.

Maturity Level	Characteristics	Priority Actions
1 - Reactive	Rule-based automation only; human-driven triage; no AI governance framework	Establish governance; audit current tooling; begin behaviour-based pilot
2 - Informed	ML-assisted triage; documented AI authority tiers; basic model monitoring	Deploy behaviour-based detection; initiate AI red-team programme
3 - Proactive	Autonomous Tier 1 actions; adversarial robustness testing; AI provenance tracking	Expand autonomous authority; deploy deception technology; build AI-speed response
4 - Adaptive	Full AI-vs-AI capability; continuous model retraining; real-time policy adaptation	Optimise feedback loops; mature governance; contribute to sector intelligence sharing

> *"The organisations that will be most resilient are not those with the largest security budgets, but those that establish the governance, talent, and architectural foundations for AI-speed security operations before they face an AI-speed attack."*

SOURCES AND REFERENCES

1. **Gartner**. *Predicts 2025: AI in Cybersecurity*. https://www.gartner.com/en/cybersecurity

2. **IBM Security**. *Cost of a Data Breach Report 2024*. https://www.ibm.com/reports/data-breach

3. **CrowdStrike**. *Global Threat Report 2024*. https://www.crowdstrike.com/global-threat-report/

4. **Google Mandiant**. *M-Trends 2024 Threat Intelligence Report*. https://www.mandiant.com/m-trends

5. **NIST**. *Adversarial Machine Learning: A Taxonomy (NIST AI 100-2)*. https://airc.nist.gov/

6. **MITRE ATLAS**. *Adversarial Threat Landscape for AI Systems*. https://atlas.mitre.org/

7. **ENISA**. *AI Cybersecurity Challenges - Threat Landscape for AI (2023)*. https://www.enisa.europa.eu/

8. **Verizon**. *Data Breach Investigations Report 2024*. https://www.verizon.com/business/resources/reports/dbir/

9. **European Commission**. *EU AI Act - Official Text (2024)*. https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence

10. **Recorded Future**. *AI-Powered Threat Intelligence - Annual Review 2024*. https://www.recordedfuture.com/

11. **UK AI Safety Institute (AISI)**. *Our Evaluation of Claude Mythos Preview's Cyber Capabilities (2024)*.
https://www.aisi.gov.uk/blog/our-evaluation-of-claude-mythos-previews-cyber-capabilities

Partner with Digital Resilience

Digital Resilience is a leading cybersecurity advisory firm specialising in helping organisations build proactive, resilient security architectures that are fit for the age of agentic AI. Our services span threat intelligence, security architecture review, AI governance framework development, and hands-on red-team exercises targeting AI-dependent security systems.

Whether you are assessing your current exposure to agentic AI threats, designing an autonomous security operations roadmap, or seeking to validate the resilience of your defensive AI estate, our team of practitioners bring the depth of technical expertise and strategic perspective to accelerate your programme.

Protecting your Business Ecosystem

yaron@digitalresilience.co.za | www.digitalresilience.co.za