Transitioning from Reactive to Autonomous AI Cyber Defense

Dan Ungureanu | Cyber Exercises Branch Head, CR14 · Tallinn, Estonia

Session: The Impact of Artificial Intelligence on Cybersecurity and Innovative Trends

SPEAKER

IV NATIONAL CYBERSECURITY FORUM

04.06.2026 · BAKU

The attack-defence gap is real and measurable

60%

faced AI-powered attack

in the past year

7%

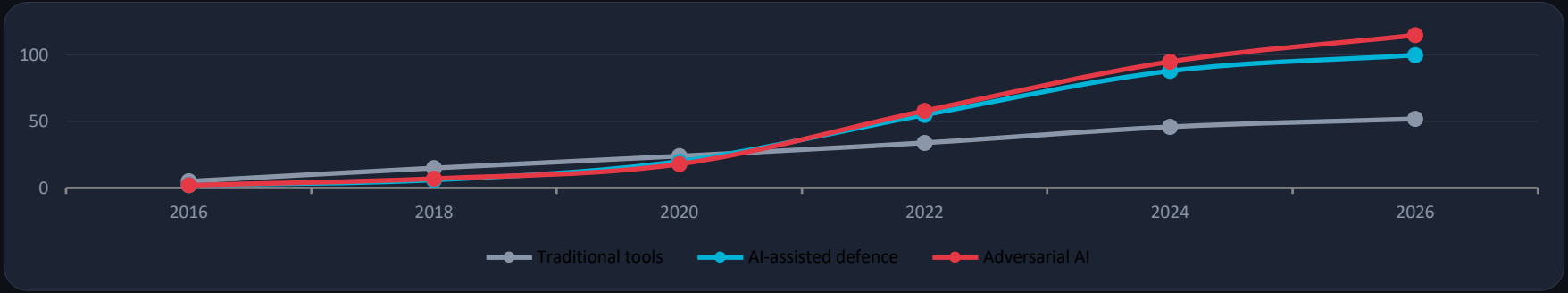
have AI-enabled defence

deployed today

+41

AZ AI Readiness rank jump

111th → 70th globally



Composite index: Oxford Insights GARI, Verizon DBIR, ENISA Threat Landscape (directional, not a single dataset)

↗ BCG report bcg.com/publications/2025/ai-raising-stakes-in-cybersecurity · Oxford Insights oxfordinsights.com/ai-readiness/government-ai-readiness-index-2025

THE CORE QUESTION

Every security team faces the same two questions

TODAY

How can AI help our team respond faster?

- **Alert triage at scale** without burning out analysts
- **Auto-generated playbooks** from incident context
- **Instant hardening** on first compromise indicator

EVENTUALLY

What would a fully autonomous defender look like?

- **Detect. Decide. Act.** no human in the loop for routine incidents
- **End-to-end autonomous** zero analyst touch on standard cases
- **Continuous self-test** adversarial, not annual

Locked Shields 2024–2025, 40+ nations (CCDCOE): ccdcoe.org/locked-shields · AZ AI Strategy decree 19 March 2025: en.apa.az ↗

Autonomy is a spectrum, not a binary switch

L1 ASSISTED

AI surfaces alerts.
Humans decide everything.

Where most teams are today

L2 AUGMENTED

AI recommends and runs pre-approved playbooks.
Humans handle escalations.

Realistic 12–18 month target

L3 AUTONOMOUS

AI detects, decides, acts within policy boundaries.
Humans set the rules.

Tested in exercises. Emerging in production.

Governance must be defined before expanding autonomy from L1 to L2 or L3.

AI CAPABILITIES — TESTED

Four categories validated under adversarial conditions

Agentic Hardening

Browser / host agent

~30% MTTR reduction

Microsoft/Forrester

Traffic Intelligence

Fine-tuned LLM

Detects C2 IDS misses

CCDCOE AI Research, LS exercises

Adaptive Blocking

WAF rule generator

86% XSS bypasses blocked

GenXSS, arXiv Apr 2025

Situational Awareness

AI reasoning agent

Significant report-time cut

Multiple vendor SOC deployments 2025

Three pillars of the reactive to autonomous shift

DATA

- **One pipeline**
  endpoint, cloud, OT, network
- **Quality over volume**
  better signals, not more alerts
- **AzInTelecom H200**
  national compute for government AI

MODELS

- **Triage + classify**
  LLM-assisted, first alert wave
- **Agentic response**
  WAF rules and hardening without tickets
- **Adversarial testing**
  stress-test before production

GOVERNANCE

- **Rules first**
  define AI action boundaries before deploying
- **Log everything**
  immutable audit trail
- **AZ leads**
  6 standards adopted, SSSCIS mandate active

AzInTelecom H200: trend.az/azerbaijan/society/4176621.html · AZ AI Standards (AZSTAND): en.trend.az/azerbaijan/society/4053529.html

What can go wrong — exercise observations

! Prompt injection

Adversaries craft inputs that flip AI triage decisions. New attack surface — not present in rule-based tools.

! Automation-driven cascades

Autonomous remediation can affect large portions of infrastructure in seconds. Human override cannot engage at machine speed.

! Attribution gap

When AI acts, rules of engagement and legal accountability become unclear. Courts and commands are not yet designed for this.

! Governance added last

Teams that deploy without defined boundaries cannot explain or roll back what their AI did. Most preventable failure mode.

Red teams targeted AI coordination logic, not just endpoints. Deploying AI ≠ securing it. NIST AI 100-1 ↗

Why live exercises are where AI defence matures

Multi-sector scope

IT, OT, 5G and cloud running simultaneously

Adversarial pressure

Red teams use AI-assisted attack chains

40+ nations

Locked Shields 2024–2025 (CCDCOE)

Rich telemetry

Thousands of events/hour, real networks

Safe to fail

AI can misfire. Nothing breaks in production.

Measured outcomes

MTTR, false positive rate, containment speed

Locked Shields — world's largest live-fire cyber exercise: ccdcoe.org/locked-shields

EVIDENCE OVER HYPE

What rigorous AI cyber research actually produces

Open code & tools

- **Reproducible prototypes**
  browser agents, WAF generators, LLM classifiers
- **Community-testable**
  break it, improve it, trust it
- **Governance-ready**
  open code reduces black-box concerns

Peer-reviewed findings

- **Exercise-derived data**
  CyCon work on agentic AI for autonomous cyber defence
- **Reproducible baselines**
  what worked, what failed, exact conditions
- **Policy-relevant**
  AZ Strategy prioritises international alignment

Labelled datasets

- **Live exercise traces**
  network captures, alert logs, attack data
- **Adversarially generated**
  conditions lab benchmarks cannot replicate
- **Research-curated**
  some groups have built datasets from exercise data

Live exercises have become research infrastructure. The training is almost a side effect.

AZ AI Strategy: en.apa.az/ict/azerbaijan-approves-artificial-intelligence-strategy-for-2025-2028-463305

KEY TAKEAWAYS

The question is not whether AI will defend

it is who sets the rules.

- **AI is operational infrastructure** not a research project — plan accordingly
- **Tested capabilities exist now** agentic hardening, LLM triage, WAF generation, situational awareness
- **Governance before deployment** AZ: 6 standards, SSSCIS mandate, deputy-minister accountability
- **Exercises = the proving ground** the only safe environment at real adversarial scale

Dan Ungureanu

Cyber Exercises Branch Head · CR14
Tallinn, Estonia

linkedin.com/in/danungureanu

GIAC GPEN · GCIA · GCED · GCIH