

CURATED STREAM

AI and Cybersecurity

Exploring the intersection of AI and cybersecurity

Contents

01 Introduction to AI

02 Understanding Cybersecurity

03 AI-Driven Cybersecurity Solutions

04 Challenges and Risks of AI in Cybersecurity

05 Implementing AI in Cybersecurity Practices

06 Future Trends in AI and Cybersecurity

What is AI?

Artificial Intelligence is the branch of computer science focused on building systems that perform tasks normally requiring human intelligence.

Problem-solving
Finding solutions to complex issues

Learning
Gaining knowledge from experience

Reasoning
Drawing logical conclusions from data

Perception
Interpreting images, speech, signals

Language
Processing and generating language

Machine Learning (ML)

A subset of AI that uses algorithms to let computers learn from data and improve over time without explicit programming. Spam filters, for example, learn to recognise phishing emails by analysing patterns in message content.

Deep Learning (DL)

A specialised subset of ML built on multi-layer neural networks for large, complex datasets. It excels at image recognition, speech processing, and anomaly detection in cybersecurity.

Introduction to AI and Cybersecurity



AI and cybersecurity are converging fast.

AI speeds threat detection, predictive analytics, and automated response — and at the same time it opens new attack surfaces.

This stream walks through both sides: how AI is deployed in defence, the risks it introduces, and how organisations are adapting in 2026.

+32%
rise in prompt-injection attempts (Nov 2025 to Feb 2026)

13%
of organisations reported an AI-related breach (IBM 2025)

87%
of security pros are seeing more AI-driven threats (Darktrace)

PART 01 — UNDERSTANDING CYBERSECURITY

Threats, defences, and where AI fits in

Common threats and vulnerabilities

Phishing

Deceptive emails, sites, and messages that trick people into revealing data.

Malware

Viruses, ransomware, and spyware that damage systems or steal information.

Password attacks

Attempts to steal or crack credentials to gain unauthorised access.

Insider threats

Employees or contractors causing harm, deliberately or otherwise.

Man-in-the-middle

Intercepting traffic between two parties — via Wi-Fi, spoofing, or proxies.

Advanced persistent threats

Long-running intrusions that lie dormant until activated.

Zero-day exploits

Attacks on unknown vulnerabilities before a patch is available.

AI-enabled attacks

Prompt injection, deepfake-driven phishing, and model abuse — new in 2025-26.

Essential cybersecurity measures

Access control

- Multi-factor authentication (MFA) — something you know, have, and are
- Zero Trust Architecture (ZTA) — never trust by default; validate every request

Encryption

- Strong encryption for data in transit and at rest
- Post-quantum cryptography to future-proof against quantum attacks

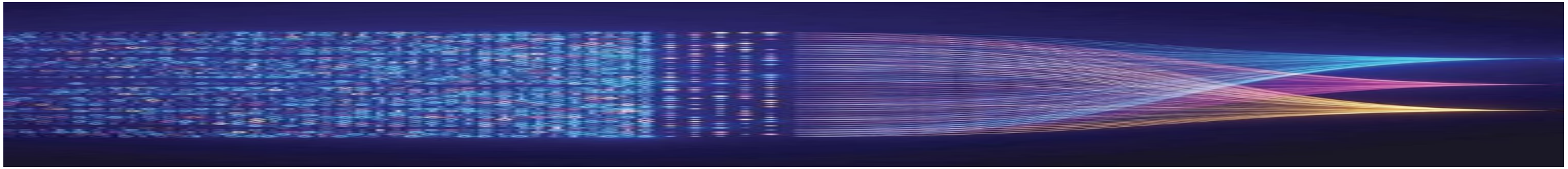
Behavioural analytics

- AI watches user and system behaviour for unusual patterns
- Example: an account pulling large volumes of files at odd hours

PART 02 — AI-DRIVEN CYBERSECURITY SOLUTIONS

How AI defends modern systems

AI techniques used in cybersecurity



ML

Machine Learning

Detects anomalies, phishing, and malware activity. Drives predictive analytics and automated incident response.

NLP

Natural Language Processing

Reads email, chat, and document content to flag phishing, social engineering, and policy breaches. Now extending to multimodal — text, image, audio, video.

DL

Deep Learning

Finds threats hidden in large datasets. Powers intrusion detection, behavioural baselining, and content filtering at scale.

Advanced expert-system components

ATP — Advanced Threat Protection

ATP platforms prevent, detect, and respond to sophisticated cyber threats by combining several security layers in one system.

They ingest network, identity, endpoint, and email telemetry — then apply machine learning to spot patterns that traditional tools miss.

Examples: Microsoft Defender, CrowdStrike Falcon, and specialised next-gen XDR platforms.

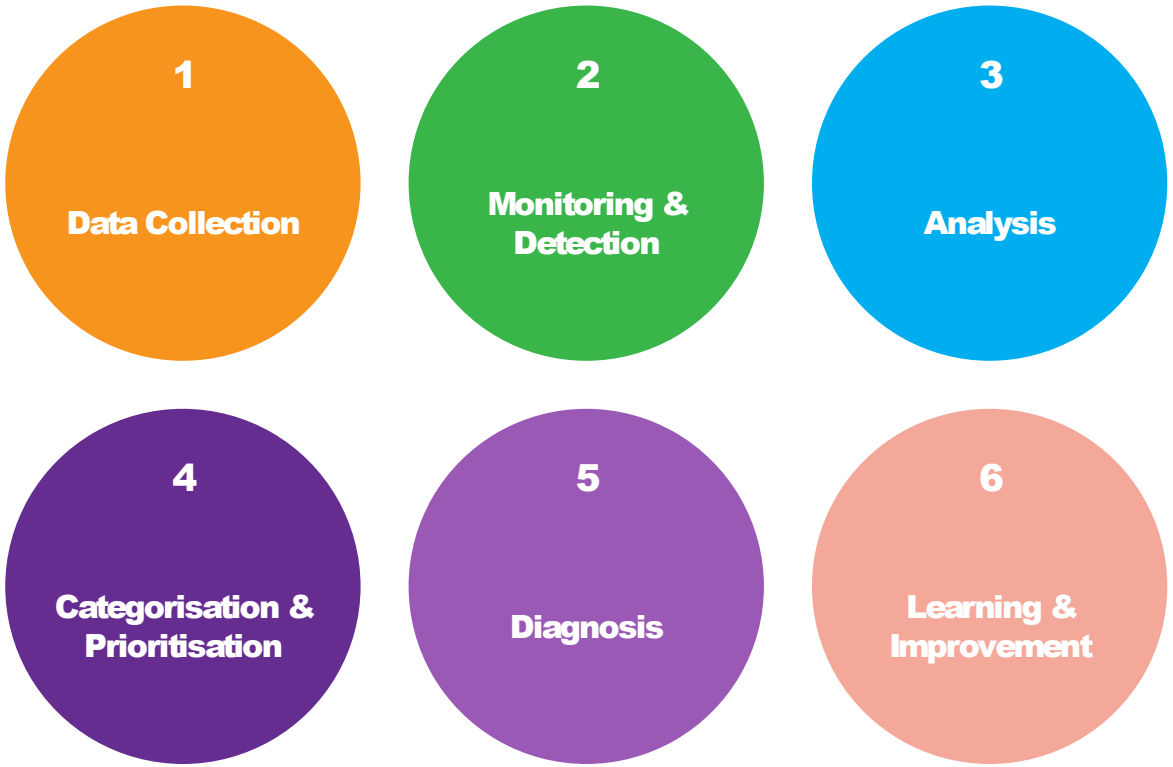
Zero Trust Architecture (ZTA)

A security model that assumes no user or device is trustworthy by default. Every access request is continuously verified — identity, device posture, and context.

Widely used in network and identity stacks: Cisco Zero Trust, Palo Alto Prisma Access, Okta Identity Cloud.

In 2026, ZTA increasingly extends to AI: agents and tools authenticate with the same scrutiny as users.

Automated response systems



How it works

Continuously scan the network — collect telemetry, monitor for anomalies.

Triage with rules, ML, and AI — separate false positives from real threats.

Act automatically — isolate hosts, block IPs, kill suspect processes by severity.

Integrate with SIEM and SOAR platforms for whole-of-stack visibility.

Learn from each incident — detection improves over time.

Benefits

Faster mitigation · 24/7 cover · less human error · real-time intelligence · frees analysts for complex work

Real-world applications

1 Threat detection

CrowdStrike Falcon and Darktrace use ML to spot threats in real time — Darktrace reports 87% of security pros now see more AI-driven threats.

2 Phishing detection

NLP reads email content and metadata to catch phishing and business-email compromise — including AI-written variants.

3 Automated incident response

Microsoft Sentinel, Splunk SOAR, and Google SecOps cut response times by orchestrating actions across the stack.

4 Fraud detection

Banks score transactions in milliseconds against AI baselines of normal customer behaviour.

5 Endpoint security

Behaviour-based endpoint AI replaces signature-only antivirus — the standard for laptops and phones in 2026.

6 AI governance & runtime control

New category: tools that block prompt injection, tool misuse, and identity impersonation as agents act.

PART 03 — CHALLENGES AND RISKS

What AI breaks when it joins the defence

The 2026 AI threat landscape

#1

Prompt injection

is the top risk in the OWASP Top 10 for LLM Applications (2025 edition).

+32% rise in attempts Nov 2025 → Feb 2026.
+90% rise in customer deployments showing potential signals.

HashJack (Nov 2025)

Novel attack exploiting AI browsers through URL-fragment payloads — the AI reads attacker-controlled content the user never sees.

ShadowLeak (Sep 2025)

Email-agent attack that exfiltrated personal data via ChatGPT. Patched, but a clear marker of the new attack surface.

Indirect prompt injection

Now the dominant vector. The model reads untrusted content (a webpage, a document, a tool result) that issues hidden instructions.

Potential biases in AI algorithms

What is bias in AI?

AI bias occurs when algorithms make decisions influenced by prejudice or incomplete data. It can come from:

- Skewed training datasets
- Implicit human bias during model development
- Narrow perspectives in AI design teams

Impact

- False positives / negatives — legitimate activity flagged or real threats missed
- Discrimination — biometric systems denying access to certain groups

Mitigating bias

1. Diverse, representative datasets

Avoid skew at the source — sample widely, document gaps.

2. Regular algorithm audits

Run ongoing fairness reviews after deployment, not just before.

3. Human oversight

Keep a human in the loop for high-stakes decisions. Note: the human's own assumptions matter.

4. Transparent decision logs

Make every model decision auditable so disputes can be reviewed.

Adversarial attacks on AI systems

Deliberate attempts to deceive or manipulate AI models by exploiting their vulnerabilities. They undermine integrity, reliability, and trust.

Evasion

Inputs crafted to push a model to the wrong decision (adversarial examples).

Poisoning

Malicious data injected during training to teach the model wrong behaviour.

Model inversion

Reconstructing sensitive training data from the model's responses.

Model stealing

Repeated queries used to clone the model's behaviour into an unauthorised copy.

Prevention

Adversarial training

Expose the model to attack examples during training so it learns to resist them.

Input validation

Filter inbound prompts and content — flag injection patterns at the perimeter.

Data sanitisation

Verify, clean, and version training data before it lands in the pipeline.

Model hardening & monitoring

Rate-limit, log every call, and use AI runtime guards to catch tool misuse live.

Ethics and data privacy

Personal data collection

Cybersecurity AI runs on a lot of personal data. That makes privacy and consent a first-class design concern, not an afterthought.

EU AI Act fully applicable: 2 August 2026

- Informed consent from users
- Transparency on how decisions are made
- Organisational accountability for AI outputs
- Conformity assessments and CE marking for high-risk systems
- Penalties for non-compliance — enforced by the European AI Office

Ethical AI practices

1. Privacy by design

Minimise data collected; encrypt by default; clear retention rules.

2. Transparency

Document model, training data, intended use, and known limits.

3. Regular audits

Independent reviews of fairness, accuracy, and security posture.

4. NZ guidance

Follow NZ Digital Government's Responsible AI Guidance for the Public Service.

5. IBM 2025 reality check

97% of breached orgs lacked proper AI access controls — start with controls.

PART 04 — IMPLEMENTING AI IN PRACTICE

From pilot to production

Steps to integrate AI into existing systems

1

Data preparation

Collect, clean, and pre-process the data. Document sources and gaps.

2

Model development

Select, train, and optimise models. Validate on held-out data and red-team for adversarial inputs.

3

Deployment & monitoring

Integrate with existing systems. Monitor accuracy, drift, fairness, and security continuously.

Best practices for AI in cybersecurity

1

Security across the AI lifecycle

Build threat modelling, secure coding, and red-teaming into every stage of model development.

2

Prevent data leakage

Encrypt training data, classify by sensitivity, restrict who and what can access models.

3

Ensure compliance

Map to GDPR, the EU AI Act, NZ Privacy Act, and your sector's controls — early.

4

Monitor, iterate, validate

Track accuracy, drift, fairness, and adversarial signals after launch — not just at go-live.

5

Human + AI collaboration

Treat the AI as an analyst's tool, not a replacement. Escalation paths to humans for ambiguous calls.

PART 05 — FUTURE TRENDS

What's coming next

Emerging AI in cybersecurity

LLMs in defence

Modern LLMs — Claude Fable 5 / Opus 4.8, GPT-5.6, Gemini 3.5 Flash — read code, logs, and email at scale to spot threats faster than rule-based systems.

- 1M-token context windows for whole-codebase review
- Multimodal — analyse screenshots, recordings, and PDFs together
- Used in agentic SOC workflows: triage, enrich, recommend, act
- Run on-prem or in private clouds for sensitive workloads

New threat surfaces

Deepfake-driven fraud

Voice and video clones used in CEO-fraud, call-centre impersonation, and family-emergency scams.

AI-written phishing

Grammatically perfect, context-aware lures — old 'spot the typo' advice no longer works.

Agentic abuse

Attackers chain prompts to weaponise the victim's own AI agent against itself.

Browser-based AI attacks

HashJack-style attacks on AI browsers reading attacker-controlled URL content.

On the horizon

Quantum computing

Threat: breaks much of today's public-key crypto.
Opportunity: post-quantum cryptography (PQC) and ultra-fast anomaly detection across massive datasets.

Federated learning

Decentralised training — banks can build shared fraud models without sharing customer data. Privacy + collective intelligence.

Decentralised AI

Blockchain-anchored model registries and audit logs. Tamper-evident records of what model decided what, and why.

Agentic AI defenders

AI agents that hunt threats across your stack — opening tickets, isolating hosts, and writing post-incident reports autonomously.

Predictions and future outlook

1

AI + cybersecurity will keep co-evolving

Defenders will rely more on ML, NLP, and agentic AI; attackers will use the same toolbox. Speed of iteration becomes the moat.

2

AI-powered security analytics

More signal, less noise. Models will explain their findings in natural language, making the SOC accessible to a wider workforce.

3

Autonomous response

Automated triage, containment, and remediation will move from low-risk patterns into higher-risk ones — under tighter governance.

NZ context

AI could boost NZ GDP by up to NZ$108 billion by 2038 (MBIE analysis). Building the security capability to use AI safely is core to that opportunity.

Preparing for the evolving landscape

Stay informed

Keep up with where AI and cybersecurity are heading — the landscape changes faster than most policy cycles.

- OWASP Top 10 for LLM Applications (2025 ed.)
- EU AI Act — full obligations from 2 Aug 2026
- IBM Cost of a Data Breach (annual)
- NZ Digital Government: Responsible AI Guidance
- Major vendor threat reports (Microsoft, Darktrace, CrowdStrike)

Adapt practices

Match the change in your tools, processes, and people. Strong fundamentals still matter most.

- Treat AI systems as critical assets — inventory, patch, monitor
- Bake in identity, access, and runtime controls from day one
- Test for prompt injection and data leakage like any other vuln
- Run table-top exercises that include AI-enabled scenarios
- Train staff: phishing is now AI-perfected

Key takeaways

Opportunities

AI enables faster threat detection, deeper analytics, and automated response — multiplying analyst capacity.

Challenges

AI itself becomes an attack surface. Prompt injection (OWASP #1) and data privacy are now first-tier risks.

Future outlook

Ethics, governance, and emerging tech — quantum-safe crypto, federated learning, agentic defenders — will define the next phase.

Call to action

Stay informed. Invest in AI alongside the controls to use it safely. Foster collaboration between human expertise and AI capability — that combination is the strongest cybersecurity posture organisations can build in 2026.