AI Agents: Security Defense or Biggest Threat?

Presented by **Michael Ifeanyi**

BSides Toronto

October 2025

Who Am I?

Solutions Engineer

Supporting enterprise customers with Kubernetes networking and security in cloud environments

Independent Researcher

Passionate about AI security

About This Talk

01

AI Threat Landscape

(what we're facing)

02

Defense Capabilities

(what research shows works)

03

Conceptual Demo

(seeing AI security in action)

04

Implementation Reality

(how to actually do it)

05

Open Discussion

(learning from you)

This talk combines academic research synthesis with conceptual demonstration to provide an honest assessment of AI agents in security. You'll get research-backed insights and hands-on exploration, not vendor promises or production war stories.

Goal: Cut through hype, share research, discuss reality

The AI Threat Reality

55%

More Effective

AI phishing improvement in effectiveness vs humans (Hoxhunt, 2025)

67%

AI-Powered

Of phishing attacks used AI in 2024 (CybelAngel)

95%

Cost Savings

Cost savings for attackers using AI (Harvard Business Review)

AI campaigns achieve 42% higher success rates. Attacks evolve faster than traditional defenses can adapt.

Sources:

- Hoxhunt AI Phishing Research (2025)
- CybelAngel External Threat Report (2025)
- StrongestLayer (2025)
- TechMagic (2025)

Meanwhile, security teams are like...

... as AI-powered attacks surged by 67% in 2024 (CybelAngel).

Defense Industry Benchmarks

Speed

- **98 days faster** incident detection and containment (IBM/Ponemon 2024)

Efficiency

- 25% of analyst time wasted on false positives (Exabeam).
- AI automation saves **$2.2M average** per breach (IBM 2024)

Accuracy

- **97.2% detection accuracy** achievable (Wazuh ML research).

Sources:
- IBM Cost of a Data Breach Report 2024
- Exabeam Study (2019)
- Wazuh ML Research (2025)

Conceptual Demo

AI Security Dashboard Demo

01

Real-time Metrics

Threats processed, accuracy rates, response times

02

Live Threat Detection

Continuous monitoring with severity scoring

03

Email Analysis

Confidence scoring with reasoning explanations

04

Human Escalation

12 items requiring judgment vs 950 auto-processed

Conceptual demo illustrating AI security operations based on industry research

How AI Detection Works

Endpoint Analysis

- Process behavior patterns
- Memory analysis for fileless attacks
- Network communications monitoring
- Ransomware detection (3-5 files, not 3,000)

Network Monitoring

- Traffic flow analysis
- DNS anomaly detection
- API usage patterns
- Non-human timing signatures

Key Insight: AI detects behavior patterns, not just known signatures

Human-AI Collaboration

AI Handles (1000+ items)

- Known threat patterns
- High-confidence decisions
- Automated triage
- Instant response

Humans Handle (12 items)

- Novel attack patterns
- Complex context analysis
- Strategic decisions
- Unknown unknowns

> "Novel attack pattern detected - requires human expertise"

The Security Professional's Dilemma

AI excels at handling the high volume of routine tasks, freeing human experts for critical analysis and decision-making on complex, novel threats.

Implementation Reality

1

Phase 1 (Weeks 1-2)

Foundation: Audit tools, establish baselines

2

Phase 2 (Month 1)

Basic Defense: Deploy one capability (start with email)

3

Phase 3 (Months 2-3)

Integration: Connect systems, automate response

Reality Check: First month: 40-50% false positives. Month 3: 90%+ accuracy after tuning.

Biggest challenges: Data quality, integration, organizational resistance

Deployment: Expectations vs. Reality

Expectations

Reality

Understanding these initial challenges is crucial for successful AI security integration.

Open-Source Security Tools

Email Security

Elastic Security AI, SpamAssassin for threat detection

Network Monitoring

Suricata, Zeek for traffic analysis

Endpoint Protection

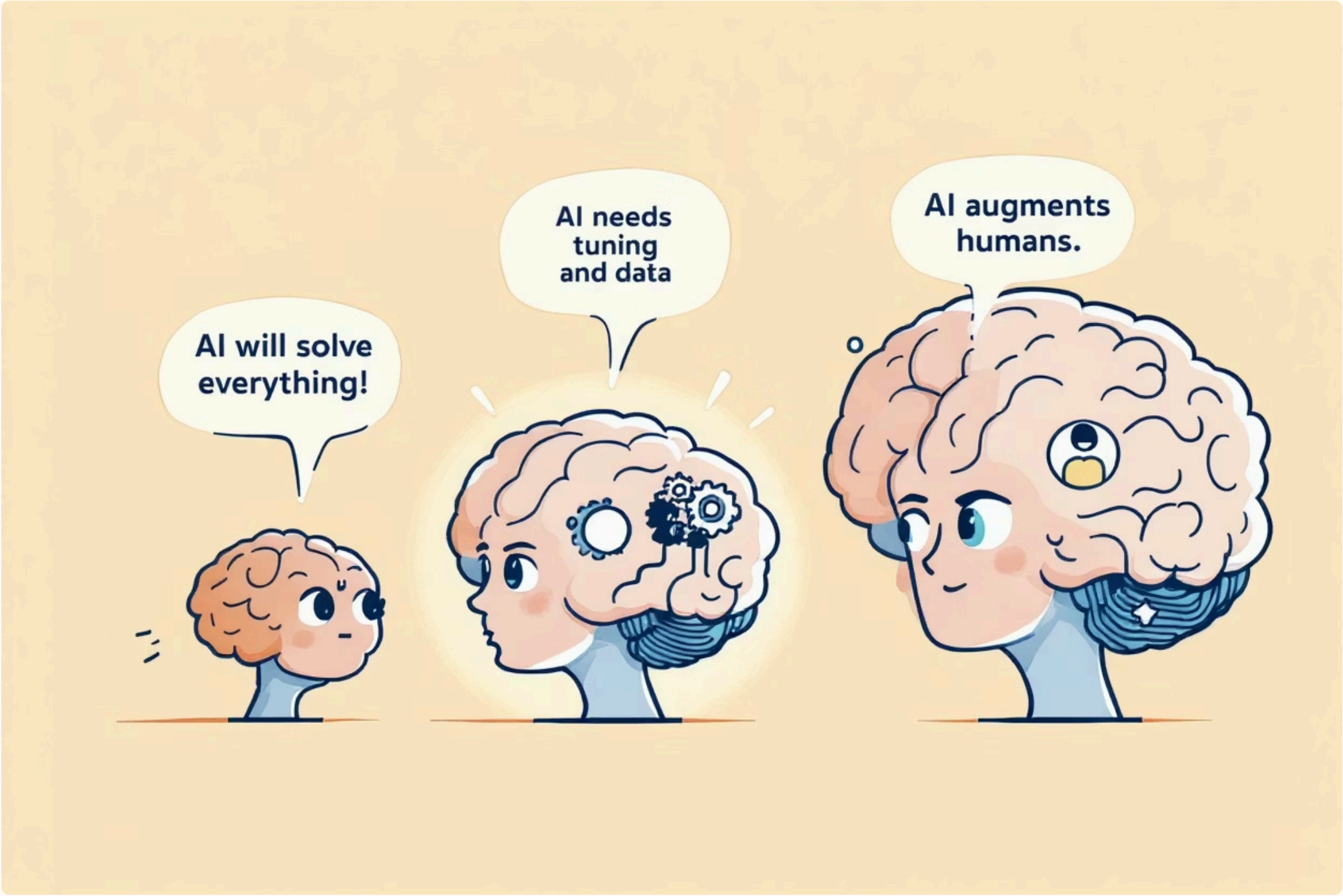
Wazuh (97.2% accuracy), OSSEC

SIEM Integration

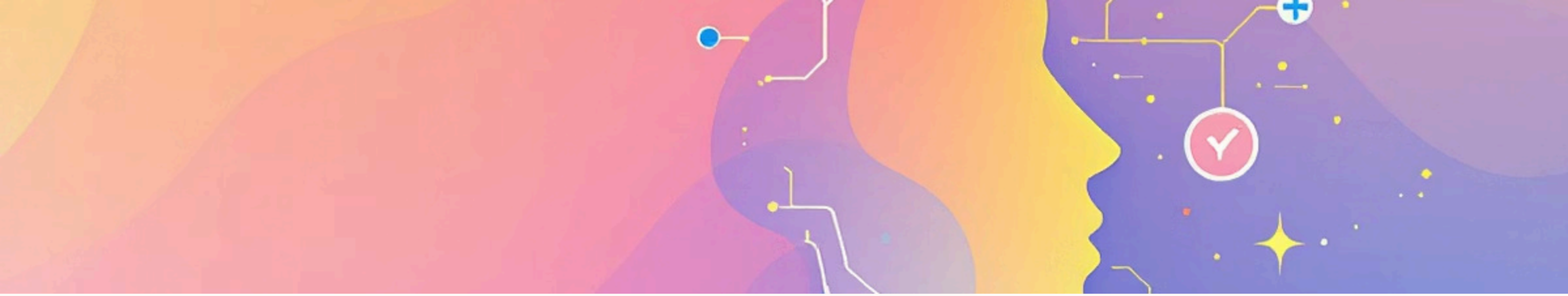
Elastic Security, Wazuh platforms

Research Sources: IBM/Ponemon, Hoxhunt, NIST AI Risk Framework, OWASP Guidelines

Hype vs. Reality: The AI Security Evolution



Moving beyond initial expectations, we find the true potential of AI in security lies in intelligent augmentation and continuous refinement.



Key Takeaways & Next Steps

What Research Shows

- ✅ AI defense achieves 97%+ accuracy
- ✅ 98 days faster incident containment
- ✅ Processes millions of events at scale

The Reality

- ⚠️ 2-3 months tuning to reach maturity
- ⚠️ Data quality is 80% of the challenge
- ⚠️ Human oversight remains critical

> **Bottom Line:** AI augments analysts, doesn't replace them

Thank You

🤝 Let's Connect

LinkedIn

linkedin.com/in/mifeanyi/

Email

michael@michaelifeanyi.com