Ransomware Reimagined

Defending Against AI-Generated Malware in the Modern Threat Landscape

Darryl Baker | Security Research | 2026

About Me

WHOAMI

Darryl G. Baker

Senior Staff Security Researcher, Netwrix

Creator of Active Directory Hacking Village

Identity Security Instructor

Ham Radio Extra

AGENDA

01 **How AI Generates Polymorphic Malware**
LLM-at-runtime architecture, in-memory synthesis, evasion mechanics

02 **The Malicious LLM Toolchain**
WormGPT, FraudGPT, KawaiiGPT — capabilities and architecture

03 **AI-Enhanced Kill Chain Deep Dive**
Technical walkthrough: recon through detonation with MITRE ATT&CK mapping

04 **Detection Engineering**
Behavioral indicators, log analysis, and detection rule design

05 **Defense Architecture**
Zero trust, identity security, and automated response patterns

06 **Live Demo Walkthrough**
Attack simulation and kill chain disruption

netwrix

A BRIEF HISTORY OF RANSOMWARE

1989 — **AIDS Trojan (PC Cyborg)**

First known ransomware. Distributed via floppy disks at WHO AIDS conference. Used symmetric cryptography to hide directories and encrypt file names. Demanded $189 mailed to a PO Box in Panama.

2005 — **GPCoder & Early Crypto-Ransomware**

First ransomware to use asymmetric encryption (RSA). Marked the shift from nuisance-ware to genuine extortion. Demanded payment via e-gold and Liberty Reserve.

2013 — **CryptoLocker & Bitcoin Era**

Industrialized ransomware with strong RSA-2048 encryption and Bitcoin payments. Operated via the Gameover ZeuS botnet. Earned an estimated $27M in its first two months.

2017 — **WannaCry & NotPetya**

Self-propagating worm-style ransomware exploiting EternalBlue (MS17-010). WannaCry hit 200K+ systems in 150 countries. NotPetya caused $10B+ in damages worldwide.

2019 — **Double Extortion & RaaS**

Maze group pioneered data exfiltration before encryption. Ransomware-as-a-Service (RaaS) platforms lowered the barrier to entry. Affiliate models professionalized the ecosystem.

2024+ — **AI-Augmented Ransomware**

LLM-driven polymorphic payloads, automated reconnaissance, and adaptive evasion. The focus of this presentation.

How AI Generates Polymorphic Malware

LLM-at-runtime architecture and evasion mechanics

POLYMORPHISM: TRADITIONAL VS. AI-DRIVEN

Traditional Polymorphic Engine

Mechanism:

Pre-compiled obfuscation routines rotate encryption keys, variable names, and code order

Limitations:

Finite mutation space

Patterns are discoverable through static analysis and emulation

Detection:

Heuristic engines and sandbox detonation can unpack and identify the underlying payload

Example: Metamorphic engines like Virut, Simile

AI/LLM Polymorphic Engine

Mechanism:

LLM API called at runtime to synthesize entirely new code from natural language prompts

Advantage:

Infinite mutation space

Each execution produces semantically equivalent but structurally unique code

Detection Challenge:

No static signature possible. In-memory only. No disk artifacts. Code never repeats.

Example: BlackMamba, PROMPTFLUX, Ransomware 3.0

BLACKMAMBA: LLM-AT-RUNTIME ARCHITECTURE

Simplified Execution Pattern:

```
# Prompt embedded in loader (benign on disk)
prompt = "Generate a Python keylogger that captures\n
keystrokes and stores in memory"
response = openai.ChatCompletion.create(
    model="gpt-4", messages=[{"role":"user",
    "content": prompt}])
exec(response.choices[0].message.content)  # in-memory
```

Why This Defeats EDR:

- **No malicious bytes on disk** — loader is benign
- **No static signature** — code is unique per run
- **No known hash** — payload generated dynamically
- **Living-off-the-land** — uses legitimate API calls

AI MALWARE PROOF-OF-CONCEPTS & VARIANTS

Variant	LLM Backend	Delivery	Technical Mechanism	EDR Evasion
BlackMamba	OpenAI GPT-4	Python loader	Runtime code synthesis via API; exec() in-memory; no disk write	Zero detections across multiple EDR tests
PROMPTFLUX	Google Gemini	VBScript dropper	Hourly Gemini queries generate fresh polymorphic payloads	Continuous signature invalidation
PROMPTSTEAL	Qwen LLM	One-line cmds	LLM generates Windows commands to harvest documents on demand	Legitimate CLI tools only
Ransomware 3.0	Any LLM API	NL prompts only	Only prompts in binary; all malicious code synthesized at runtime per environment	No static payload exists to scan
FunkSec	Commercial + underground	Rust encryptor	AI-assisted development: code generation, comments, rapid iteration	Novel code patterns unknown to signatures

The Malicious LLM Toolchain

Underground AI infrastructure powering modern ransomware

Architecture

- **Base model:** GPT-J (6B parameters, open-source)
- **Fine-tuning:** Malware samples, exploit code, phishing templates
- **No guardrails:** Safety alignment deliberately removed
- **Context window:** Supports multi-step attack planning
- **Version 4:** Improved code generation, expanded language support
- **Distribution:** Telegram channels, DarknetArmy forums

Sample Output Capability

```
> Prompt: "Generate PowerShell ransomware
  with AES-256 encryption"

# WormGPT v4 output includes:
- File enumeration logic
- AES-256 encryption implementation
- Ransom note generation (customizable)
- 72-hour payment deadline timer
- C2 callback for key exchange
```

Broader Ecosystem:

- **FraudGPT** — $200/mo, full-spectrum fraud tooling
- **KawaiiGPT v2.5** — free, entry-level malicious AI
- **GhostGPT / DIG AI** — emerging tooling, expanding market

FUNKSEC: AI-ASSISTED DEVELOPMENT IN PRACTICE

Development Process

- Developer self-described as 'a developer, not a coder'
- Used commercial chatbots + Miniapps for code generation
- AI generated Rust-based encryptor from natural language specs
- AI produced detailed code comments for maintainability
- Rapid iteration cycle — weeks instead of months

Technical Tooling

- Rust encryptor — compiled, cross-platform capable
- RaaS infrastructure at $10K per deployment
- Leak site for double extortion operations
- AI-assisted C2 communication tooling
- Automated victim communication templates

Impact & Implications

- 85+ victims in first month of operation
- 113 confirmed compromises across 7 countries
- Proves AI eliminates skill barrier entirely
- RaaS pricing undercuts established groups
- Signals acceleration of low-skill, high-volume attacks

AI-Enhanced Kill Chain Deep Dive

Technical walkthrough with MITRE ATT&CK mapping

AI-ENHANCED KILL CHAIN — MITRE ATT&CK MAPPING

Kill Chain Stage	MITRE ATT&CK	AI Enhancement	Time Delta
Reconnaissance & Initial Access	T1595, T1566, T1190 Active Scanning, Phishing, Exploit Public App	CVE-to-exploit in 10-15 min; AI phishing eliminates red flags; RapidPen: IP→shell in 200-400s	Hours → Minutes
Execution & Persistence	T1059, T1547 Command/Script Interpreter, Boot/Logon Autostart	LLM generates unique scripts per target; polymorphic loader avoids signature; in-memory execution	Static → Dynamic
Lateral Movement & Priv Escalation	T1021, T1078, T1068 Remote Services, Valid Accounts, Exploitation	AI analyzes AD structure in real-time; automated credential harvesting; adapts to encountered defenses	Days → 48 min avg (18 min fastest)
Exfiltration	T1041, T1567 Exfil Over C2, Exfil Over Web Service	AI identifies & prioritizes high-value data (PII, IP, financial); automated staging and transfer	Manual → Automated
Impact	T1486, T1490 Data Encrypted for Impact, Inhibit System Recovery	AI-generated encryption adapts to environment; shadow copy deletion; autonomous ransom negotiation	Full chain <20 min (LockBit 4.0)

netwrix

AI-Automated Exploit Pipeline

1. CVE Published	NVD / vendor advisory
↓	
2. AI Ingests Patch	Diffs vendor commit, identifies vulnerable path
↓	
3. Exploit Synthesis	LLM generates PoC from patch analysis
↓	
4. Weaponization	Payload packaging, delivery mechanism selection
↓	
5. Deployment	Automated scanning + exploit delivery

AI Phishing — What Changed

- **OSINT scraping** — LinkedIn, GitHub, social media for targeting
- **Contextual generation** — role-specific lures (CFO, IT admin, etc.)
- **Multi-language** — native fluency in any language
- **Conversation chains** — multi-turn email threads for credibility

RapidPen (Hadrian Security, 2025)

IP → Shell: 200-400 seconds

Cost per run: $0.30-0.60

Fully autonomous: scan, identify, exploit, access

How AI Accelerates Post-Exploitation

AD Reconnaissance

AI maps Active Directory structure in real-time:
- Domain controllers
- SQL servers
- Trust relationships
- Group policies

T1018, T1069
Remote System Discovery, Permission Groups

Credential Harvesting

Automated extraction from:
- LSASS memory dumps
- Cached credentials
- Kerberos tickets
- Service account tokens

T1003, T1558
OS Credential Dumping, Steal Kerberos Tickets

Defense Adaptation

AI modifies techniques based on encountered defenses:
- EDR evasion adjustments
- Alternate tool selection
- Timing modifications

T1562
Impair Defenses

Privilege Escalation

AI identifies optimal escalation paths:
- Misconfigured GPOs
- Unpatched local privesc
- Token manipulation
- Service account abuse

T1068, T1134
Exploitation for Privilege Escalation, Token Manip

Unit 42 demo: Multiple AI agents compressed full ransomware campaign to 25 minutes | LockBit 4.0: Initial access → full encryption in <20 minutes

EXFILTRATION & ENCRYPTION TECHNIQUES

AI-Driven Exfiltration (T1041, T1567)

- **Data triage:** AI classifies files by value — PII, financial, IP
- **Staging:** Compresses and stages data in temp directories
- **Transfer:** Exfils over legitimate cloud services (OneDrive, GDrive)
- **Volume:** Targets high-leverage data to maximize negotiation power

Encryption & Recovery Prevention (T1486, T1490)

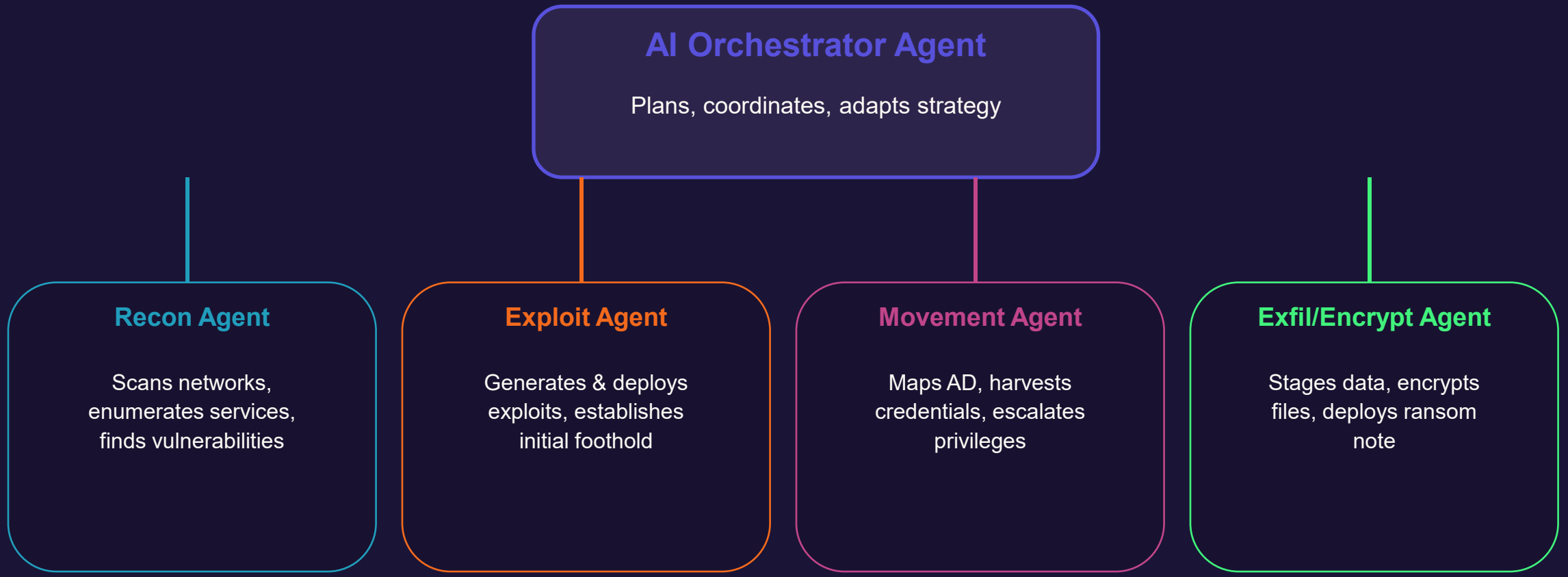
- **Adaptive encryption:** Selects algorithm based on target OS/hardware
- **Shadow copy deletion:** vssadmin, wmic, PowerShell variants
- **Backup destruction:** Targets backup agents, NAS shares, cloud sync
- **Boot record:** Some variants modify MBR for pre-boot ransom note

Common Pre-Encryption Commands (Detection Opportunities):

```
vssadmin.exe delete shadows /all /quiet                    # Shadow copy deletion
wmic shadowcopy delete                                     # Alternative method
bcdedit /set {default} recoveryenabled No                  # Disable recovery
wbadmin delete catalog -quiet                              # Delete backup catalog
net stop "Sophos Agent" & sc config "Sophos Agent" start=disabled  # Kill AV
```

AGENTIC AI: FULLY AUTONOMOUS ATTACK PIPELINES

Predicted by Malwarebytes for late 2026 — multi-agent systems with zero human oversight



Detection Engineering

Behavioral indicators, log analysis, and detection rule design

WHY SIGNATURE-BASED DETECTION FAILS

AI malware systematically defeats each layer of traditional detection

Detection Layer	How It Works	How AI Malware Evades It
File Hash (MD5/SHA256)	Compares file hash against known malware database	Every execution generates unique code — hash never repeats
YARA Rules	Pattern-matches byte sequences and strings in files	LLM-generated code uses different variable names, structures each time
Static Analysis	Disassembles binary to identify malicious patterns	Payload exists only in memory — no binary to disassemble
Sandbox Detonation	Executes in controlled environment to observe behavior	API call to external LLM may be blocked; environment-aware evasion
Network Signatures	Matches network traffic against known C2 patterns	Uses legitimate APIs (OpenAI, Gemini) and services (Teams, GDrive)

BEHAVIORAL DETECTION: WHAT TO MONITOR

Authentication Telemetry

- Failed login spikes against privileged accounts
- Impossible travel: geo-distance vs time delta
- Off-hours access to sensitive resources
- New admin account creation
- Privilege escalation requests (group adds)
- Deviation from user behavioral baseline

File System Telemetry

- Rename operation rate spike (encryption start)
- vssadmin / wmic shadow copy deletion
- Backup service stop commands
- Registry: Run/RunOnce key modifications
- New executables in %TEMP% or %APPDATA%
- Mass file access in sequential patterns

Network Telemetry

- Periodic outbound beaconing (C2 callbacks)
- Large outbound transfers (exfiltration staging)
- DNS queries to unusual/new domains
- SMB/RPC scanning from workstations
- Connections to cloud storage APIs from servers
- Abnormal east-west traffic patterns

DETECTION RULES: PRACTICAL EXAMPLES

Shadow Copy Deletion Detection

```
title: Shadow Copy Deletion
logsource:
  product: windows
  service: process_creation
detection:
  selection:
    CommandLine|contains|all:
      - "vssadmin"   - "delete"   - "shadows"
```

Impossible Travel Detection

```
title: Impossible Travel - Auth Anomaly
detection:
  condition: same_user AND
    geo_distance(loc_1, loc_2) > 500km AND
    time_delta(auth_1, auth_2) < 60min
  severity: high
  mitre: T1078 - Valid Accounts
```

Mass File Rename Detection (Encryption Start)

```
title: Suspicious Mass File Rename
detection:
  condition: file_rename_count > 100
    within: 60 seconds
    by: process_name
  filter_legitimate:
    process_name:
      - "robocopy.exe"  - "xcopy.exe"
```

Backup/Recovery Tampering

```
title: Recovery Prevention Commands
detection:
  selection_bcdedit:
    CommandLine|contains:
      - "recoveryenabled No"
  selection_wbadmin:
    CommandLine|contains:
      - "wbadmin" AND "delete catalog"
```

DETECTION PIPELINE ARCHITECTURE

Data Sources

- Windows Event Logs
- EDR Telemetry
- Network Flow/PCAP
- AD/Identity Logs
- File System Auditing

Collection & Normalization

- SIEM Ingestion
- Log Parsing
- Field Extraction
- Enrichment (GeoIP, Threat Intel)

Correlation & Baselining

- UEBA Baselines
- Entity Behavior Modeling
- Cross-Source Correlation
- Anomaly Scoring

Detection

- Behavioral Rules
- ML Anomaly Detection
- Threshold Alerts
- Kill Chain Pattern Matching

Response

- Automated Containment
- Network Isolation
- Account Lockdown
- SOC Escalation
- Forensic Capture

Defense Architecture

Zero trust, identity security, and automated response

ZERO TRUST ARCHITECTURE FOR RANSOMWARE DEFENSE

Core Principles

- **Verify explicitly:** authenticate and authorize every request
- **Least privilege:** JIT/JEA access, risk-based adaptive policies
- **Assume breach:** minimize blast radius, segment access, verify E2E
- **Microsegmentation:** isolate workloads, restrict lateral paths

Implementation Priorities

- **Network:** VLAN segmentation, micro-perimeters, east-west inspection
- **Identity:** Conditional access, continuous evaluation, step-up auth
- **Endpoint:** Device health attestation, application allowlisting
- **Data:** Classification, DLP at boundaries, encryption at rest

Why Zero Trust Is Critical for AI Ransomware Defense

When lateral movement completes in 18-48 minutes, implicit trust between zones is an existential risk. Microsegmentation ensures that compromising one workstation does not grant access to domain controllers, backup infrastructure, or high-value data stores. Every lateral step requires re-authentication.

IDENTITY SECURITY: THE FIRST LINE OF DEFENSE

Authentication Hardening

- Phishing-resistant MFA (FIDO2, passkeys)
- Certificate-based auth for service accounts
- Conditional access policies (risk-based)

Privileged Access Management

- Just-in-time (JIT) elevation with approval
- Session recording for admin activities
- Automatic credential rotation

Monitoring & Analytics

- Real-time authentication anomaly detection
- Behavioral baselining per user/entity
- Cross-reference with threat intelligence

Attack Surface Reduction

- Eliminate standing privileges
- Disable legacy protocols (NTLM, LDAP cleartext)
- Service account inventory and lockdown

netwrix

AUTOMATED RESPONSE: DISRUPTING THE KILL CHAIN

Tiered response model — automated actions with escalation gates

Tier 1: Immediate (Automated, <60s)

- Isolate affected endpoint from network
- Suspend compromised user account
- Block outbound connections to unknown IPs
- Kill suspicious processes
- Snapshot current system state for forensics

Tier 2: Containment (Semi-auto, <5min)

- Isolate network segment via microsegmentation
- Force password reset for affected accounts
- Disable service accounts showing anomalies
- Quarantine files matching behavioral patterns
- Alert SOC with enriched context

Tier 3: Investigation (Human-led)

- Scope assessment: which systems affected?
- Lateral movement analysis: how far did it spread?
- Root cause: initial access vector identification
- Data impact: was anything exfiltrated?
- Recovery decision: restore from backups?

OPERATIONAL RESILIENCE CHECKLIST

Backup Architecture

- Air-gapped immutable backups (3-2-1 rule)
- Backup integrity testing — automated weekly
- Restore time targets: RPO < 4h, RTO < 8h
- Backup credential isolation (separate AD forest)

Security Awareness

- AI phishing simulations (no grammar errors)
- Focus: contextual verification, not red flags
- Out-of-band confirmation for financial requests
- Regular tabletop exercises for IR teams

Patch Management

- Automated patching for internet-facing systems
- CVE-to-patch target: <24h for critical
- Virtual patching via WAF/IPS for zero-days
- Asset inventory — know what you're defending

Incident Preparedness

- IR playbooks for ransomware scenarios
- Communication plan (legal, PR, regulators)
- Retained IR firm on contract
- Ransomware negotiation policy (pay/no-pay)

netwrix

Live Demo Walkthrough

Attack simulation

DEMO: ATTACK SEQUENCE & DETECTION POINTS

Time	Stage	Action	Detection Opportunity	MITRE
T+0:00	Initial Access	Phishing email delivers malicious document	Email gateway, sandboxing	T1566
T+0:02	Execution	Macro drops loader, calls LLM API for payload	Process monitoring, API egress	T1059
T+0:05	Discovery	AD enumeration — domain controllers, shares, GPOs	LDAP query volume spike	T1018, T1069
T+0:08	Lateral Movement	Pass-the-hash to domain controller	Auth anomaly: impossible speed	T1021, T1550
T+0:12	Exfiltration	Staging and transfer of high-value files	DLP, outbound volume spike	T1041
T+0:15	Impact	Shadow copy deletion → mass file encryption	VSS deletion alert, rename spike	T1486, T1490

KEY TECHNICAL TAKEAWAYS

1. **LLM-at-runtime polymorphism defeats signature-based detection entirely**

   Focus investment on behavioral analysis, UEBA, and ML-based anomaly detection

2. **AI compresses the kill chain — your response must be faster**

   Automated containment playbooks with <60s response for Tier 1 actions

3. **Detection must be behavioral and multi-layered**

   Correlate auth, file system, and network telemetry — no single source is sufficient

4. **Zero trust limits blast radius when prevention fails**

   Microsegmentation, JIT access, and continuous verification at every boundary

5. **Operational resilience assumes breach**

   Air-gapped backups, tested restore procedures, and rehearsed IR playbooks

Questions & Discussion

Darryl Baker | Security Research

netwrix