BLACKHACK SOCIETY | HOME FOR ETHICAL HACKERS

AI & Cybersecurity

Navigating the New Threat Landscape

An Executive Briefing for Business Leaders

AI in Action | Sugar Hill, GA | April 2026

THE THESIS

AI is the most consequential force multiplier in cybersecurity history. It is simultaneously the greatest weapon in the attacker's arsenal and the only scalable defense mechanism available to organizations of any size. The question is no longer whether to adopt AI — it is whether you will wield it before someone wields it against you.

01

Understand the Threat

AI-powered attacks have increased 1,210% in 2025. Attackers are faster, cheaper, and indistinguishable from legitimate activity.

02

Integrate, Don't Outsource

AI should augment your team's judgment — not replace it. The 88% adoption rate masks a scaling gap where only 6% are high performers.

03

Defend with AI

Defenders using AI detect breaches 80 days faster and save $1.9M per incident. The arms race is real and the defenders are winning — if they show up.

Source: Vectra AI, McKinsey State of AI 2025, IBM Cost of a Data Breach 2025

AGENDA

0:00	**The State of Play**	Where AI and cybersecurity collide in 2026
0:05	**The Attacker's Playbook**	How threat actors are weaponizing AI right now
0:12	**The Defender's Arsenal**	AI-powered tools that are changing the game
0:20	**Integration, Not Outsourcing**	Building an AI-augmented security posture
0:27	**The Executive Playbook**	Governance, frameworks, and what to do Monday morning
0:33	**Q&A**	Open discussion

01

The State of Play

Where AI and cybersecurity collide in 2026

THE AI ADOPTION PARADOX

Everyone is adopting AI. Almost nobody is doing it well.

88%
of organizations use AI in at least one business function

McKinsey 2025

6%
qualify as AI "high performers" who generate measurable ROI

McKinsey 2025

83%
lack automated controls for AI deployments

HiddenLayer 2026

98%
of organizations have unsanctioned shadow AI usage

Cyberhaven 2025

Source: McKinsey State of AI 2025, HiddenLayer 2026 Report, Cyberhaven

THE MARKET IS MOVING FAST

AI Cybersecurity Market

- $29B in 2025 → $167B+ by 2034
- 18–22% CAGR across all estimates
- $244.2B total security spending in 2026
- AI security tools: $49B → $160B by 2029

Breach Economics

- $4.44M global average breach cost
- $10.22M average in the United States
- +$670K added cost from shadow AI
- −$1.9M savings with AI security tools

KEY INSIGHT

Organizations deploying AI-driven security detect breaches 80 days faster, contain them in 97 days fewer, and spend $1.9M less per incident than those relying on traditional tools. For SMBs, the math is existential: a single breach at $4.44M can be a business-ending event.

Source: IBM Cost of a Data Breach 2025, Gartner, MarketsandMarkets

02

The Attacker's Playbook

How threat actors are weaponizing AI in 2025–2026

AI-POWERED ATTACK VECTORS

AI-Generated Phishing

82.6% of phishing emails now use AI

60% higher click-through rate vs. traditional phishing

Cost to attackers: as low as $75

Deepfake & Voice Cloning

$200M+ losses in Q1 2025 alone

3 seconds of audio creates 85% voice match

680% YoY increase in deepfake incidents

Prompt Injection

540% surge in prompt injection vulns

Found in 73% of production AI systems

CVEs at CVSS 9.3–9.8 in Copilot, Cursor, GitHub

Source: Zensec, Resemble AI, HackerOne 2025, Obsidian Security

REAL-WORLD AI ATTACKS: CASE STUDIES

$25M

Hong Kong CFO Deepfake

Attackers used AI to replicate the CFO's face and voice on a live video call. Finance team transferred $25 million before discovering the deception.

320 CASES

North Korean AI Infiltration

State-sponsored operatives used GenAI to fabricate resumes, simulate personas, and generate interview materials to infiltrate Western companies as remote employees.

$12.5B

AI Scam Epidemic

Total U.S. consumer fraud losses in 2025. AI scams surged 1,210% while traditional fraud grew 195%. AI-powered fraud is now the dominant attack vector.

Source: The Network Installers, Tech Prescient, FTC / Fortune

ATTACKER ECONOMICS HAVE FUNDAMENTALLY SHIFTED

BEFORE AI		WITH AI
Phishing required skilled copywriters	→	AI generates perfect phishing at $75/campaign
Voice impersonation was near-impossible	→	3 seconds of audio creates an 85% voice match
Pentesting required expert humans	→	AI agents find first exploit in 18 minutes
Vulnerability research took months	→	AI finds thousands of zero-days autonomously
Attacks were manual, slow, expensive	→	28M AI-driven attacks projected in 2025
Required technical skill to execute	→	CaaS lets anyone rent AI attack tools

Source: DEF CON 33 / OpenAI-Stanford, Anthropic Mythos, DeepStrike

SIGNAL FROM THE FRONTIER: CLAUDE MYTHOS

April 7, 2026 — Anthropic announced Claude Mythos Preview, a model so capable in offensive security that it was deemed too dangerous for public release.

Thousands

of zero-day vulnerabilities found across every major OS and browser

27-Year Bug

discovered in OpenBSD — a system renowned for security

83%

first-attempt success rate on autonomous exploitation

10-100x

the output of an elite human vulnerability research team

Project Glasswing: 40+ organizations (AWS, Apple, CrowdStrike, Microsoft, Google) received gated defensive access. $100M in credits committed. Goal: give defenders a head start before Mythos-class models reach adversaries.

Source: Anthropic, Fortune, The Hacker News

03

The Defender's Arsenal

AI-powered tools that are changing the game for security teams

DEFENDERS WITH AI ARE WINNING

80
DAYS
Faster breach detection

97
DAYS
Faster breach containment

$1.9M
Saved per breach

20x
Improvement in MTTD

LEADING AI SECURITY PLATFORMS

Platform	Key Capability
CrowdStrike Charlotte AI	7 autonomous agents, AgentWorks ecosystem
SentinelOne Purple AI	338% ROI, 55% faster remediation (IDC)
Microsoft Security Copilot	68% fewer reopened incidents, 550% faster phishing detection
Darktrace ActiveAI	Self-learning AI maps normal behavior, detects anomalies in real-time

Source: IBM 2025, Stellar Cyber, IDC / SentinelOne

THE AI ARMS RACE: OFFENSE VS. DEFENSE

ATTACKERS

- AI-generated phishing at industrial scale
- Deepfake voice/video for exec impersonation
- Autonomous vulnerability discovery
- AI-powered ransomware target selection
- Polymorphic malware evading signatures
- Context poisoning in RAG systems

VS

DEFENDERS

- AI-powered SOC reduces MTTD by 20x
- Behavioral analysis detects novel threats
- Automated Tier-1 triage and containment
- AI pentesting at $550/day vs. $2K+ human
- Predictive threat intelligence at scale
- Real-time deepfake detection systems

The asymmetry is closing. AI defenders now match attacker speed at a fraction of the cost — but only for organizations that have invested in the capability.

Source: Stellar Cyber, DEF CON 33, HackerOne 2025

FROM THE FRONT LINES: DEF CON 32 & 33

The world's largest hacker conferences are ground zero for AI security research.

DEF CON 33

AI vs. Human Pentesting

OpenAI/Stanford tested AI agents vs. human pentesters on a live university network. AI reached first exploit in 18 minutes; humans took 45. But humans found context-driven vulns that AI missed entirely.

DEF CON 33

Claude in CTF Competitions

Anthropic's Claude scored top 3% in picoCTF and solved 19/20 challenges in Humans vs. AI — but failed at elite competitions requiring deep domain expertise and multi-step reasoning.

DEF CON 33

Shadow Data Extraction

Researchers demonstrated PII and SSN extraction via model inversion attacks on fine-tuned models and embedding inversion on RAG vector databases. Current scanning tools miss this entirely.

DEF CON 32

GRT2: Large-Scale LLM Red Teaming

First-ever "bug bash" for LLMs at scale. Participants evaluated real model behavior across attack surfaces — not just single CTF captures. Government and corporate partners co-hosted.

Source: HackerOne DEF CON 33, AI Village, Google Security Blog

HACKERONE: AI VULNERABILITY LANDSCAPE

210%
Spike in AI vulnerability reports

540%
Surge in prompt injection vulns

$81M
Total bounties paid in 2025

KEY FINDINGS FROM THE 2025 HACKER-POWERED SECURITY REPORT

- Autonomous hackbots submitted 560+ valid vulnerability reports — AI is now a legitimate security researcher
- $3 billion in estimated breach losses avoided through the HackerOne platform
- AI vulnerability rewards jumped 339% year-over-year as organizations prioritize AI-specific bug bounties
- Prompt injection is now the fastest-growing attack category, displacing traditional web application flaws

Source: HackerOne 2025 Report, HackerOne Press Release

04

Integration of AI into your Workflows

The right way to bring AI into your security program

AI AUGMENTS JUDGMENT. IT DOES NOT REPLACE IT.

DEF CON 33 proved it: AI agents reached first exploit in 18 minutes vs. 45 for humans. But humans found context-driven vulnerabilities that AI missed entirely. The winning model is human intuition + AI speed.

Let AI Handle

- Tier-1 alert triage and dedup
- Log correlation at scale
- Phishing email classification
- Vulnerability scanning
- Baseline behavior modeling

Human + AI Together

- Incident investigation
- Threat hunting
- Penetration testing
- Policy enforcement
- Vendor risk assessment

Keep Human

- Strategic decision-making
- Crisis communications
- Legal and compliance calls
- Ethical judgment
- Stakeholder relationships

Source: DEF CON 33 / OpenAI-Stanford, McKinsey

THE SHADOW AI PROBLEM

Your employees are already using AI. The question is whether you know about it.

98%
of organizations have unsanctioned AI usage

68%
surge in shadow GenAI usage in 2025

$670K
added breach cost from high shadow AI exposure

223
sensitive data incidents per company per month

63%
of organizations lack AI governance policies

76%
see shadow AI as a definite/probable problem

Source: Cyberhaven, IBM 2025, HiddenLayer 2026

OWASP TOP 10 FOR LLM APPLICATIONS (2025)

If your organization deploys LLMs, these are your threat vectors.

#	Vulnerability	Risk
LLM01	Prompt Injection	Manipulating LLMs via crafted inputs
LLM02	Sensitive Info Disclosure	Leaking PII, proprietary data in responses
LLM03	Supply Chain	Compromised training data or model repos
LLM04	Data & Model Poisoning	Corrupting training data to embed backdoors
LLM05	Improper Output Handling	Trusting LLM output without validation
LLM06	Excessive Agency	Over-permissioned autonomous agents
LLM07	System Prompt Leakage	Exposing internal instructions & logic
LLM08	Vector & Embedding Flaws	Exploiting RAG pipelines and vector DBs
LLM09	Misinformation	Hallucinations used for social engineering
LLM10	Unbounded Consumption	Resource exhaustion attacks on LLM infra

Source: OWASP LLM Top 10

MITRE ATLAS: AI THREAT TAXONOMY

The ATT&CK framework for AI systems. 14 tactics covering the full AI attack lifecycle.

- Reconnaissance — Probing model APIs and behavior
- Resource Development — Building adversarial tools
- Initial Access — Prompt injection, API exploitation
- Execution — Malicious code via model outputs
- Persistence — Backdoors in fine-tuned models
- Exfiltration — Extracting training data and PII
- Impact — Model degradation, denial of service
- ML Attack Staging — Crafting adversarial inputs

OCTOBER 2025 UPDATE: 14 NEW AGENT-FOCUSED TECHNIQUES

- Memory manipulation in AI agents
- RAG credential harvesting
- Tool invocation abuse
- Multi-agent communication hijacking

Source: MITRE ATLAS

AI SUPPLY CHAIN: THE HIDDEN ATTACK SURFACE

2024	**Hugging Face Malicious Models**	3,000+ harmful files found in public model repos. Most orgs never scan before deployment.
2024	**NullBulge / LockBit Campaign**	Ransomware operators embedded malware inside AI model files on trusted ML platforms.
2025	**nullifAI Bypass Technique**	Attackers bypass Hugging Face model safety scanners entirely, making detection ineffective.
2025	**ShadowRay 2.0**	230,000+ exposed Ray ML servers. Attackers used compromised inference infra for cryptomining.
2025	**Ultralytics Compromise**	Popular CV library backdoored via CI/CD pipeline. Downloaded by thousands before detection.

97% of organizations use models from public repos. Only **49%** scan them before deployment.

Source: HiddenLayer, CSO Online

05

The Executive Playbook

Governance, frameworks, and what to do Monday morning

THREE FRAMEWORKS EVERY EXECUTIVE MUST KNOW

U.S. VOLUNTARY

NIST AI RMF

- GOVERN: Policies, roles, risk culture
- MAP: Identify AI risks in context
- MEASURE: Track and quantify risk
- MANAGE: Prioritize and mitigate

MANDATORY (IF YOU SERVE EU)

EU AI ACT

- Risk-tiered: Minimal to Unacceptable
- Fines up to 7% global revenue
- Applies to U.S. firms with EU users
- Key deadlines through Aug 2026

INTERNATIONAL STANDARD

ISO 42001

- AI Management System standard
- 42 Annex A controls
- Certifiable (audit-ready)
- Aligns with ISO 27001 approach

These are not mutually exclusive. Leading organizations layer NIST AI RMF (operational), ISO 42001 (certifiable), and EU AI Act readiness (regulatory) into a unified governance program.

Source: NIST AI RMF, EU AI Act, ISO 42001

THE REGULATORY LANDSCAPE IS ACCELERATING

FEDERAL

- Jan 2025: Executive Order emphasizing innovation over regulation
- Dec 2025: DOJ AI Litigation Task Force established
- Mar 2026: White House framework calling for federal preemption of state laws

STATE-LEVEL (PATCHWORK)

State	Law / Status	Key Requirement
California	SB 53	Frontier model safety testing
Colorado	AI Act (Feb 2026)	Algorithmic discrimination protections
Illinois	BIPA + AI amendments	Biometric data protections for AI
NYC	Local Law 144	Automated hiring tool audits

Source: CISA, Anthropic RSP v3.0

WHAT TO DO MONDAY MORNING

Seven high-impact actions that require leadership — not a large budget.

1	**Audit Your AI Footprint**	Map every AI tool in use — sanctioned and unsanctioned. You cannot secure what you cannot see.
2	**Establish an AI Governance Policy**	63% of organizations lack one. Define acceptable use, data handling, and risk ownership before an incident forces your hand.
3	**Deploy AI-Augmented Security**	AI SOC tools save $1.9M per breach and detect threats 80 days faster. Start with Tier-1 triage automation.
4	**Test Your Deepfake Readiness**	Implement out-of-band verification for all wire transfers and executive requests. AI voice cloning needs only 3 seconds of audio.
5	**Scan Your AI Supply Chain**	97% use public model repos; 49% scan them. Treat AI models like software dependencies — verify before deploy.
6	**Train Your People**	AI-generated phishing has a 60% higher click rate. Update security awareness training to include AI-specific attack vectors.
7	**Engage Expert Partners**	The threat landscape moves faster than any single team can track. Leverage vCISO services, managed detection, and AI-specific red teaming.

AI SECURITY READINESS ASSESSMENT

Where does your organization stand? Score each dimension 1–5.

Dimension	Assessment Question	1 (Low)	5 (High)
AI Governance	Do you have formal policies governing AI use, procurement, and risk?	No policies	Mature program
Threat Visibility	Can you detect AI-powered attacks (deepfakes, AI phishing, prompt injection)?	No capability	AI-augmented SOC
Supply Chain	Do you scan and validate AI models and third-party AI integrations?	No scanning	Full SBOM for AI
Workforce Readiness	Is your team trained on AI-specific threats and tools?	No training	Continuous upskilling
Incident Response	Does your IR plan address AI-specific scenarios?	Not covered	AI-specific playbooks

Score 5 or below? Your organization has critical gaps. A focused 90-day roadmap can close them before adversaries exploit them.

KEY TAKEAWAYS

01 AI has permanently changed the threat landscape. Attacks are faster, cheaper, and harder to detect than at any point in history.

02 The adoption paradox is real. 88% of organizations use AI, but only 6% generate measurable security ROI. The gap between adopters and high performers is where risk lives.

03 Integration beats outsourcing. The winning model is human judgment augmented by AI speed — not full automation. DEF CON 33 proved it.

04 Defenders with AI are winning. $1.9M saved per breach, 80 days faster detection, 20x MTTD improvement. The tools exist. Deployment is the bottleneck.

05 Governance is non-negotiable. Shadow AI, supply chain risks, and regulatory acceleration mean that inaction is the highest-risk posture available.

BLACKHACK SOCIETY

Home for Ethical Hackers

We help organizations navigate the AI-security landscape with **clarity, capability, and confidence.**

- vCISO & Security Leadership
- Security Awareness Training
- Compliance Programs (SOC 2, ISO 27001, NIST)

Aaron Butler on LinkedIn:

getblackhack.com
founder@getblackhack.com

06

Appendix

Additional data, frameworks, and source references

APPENDIX: CISA AI SECURITY GUIDANCE

May 2025: Joint advisory from CISA, NSA, FBI, and Five Eyes partners on AI data security.

- Ensure data supply chain integrity for AI training pipelines
- Implement data provenance tracking for all model inputs
- Protect against data poisoning via input validation
- Monitor for data drift that degrades model performance
- Apply least-privilege to AI system access controls
- Conduct regular adversarial testing of AI models
- Maintain AI-specific incident response playbooks
- Log and audit all AI system decisions for accountability

Source: CISA AI

APPENDIX: ANTHROPIC RSP v3.0 & SAFETY RESEARCH

RESPONSIBLE SCALING POLICY — AI SAFETY LEVELS

Level	Status	Safeguards
ASL-2	Current standard models	Basic safeguards; detailed requirements defined
ASL-3	Activated May 2025	Input/output classifiers blocking WMD content; dual-layer jailbreak shield (86% to 4.4% attack success)
ASL-4+	Future frontier models	Defense against state-level actors; some mitigations not yet technically feasible

KEY SAFETY RESEARCH (2026)

- Mechanistic Interpretability: Attribution graphs tracing features from prompt to response; goal to detect model deception by 2027
- Model Organisms of Misalignment: 16 frontier models resorted to blackmail and harmful behaviors when facing goal conflicts
- Chain-of-Thought Hijacking: Jailbreak success jumps from 27% to 80%+ with extended reasoning chains (Anthropic/Oxford/Stanford)
- Anthropic Fellows: AI agents autonomously found $4.6M in blockchain vulnerabilities and two novel zero-days

Source: Anthropic RSP v3.0, MIT Technology Review, Fortune

APPENDIX: KEY INDUSTRY REPORTS 2025-2026

IBM COST OF A DATA BREACH 2025

- $4.44M global average breach cost
- $10.22M average in the United States
- $1.9M savings with extensive AI/automation
- 80 days faster detection with AI tools
- 258-day average breach lifecycle
- $670K added cost from shadow AI exposure

VERIZON DBIR 2025

- 22,052 incidents; 12,195 confirmed breaches
- 68% involved a human element
- BEC losses totaled $6.3 billion
- Espionage-related breaches up 163%
- 37% rise in AI-assisted BEC incidents
- Largest dataset ever analyzed by DBIR

ADDITIONAL INTELLIGENCE

- HiddenLayer 2026: 1 in 8 AI breaches linked to agentic systems; 35% from malware in model repos; 31% don't know if they've been breached
- Gartner 2026: $244.2B total security spending; AI SOC agents at Peak of Inflated Expectations; 40% governance gap
- Experian 2026 Forecast: 72% of business leaders see AI-enabled fraud as a top operational challenge

Source: IBM, Verizon DBIR, HiddenLayer