CYBERCLAN™

Incident Response & Breach Prevention

24/7/364 · 15-MIN RESPONSE SLA

Human Hackers vs. AI-Powered Attacks

Cybersecurity fact vs. fiction: a 2025–2026 data briefing comparing how breaches start when a person is manipulated versus when an attacker uses an AI tool to do the manipulating.

62%

of breaches involved the human element — phishing, pretexting, and stolen credentials.

VERIZON 2026 DBIR

300,000+

ChatGPT credential sets advertised on the dark web in 2025 as infostealers expanded to AI services.

IBM 2026 X-FORCE INDEX

73 → 109

distinct extortion groups identified in one year, as AI tooling lowers the barrier to run a ransomware operation.

IBM 2026 X-FORCE INDEX

Seconds

of audio are enough to clone a voice for a fabricated crisis or executive-fraud call.

FBI / IC3 ADVISORY

	HUMAN-INITIATED ATTACKS	AI-POWERED ATTACKS
PRIMARY TECHNIQUE	Social engineering, pretexting, and phishing, increasingly impersonating a trusted contact inside an existing conversation thread.	AI-drafted phishing at scale, cloned voices, deepfake video calls, and purpose-built malicious chatbots.
WHAT IT EXPLOITS	Trust, urgency, authority, and habit, the same psychology attackers have relied on for decades.	The same human psychology, but generated and delivered at machine speed and volume.
SKILL REQUIRED	Ranges from low (mass phishing blasts) to high (tailored, researched pretexting campaigns).	Lowered sharply by ready-made criminal tools, no coding skill or fluent English needed.
TOOLING	Spoofed emails, lookalike domains, stolen credential lists, phone-based pretexting.	**WormGPT** and **FraudGPT**: guardrail-free LLMs sold on criminal marketplaces for BEC messages and malware.
REAL WORLD EXAMPLE	A business-email-compromise message impersonating a known vendor inside a live email thread.	A cloned executive voice authorizing a wire transfer on a live video call.
2025–26 DATA POINT	62% of breaches involved the human element; third-party compromise up 60% YoY to 48% of breaches.	109 active extortion groups (up from 73); 300K+ AI credentials for sale on the dark web.
BEST DEFENSE	Security awareness training, verification habits, MFA, least privilege access.	Same fundamentals, plus out-of-band verification for any voice or video payment request.

Where They Converge

NIST classifies prompt injection, hiding malicious instructions in content an AI later reads, as a distinct form of adversarial ML: it manipulates a model through its normal input channel, not by breaking code. That's structurally the same move as social engineering a person. Either way, fundamentals, patching, MFA, verification habits, stop most of both attack types.

Three Controls That Stop Most of Both Attack Types

1. **Out-of-band verification.** Confirm any payment, credential, or urgent request through a second channel — never trust a single call, email, or video request alone.
2. **MFA + least privilege.** Multi-factor authentication and tightly scoped access limit what a single stolen credential or cloned voice can actually reach.
3. **Ongoing awareness training.** Teams that regularly practice spotting pretexting and AI-generated lures catch them faster, before money or data moves.

Sources

Verizon 2026 Data Breach Investigations Report · IBM 2026 X-Force Threat Intelligence Index · CISA, Avoiding Social Engineering and Phishing Attacks · NIST / IBM, How AI Can Be Hacked with Prompt Injection · Rapid7, What Is WormGPT? · FBI/IC3 PSA, Criminals Use Generative AI to Facilitate Financial Fraud (2024).