SECURITY INSIDER

Wie Angreifer KI-Systeme austricksten und Ransomware durchschleusten

Warum KI in der IT-Sicherheit niemals allein entscheiden darf

05.02.2026 Ein Gastbeitrag von Markus Neumaier 6 min Lesedauer

Ein EDR-System blockierte zunächst Schadsoftware auf einem Domain Controller, doch die Angreifer probierten es wiederholt mit neuen Dateinamen. Das lernende Modell stufte das Muster schließlich als harmlos ein und ließ die Ransomware durch. Fünf Stunden später begann die Verschlüsselung. Der Fall zeigt, warum KI ohne menschliche Kontrolle getäuscht werden kann.

Angreifer probierten wiederholt Schadsoftware auf einem Domain Controller zu platzieren, bis das EDR-System das Muster schließlich als harmlos einstufte. Fünf Stunden später begann die Verschlüsselung.
(Bild: © zephyr_p - stock.adobe.com)

Der Einsatz von KI <https://www.security-insider.de/was-ist-kuenstliche-intelligenz-ki-a-47df994b95769969bdd9467d55c47d20/> , um Cybersecurity-Angriffe abzuwehren, verbreitet sich zunehmend. Dies kann die Sicherheitsarchitektur erheblich stärken. Wird sie jedoch zur einzigen Verteidigungslinie, kann genau darin eine gefährliche Schwachstelle <https://www.security-insider.de/was-ist-eine-sicherheitsluecke-a-648842/> liegen. In mehreren jüngsten Vorfällen sortierten KI-gestützte Schutzsysteme Alarme falsch ein: echte Angriffe wurden als harmlos abgewertet, während harmlose Ereignisse dagegen so viel Lärm erzeugten, dass das Wichtige unterging. Das Schlimmste dabei: die als ungefährlich eingestuften echten Angriffe. Angreifer nutzen diese Schieflage gezielt aus. Daher sollte KI in der IT-Sicherheit niemals allein entscheiden.

Angriffe auf Unternehmensnetze beginnen selten mit einem Paukenschlag. Im dokumentierten Fall tasteten sich Täter Schritt für Schritt vor, probierten mehrere Anläufe und brachten ein lernendes Schutzsystem schließlich dazu, sie nicht mehr als Gefahr zu erkennen. Als die Alarmroutine schwieg, war der Weg frei: Daten wurden kopiert, Backups geschwächt, Systeme verschlüsselt. Die Chronik des Vorfalls zeigt, wie sich Verteidigungstechnik austricksen lässt und wie eine forensische Auswertung den Tatablauf später lückenlos macht.

Ein öffentlich erreichbarer Schwachpunkt

Der Einstieg gelang über eine verwundbare, öffentlich erreichbare Komponente im Netz. Die Angreifer nutzten eine bekannte Schwachstelle an einer exponierten Firewall <https://www.security-insider.de/was-ist-eine-firewall-a-602870/> und verschafften sich Zugang. Anschließend legten sie Werkzeuge auf dem System ab: eine „Webshell" für dauerhaften Zugriff, ein „Credential Stealer" zum Abgreifen von Zugangsdaten und das Scan-Tool Nmap zur Netzwerkerkundung. Mit den erbeuteten VPN <https://www.security-insider.de/was-ist-vpn-a-569214/> -Zugangsdaten bauten sie per Windows-Fernzugriff (Remote Desktop Protocol) eine Sitzung zum zentralen Windows-Server (Domain <https://www.security-insider.de/was-ist-eine-domaene-netzwerkdomaene-a-3dfa66c7523e71ee437d253a1fb011dd/> Controller) auf, der Konten und Rechte verwaltet. Damit war die Schaltstelle erreicht, an der sich ein Netzwerk effizient übernehmen lässt.

Die ausgetrickste Abwehr

Auf dem Domain Controller <https://www.security-insider.de/was-ist-ein-domaenencontroller-a-1021654/> stoppte zunächst ein EDR <https://www.security-insider.de/was-ist-endpoint-detection-and-response-edr-a-1044268/> -System („Endpoint Detection and Response") die Schaddatei. Doch der Gegner probierte es stur erneut mit gleichem Inhalt, aber neuen Dateinamen wieder und wieder. Dies ist ein typisches Evasion-Vorgehen gegen ML-gestützte Erkennung, bei dem ein bekannter Effekt eintritt: Sensoren passen ihre Baseline an, wenn ein Muster oft genug ohne unmittelbare Folgeschäden auftaucht. Was gestern „auffällig" war, wirkt morgen wie „gewöhnlich". Forschung und Leitlinien warnen daher nicht nur vor adversarialen Angriffen auf Modelle, sondern auch vor Fehleinschätzungen durch Auto-Tuning und unkritisches Baselining einzelner Quellen. Nach mehreren Versuchen stufte das lernende Modell das Muster als unauffällig ein und ließ die Datei passieren. Dieser praktische Missbrauch von „Adversarial Machine Learning" zeigt, dass nicht nur Schwachstellen gefährlich sind, sondern auch das Training und die Grenzwerte eines KI-basierten Schutzes. Fünf Stunden später begann die Ransomware <https://www.security-insider.de/was-ist-ransomware-a-781385/> , sich zu verteilen. Hier zeigt sich das Grundproblem: Ein System, das ohne menschliche Kontrolle seine „Normalität" aus Beobachtungen ableitet, kann antrainierte Täuschungen übernehmen. Genau für solche Grenzfälle braucht es geübte Analysten, die Alarme querprüfen, Korrelationen bilden und Stopp rufen.

Verräterische RDP-Bitmap-Caches

Die anschließende Auswertung förderte einen unscheinbaren, aber entscheidenden Beleg zutage: den RDP <https://www.security-insider.de/was-ist-rdp-a-29113d31a86f31890b562047d9b823d6/> -Bitmap-Cache. Windows speichert dabei Bildschnipsel von Fernsitzungen lokal. Aus diesen „Kacheln" ließ sich später die Aktivität im Rückblick nahezu Bild für Bild rekonstruieren. Wie in einem Daumenkino wurden einfache Befehle wie „ipconfig", aber auch ein „LSASS-Dump" sichtbar, das heißt eine Sicherung des Arbeitsspeichers des Windows-Anmeldeprozesses. Aus diesem Speicher lassen sich mit gängigen Angreifer-Tools (z.B. Mimikatz <https://www.security-insider.de/was-ist-mimikatz-a-851187/>) Anmeldedaten auslesen, die es den Angreifenden erlauben, sich im Netzwerk weiterzubewegen und Rechte zu erhöhen. So wurde nachvollziehbar, wie sich die Angreifer vom Domänenkonto aus in Backups und Gruppenrichtlinien <https://www.security-insider.de/mit-gruppenrichtlinien-in-windows-arbeiten-a-555910/> vorarbeiteten.

Vom Datensammeln zur Erpressung

Kurz nach Mitternacht begann die systematische Datensammlung. Der Täter packte Daten teils mit bis zu 32 GB pro Datei in RAR-Archive („User.rar“, „Management.rar“, „Service_SAP.rar“ und andere). Der Abtransport lief über WinSCP, ein Datei-Übertragungsprogramm, das samt Konfiguration temporär auf die Zielrechner kopiert und danach wieder entfernt wurde. Diese nüchterne Logistik verrät Planung: Wer weiß, wo Wert liegt, archiviert gezielt und räumt anschließend auf.

Gegen 4.00 Uhr morgens startete die vom Domain Controller gesteuerte Verschlüsselung <https://www.security-insider.de/was-ist-verschluesselung-a-618734/> in Wellen, die bis circa 6.30 Uhr dauerten. Die Ransomware wurde per SMB-Dateifreigaben („Server Message Block“) verteilt und mit PsExec als Dienst auf vielen Rechnern gestartet. Minuten später erschienen die bekannten Erpressernotizen („readme.txt“). Bis zur Entdeckung und Unterbrechung verging wertvolle Zeit, in der jede Minute mehr Dateien unbrauchbar machte.

Jetzt Newsletter abonnieren

Täglich die wichtigsten Infos zur IT-Sicherheit

Geschäftliche E-Mail

Mit Klick auf „Newsletter abonnieren“ erkläre ich mich mit der Verarbeitung und Nutzung meiner Daten **gemäß Einwilligungserklärung (bitte aufklappen für Details)** einverstanden und akzeptiere die Nutzungsbedingungen. Weitere Informationen finde ich in unserer Datenschutzerklärung. Die Einwilligungserklärung bezieht sich u. a. auf die Zusendung von redaktionellen Newslettern per E-Mail und auf den Datenabgleich zu Marketingzwecken mit ausgewählten Werbepartnern (z. B. LinkedIn, Google, Meta).

Aufklappen für Details zu Ihrer Einwilligung

Warum Technik allein nicht reicht

Der Fall belegt ein strukturelles Problem: Lernende Systeme sind stark – und formbar. Wird die Grenze zwischen „auffällig“ und „normal“ aber durch konsequentes Testen verschoben, können Fehlentscheidungen provoziert werden. Das ist kein Grund gegen KI, sondern ein Argument für ihren kontrollierten Einsatz: Modelle brauchen menschliche Überprüfung, klare Eskalationsregeln und unabhängige Zusatzsignale, die sich nicht alle gleichzeitig täuschen lassen. Genau darauf zielen aktuelle Empfehlungen aus der Praxis: vorbereiten, erkennen und unterbrechen.

Sofortige Gegenmaßnahmen – in drei Schritten

Vorbeugen beginnt außen, bei exponierten Diensten, die gehärtet und aktuell gehalten werden müssen. Fernzugriffe <https://www.security-insider.de/was-ist-fernzugriff-a-1079774/> gehören streng abgesichert und Netzsegmente sauber getrennt. Ebenso unverzichtbar ist eine Multifaktor-Autorisierung <https://www.security-insider.de/was-ist-authentifizierung-a-47c6dc273418ad29315eedba3a8097ca/> . So schrumpft die Angriffsfläche und mit ihr die Zahl der Wege ins Netz.

Erkennen heißt, auch schwache Signale zusammenzuführen. Daher empfiehlt es sich, Telemetrie aus EDR, Firewall, Verzeichnisdienst und weiteren Sensoren in einem Security Operations Center (SOC <https://www.security-insider.de/was-ist-ein-security-operations-center-soc-a-617980/>) rund um die Uhr zu korrelieren. Entscheidend ist die Haltung: Unklare Ereignisse werden nicht glattgebügelt, sondern hochgestuft. Wenn die Erkennung an klaren Techniken ausgerichtet wird, etwa am MITRE-ATT&CK-Katalog, der typische Angriffsbausteine beschreibt, wird das Monitoring überprüfbar und messbar. Beispiele aus diesem Fall sind unter anderem: „Exploit <https://www.security-insider.de/was-ist-ein-exploit-a-618629/> Public-Facing Application“ (Erstzugriff über exponierte Anwendung) und „OS Credential Dumping: LSASS Memory“ (Auslesen von Anmeldedaten aus dem Arbeitsspeicher).

Reagieren will geübt sein. Incident-Response-Pläne legen fest, wer bei Verdacht entscheidet, wie Systeme isoliert und in welcher Reihenfolge wiederhergestellt werden. Dazu gehören Kommunikationswege nach innen und außen, einschließlich der Frage, welche Daten mit Behörden und Dienstleistern geteilt werden dürfen. Das regelmäßige Durchspielen dieser Abläufe in Übungen verkürzt im Ernstfall die kritischsten Stunden.

Unternehmen, die eine gut vorbereitete Abwehr und eingespielte Abläufe vorweisen, nehmen den Erpressern den Taktstock aus den Händen. Anders ausgedrückt: Wenn Außenstellen gehärtet sind, die Erkennung an klaren Mustern ausgerichtet ist und die Reaktionswege sitzen, läuft ein Angriff nicht mehr nach Drehbuch der Täter, sondern nach den Regeln der Verteidigung.

Handlungsempfehlung: 24/7-CyberSOC mit Maß

Ausschlaggebend für den erfolgreichen Angriff war in diesem Fall nicht die fehlende Technik, sondern die fehlende Einordnung. Ein dauerhaft besetztes Security Operations Center (SOC) hätte die ungewöhnliche Folge von Ereignissen zusammengeführt: wiederholte Upload-Versuche bei gleichbleibendem Inhalt, ein erst geblockter und später erlaubter Ablauf, auffällige RDP-Aktivität am Domain Controller, Veränderungen an Backup <https://www.security-insider.de/was-ist-backup-a-954035/> - und Richtlinieneinstellungen. Ein SOC prüft solche Ketten, bewertet sie im Kontext und stoppt verdächtige Prozesse – bevor aus einem Einzelfehler der KI eine Kettenreaktion wird.

Um Angriffe dieser Art früh zu bremsen, muss das SOC-Sicherheitskonzept natürlich zur IT-Landschaft und den individuellen Risiken des Unternehmens passen. Das Kernelement ist die Überwachung von AV- und EDR-Signalen im Verbund mit weiteren Quellen, zum Beispiel Firewall-Logs, Verzeichnisdienst-Ereignisse und anderer Sensorik. Wo sich Muster wiederholen, wo geblockte Artefakte plötzlich durchgehen oder wo privilegierte Systeme ungewöhnlich handeln, muss automatisch eskaliert und manuell verifiziert werden. Das SOC legt dabei Schwellenwerte fest, die nicht „umerziehbar“ sind, dokumentiert Eingriffe, übt Isolations- und Wiederanlaufverfahren und arbeitet mit klaren Entscheidungswegen. So entsteht aus vielen Einzelalarmen ein Bild – und aus dem Bild die richtige Entscheidung in der richtigen Minute. Die Voraussetzung hierfür sind Menschen, die Muster prüfen, Zweifelsfälle eskalieren und Eingriffe freigeben.

Über den Autor: Markus Neumaier ist Head of CyberSOC Central EU bei Orange Cyberdefense Deutschland <https://www.orangecyberdefense.com/de/> .

(ID:50691971)