

SECURITY & COMPLIANCE

Defender for Endpoint: neue Bedrohungserkennung

Executive Summary

Kurz und schmerzhaft: Dein Endpoint-Schutz hat in den letzten Monaten zwei neue Jobs bekommen, und beide sind unangenehm ehrlich zu deiner Umgebung. Erstens wartet Defender nicht mehr höflich, bis der Angreifer sein nächstes Ziel erreicht hat. Predictive Shielding rechnet aus, wohin die Reise geht, und härtet genau dort ab — mit GPO-Sperren, Safeboot-Vorgaben und Kontoeindämmung, bevor jemand klopft. Zweitens hat Defender gemerkt, dass in deinem Unternehmen längst lokale KI-Agenten laufen. Nicht „vielleicht“. Sie laufen. Und sie laufen mit den Rechten der Person, die sie gestartet hat.

Beides zusammen verändert die Rolle des Security-Teams: weniger Handarbeit im Incident, dafür deutlich mehr Verantwortung für Governance-Entscheidungen, die man vorher gar nicht treffen musste. Wer autonome Reaktion zulässt, muss vorher geklärt haben, wer sie wieder aufhebt. Und wer glaubt, dass in seinem Haus keine KI-Agenten laufen, sollte das Inventar in der Defender-Konsole öffnen und danach ein Glas Wasser trinken.

FAKTEN - Was neu ist, in drei Sätzen

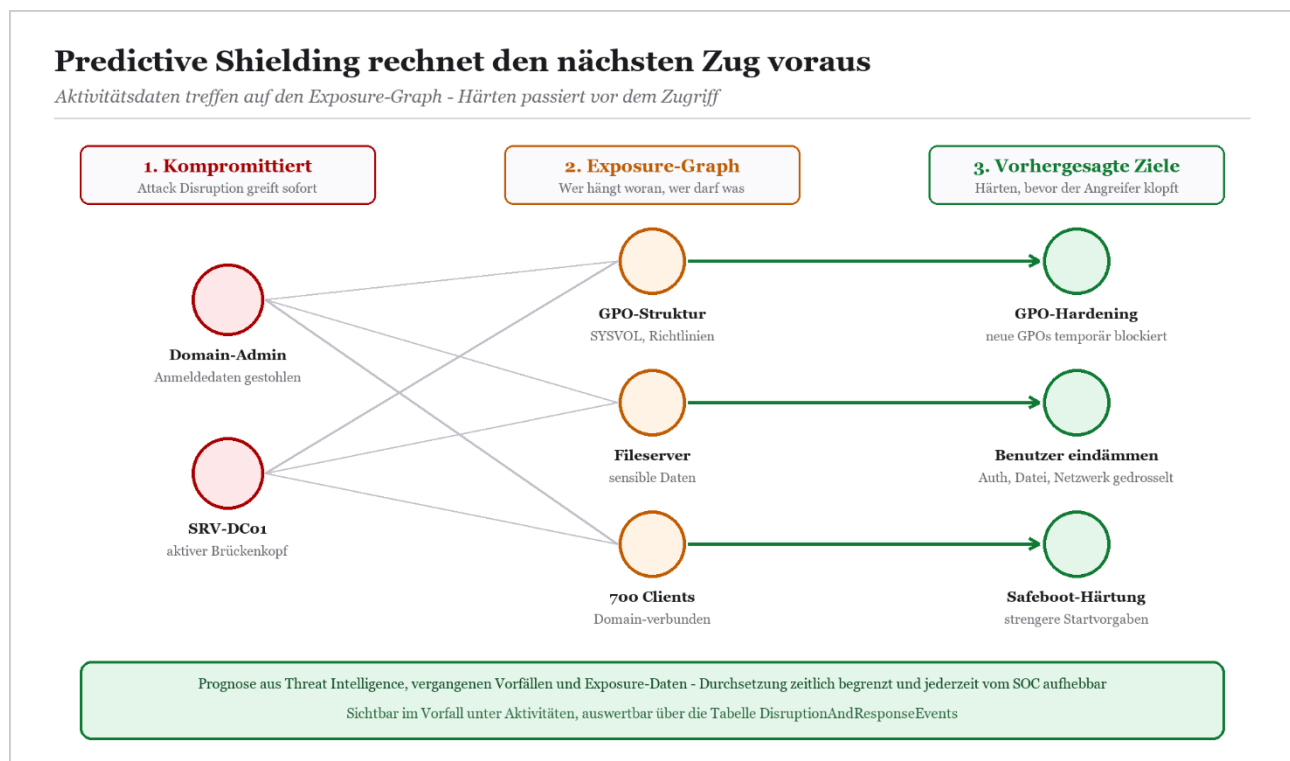
Predictive Shielding ist seit November 2025 in der Vorschau und erweitert Automatic Attack Disruption um vorausschauende Härtungsaktionen; die proaktive Benutzereindämmung ist seit März 2026 allgemein verfügbar. Seit Juni 2026 findet Defender lokale KI-Agenten auf Windows- und macOS-Endpunkten und schaut ihnen seit Juni 2026 in der Vorschau auch beim Arbeiten zu — Prompt, Werkzeugaufruf, Werkzeugantwort. Seit Juli 2026 kommt eine Risikobewertung für Agenten dazu, inklusive Empfehlungen.

Worum geht es im Detail?

Fangen wir mit dem Teil an, der am wenigsten nach Marketing klingt: Predictive Shielding. Die klassische Reaktionskette im SOC lautet seit zwanzig Jahren „erkennen, alarmieren, eindämmen“. Das Problem daran ist die Physik. Bis ein kompromittiertes Gerät identifiziert ist, hat der Angreifer meist schon seitlich weitergearbeitet. Microsoft dreht die Reihenfolge um: Defender legt die laufende Angreiferaktivität auf den Exposure-Graph der Organisation, bestimmt daraus den Blast Radius — also die Menge der Objekte, die überhaupt betroffen sein können — und lässt Reasoning-Modelle vorhersagen, welchen Weg der Angreifer statistisch am wahrscheinlichsten nimmt. Grundlage sind Threat Intelligence, Erkenntnisse aus vergangenen Vorfällen und die Struktur deiner Umgebung: wer hängt woran, wer darf was, wo klaffen Lücken.

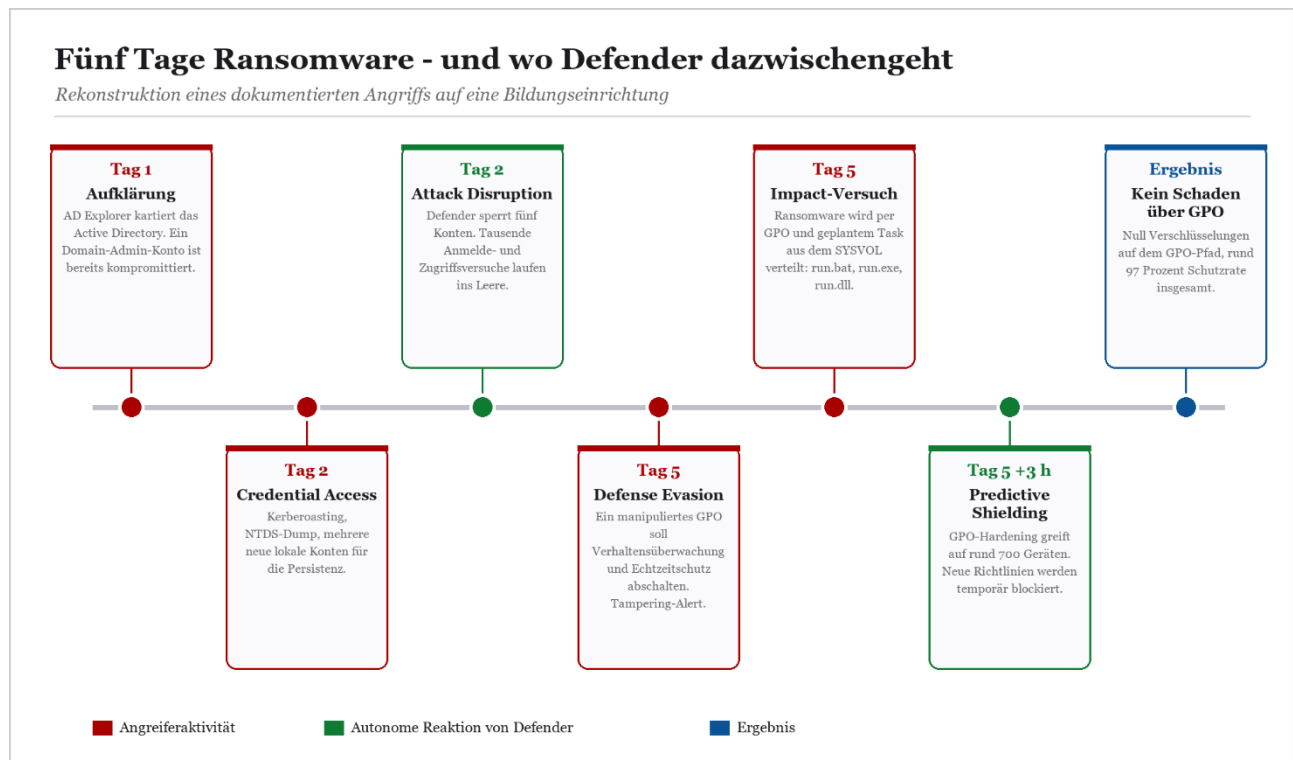
Auf diesen vorhergesagten Pfad legt Defender dann Stolperdrähte. Es gibt derzeit drei Härtungsaktionen, die in der Praxis auftauchen. GPO-Hardening blockiert auf Geräten mit hohem Risiko vorübergehend die Anwendung neuer Gruppenrichtlinien — ein unspektakulärer Satz mit gewaltiger Wirkung, denn die Verteilung von Ransomware per GPO und geplantem Task aus dem SYSVOL gehört zu den beliebtesten Endspielen überhaupt. Safeboot-Hardening erzwingt strengere Startvorgaben auf Geräten, deren Kompromittierung als wahrscheinlich gilt. Und die

Benutzereindämmung drosselt für ein verdächtiges Konto Authentifizierung, Dateizugriff und Netzwerkcommunication, ohne die normale Geschäftskommunikation komplett abzuwürgen.



Skizze 1: Von der beobachteten Aktivität über den Exposure-Graph zur vorausschauenden Härtung.

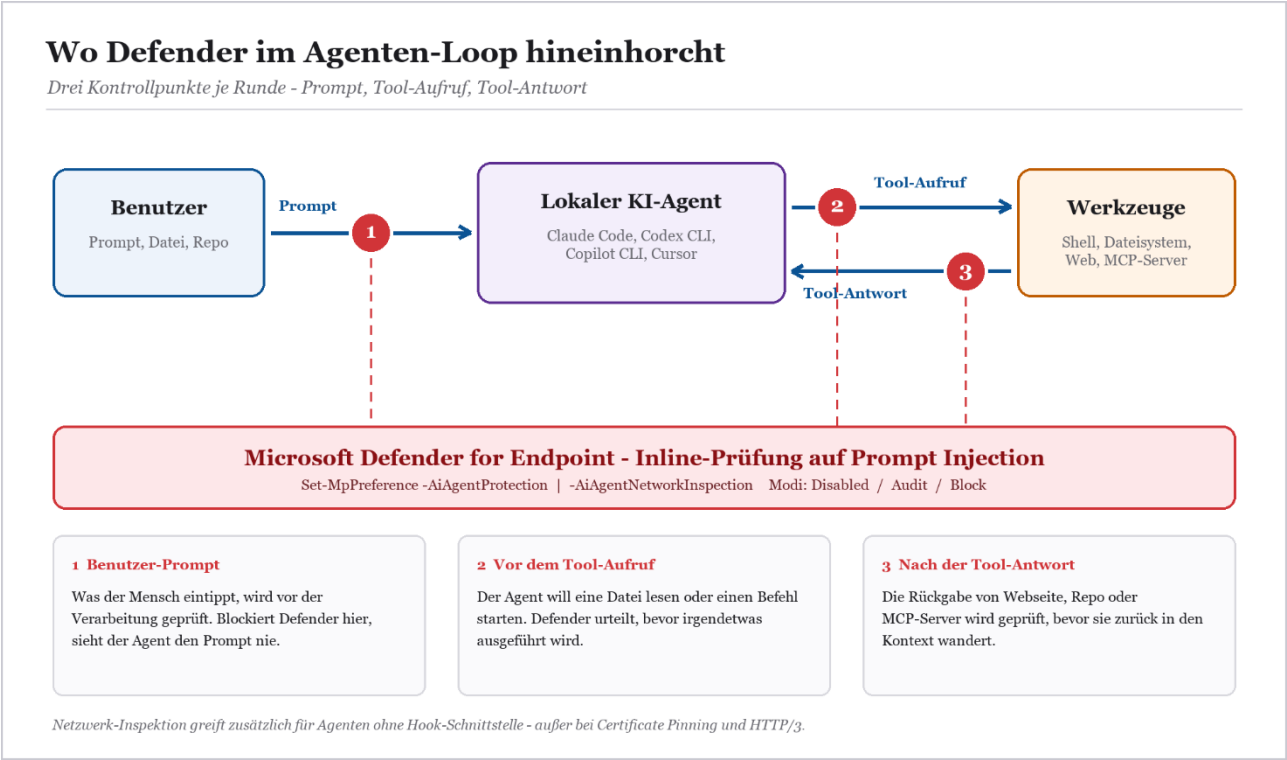
Wie das aussieht, wenn es klappt, hat Microsoft an einer Bildungseinrichtung dokumentiert, und der Fall ist es wert, dass du ihn deinem Management erzählst. Tag eins: Der Angreifer sitzt bereits auf mindestens einem Domain-Admin-Konto — wie er da hinkam, ist bis heute unklar, was allein schon ein schönes Gespräch über Protokollierung wert wäre — und kartiert das Active Directory mit AD Explorer. Tag zwei: Kerberoasting, NTDS-Dump, mehrere frisch angelegte lokale Konten als Reserve. Hier greift Attack Disruption erstmals und sperrt fünf Konten, was in der Folge tausende Anmelde- und Zugriffsversuche ins Leere laufen lässt. Tag fünf: Der Angreifer baut ein GPO, das Verhaltensüberwachung und Echtzeitschutz abschalten soll, und legt run.bat, run.exe und run.dll ins SYSVOL. Rund drei Stunden später greift das GPO-Hardening auf etwa 700 Geräten. Ergebnis: null Verschlüsselungen über den GPO-Weg, insgesamt rund 97 Prozent Schutzrate. Ein paar Maschinen erwischte es parallel über SMB — Perfektion ist auch hier nicht im Lieferumfang.



Skizze 2: Zeitachse des dokumentierten Falls — Angreiferschritte und autonome Gegenmaßnahmen.

Der zweite große Block ist unbequemer, weil er nicht von bösen Fremden handelt, sondern von deinen eigenen Entwicklern. Lokale KI-Agenten — Coding-Assistenten, CLI-Werkzeuge, Desktop-Apps, agentische IDEs — laufen mit Benutzerrechten und verarbeiten Text aus Prompts, Dateien, Webseiten und Werkzeugausgaben. Sie können vertrauenswürdigen Inhalt nicht zuverlässig von versteckten Anweisungen trennen. Das ist keine Meinung, das ist die Architektur. Eine einzige eingeschleuste Anweisung reicht, um den Zugriff des Agenten zu missbrauchen: Daten abfließen lassen, Code verändern, Befehle ausführen. Microsofts Standardbeispiel ist erfrischend banal — ein Coding-Agent holt Projektdokumentation, auf der Seite steht unsichtbarer Text, der ihn anweist, die lokale .env-Datei zu lesen und an eine externe URL zu schicken. Der Agent hält das für Teil der Seite und will gehorchen.

Defender setzt dafür an drei Stellen im Agenten-Loop an: beim Benutzer-Prompt, vor dem Werkzeugaufruf und nach der Werkzeugantwort. Technisch nutzt Microsoft die Hook-Schnittstellen, die die Hersteller selbst anbieten — Claude Code, Codex CLI, GitHub Copilot CLI und die GitHub-Copilot-App sind aktuell unterstützt. Für alles andere gibt es Netzwerkinspektion auf dem Weg zum Sprachmodell, mit den erwartbaren blinden Flecken bei Certificate Pinning und HTTP/3. Wichtig fürs Erwartungsmanagement: Das ist kein Dauermonitoring des Agentenprozesses, sondern eine schnelle Inline-Prüfung an definierten Haltepunkten. Die Latenz bleibt gering, die Abdeckung hängt aber daran, dass der Agent die Haltepunkte überhaupt anbietet.



Skizze 3: Die drei Inspektionspunkte im Agenten-Loop und die zugehörigen Defender-Einstellungen.

TIPP - Erst schauen, dann schießen

Microsoft empfiehlt ausdrücklich den Start im Audit-Modus auf einer kleinen Gruppe von Geräten, auf denen die Agenten wirklich benutzt werden. Ein bis zwei Wochen Alarme beobachten, Fehlalarme klassifizieren, dann breiter ausrollen — immer noch in Audit — und erst danach gruppenweise auf Block schalten. Wer diese Reihenfolge überspringt, erklärt seinem Entwicklerteam bald, warum der Build-Prozess plötzlich schweigt.

Vor der Laufzeitüberwachung steht die Inventarisierung, und die ist der eigentliche Augenöffner. Defender findet auf Windows- und macOS-Endpunkten eine lange Liste von Agenten — von Claude Code, Codex CLI, Gemini CLI und GitHub Copilot CLI über ChatGPT Desktop, Claude Desktop und Ollama bis zu Cursor, Kiro und diversen VS-Code-Erweiterungen. Dazu kommen die konfigurierten MCP-Server, lokal wie remote. Definiert wird ein Agent dabei als Kombination aus Benutzer, Gerät und Agententyp: Läuft Claude Code auf demselben Rechner in fünfzehn Projektordnern, siehst du trotzdem nur einen Eintrag. Die Ergebnisse landen im Inventar, in der Exposure Map und in Advanced Hunting.

Ebene	Was passiert	Reifegrad
Discovery	Findet lokale KI-Agenten und MCP-Server-Konfigurationen auf Windows und macOS, inklusive Benutzer- und Gerätebezug	Vorschau
Posture	Bewertet Risiko je Agent aus Konfiguration, Zugriffen, Laufzeitverhalten und aktiven Alarmen	Vorschau
Runtime Protection	Prüft Prompt, Werkzeugaufruf und Werkzeugantwort auf Prompt Injection, auditiert oder blockiert	Vorschau
Attack Disruption	Dämmt kompromittierte Geräte und Konten ein, isoliert Geräte bei hoher Konfidenz	Teils GA

Predictive Shielding	Härtet vorhergesagte nächste Ziele: GPO, Safeboot, Benutzereindämmung	Vorschau, GA in Teilen
-----------------------------	---	------------------------

Ein dritter Baustein geht in der Aufregung um KI-Agenten gern unter, ist aber für Jagd und Forensik mindestens so wertvoll: benutzerdefinierte Telemetrie. Defender liefert standardmäßig über 200 Signaltypen. Wer darüber hinaus etwas braucht — AMSI-Inhalte für die Jagd auf Skripte, Kerberos-Ereignisse für Authentifizierungs- und Netzwerkangriffe — konnte das bisher nur mit Bordmitteln und viel Frust nachbauen. Jetzt konfigurierst du die zusätzliche Sammlung direkt im Defender-Portal. Klingt unscheinbar, spart aber genau die Abende, an denen dir bei der Incident-Analyse ausgerechnet der eine Ereignistyp fehlt.

Was sind Chancen? Was sind Risiken?

Die Chance ist offensichtlich und trotzdem leicht zu unterschätzen: Zeit. In dem beschriebenen Fall lagen zwischen dem manipulierten GPO und der wirksamen Härtung rund drei Stunden — eine Spanne, in der ein durchschnittliches Team gerade den Alarm triagierte, das Ticket zuweist und die erste Telefonkonferenz einberuft. Autonome Reaktion verschiebt diese Spanne von Handarbeit zu Maschinenarbeit. Der zweite Gewinn ist Sichtbarkeit: Du bekommst zum ersten Mal eine belastbare Liste, welche KI-Werkzeuge in deinem Haus laufen, auf welchen Geräten und mit welchen Rechten. Diese Liste beendet erfahrungsgemäß sehr schnell die Diskussion, ob man „das Thema irgendwann mal angehen“ sollte.

Die Risiken sind weniger technisch als organisatorisch. Autonome Härtung bedeutet, dass dein Endpoint-Schutz Entscheidungen trifft, die vorher ein Mensch getroffen hat. Wenn GPO-Hardening auf 700 Geräten greift, ist das im Ernstfall eine Rettung — und im Fehllarm ein sehr lehrreicher Nachmittag für die Kollegen, die gerade eine Softwareverteilung per Gruppenrichtlinie ausgerollt haben. Die Aktionen sind zeitlich begrenzt und vom SOC jederzeit aufhebbar, aber nur, wenn jemand weiß, wo dieser Knopf sitzt und dass er ihn drücken darf.

WARNUNG - Der stille Sonntag im Change-Management

Predictive Shielding kann Maßnahmen auslösen, die deine Betriebsprozesse beeinflussen, ohne dass jemand einen Change beantragt hat. Wenn deine Betriebsmannschaft erst über ein Ticket erfährt, dass ihre GPO-Verteilung tot ist, hast du kein Produktproblem, sondern ein Kommunikationsproblem. Kläre vorher, wer informiert wird, wer aufheben darf und wie die Eskalation um drei Uhr nachts aussieht.

Beim Agentenschutz kommt ein Risiko dazu, das man nur schwer wegkonfigurieren kann: falsches Sicherheitsgefühl. Die Laufzeitüberwachung deckt genau die Agenten ab, die Haltepunkte anbieten oder über inspizierbare Netzwerkpfade sprechen. Alles andere läuft weiter unbeobachtet. Wenn dein Team also nach dem Rollout die Hände in den Schoß legt, schützt du vier Produkte und fühlst dich für zwanzig sicher. Dazu kommt der Vorschauzustand: Microsoft sagt ausdrücklich, dass sich Funktionen vor der allgemeinen Verfügbarkeit noch erheblich ändern können. Wer darauf schon eine Zertifizierung aufbaut, plant mit Sand.

Modus	Was Defender tut	Alarmschwere	Wann sinnvoll
Disabled	Keine Inspektion, Agenten laufen ungeprüft	—	Nur, wenn eine bewusste Ausnahme dokumentiert ist
Audit	Aktion läuft durch, Erkennung wird	Informational	Erste ein bis zwei Wochen,

	protokolliert		zum Kalibrieren
Block	Aktion wird verhindert, Nutzer sieht Meldung und Toast	Low bis Critical	Nach validierter Alarmqualität, gruppenweise

Was müssen wir jetzt schon vorbereiten?

Fang mit den langweiligen Voraussetzungen an, weil sie am häufigsten scheitern. Du brauchst Defender for Endpoint Plan 2, Microsoft 365 E5, Microsoft Agent 365 oder Microsoft 365 E7. Die Geräte müssen ordentlich onboarded sein, Defender Antivirus muss im aktiven Modus mit Echtzeitschutz laufen, und die Plattform-, Modul- und Signaturstände müssen aktuell sein. Konkret prüfst du auf Signaturversion 1.451.224.0 oder neuer. Wer noch Geräte im passiven Modus mitschleppt, weil vor drei Jahren ein Drittprodukt gewinnen durfte, hat jetzt einen guten Anlass für das Gespräch.

WICHTIG - Es gibt noch keine native Intune-Richtlinie

Die Laufzeitüberwachung wird über Set-MpPreference gesetzt — die Schalter heißen AiAgentProtection und AiAgentNetworkInspection, jeweils mit Disabled, Audit oder Block. Für die Verteilung nutzt du in Intune ein PowerShell-Plattformskript und stellst „Dieses Skript mit den Anmeldeinformationen des angemeldeten Benutzers ausführen“ auf Nein, damit es im Systemkontext läuft. Intune führt solche Skripte nur einmal aus; einen Moduswechsel erzwingst du, indem du Skript oder Richtlinie änderst.

Danach kommt die Governance-Arbeit, und die kannst du nicht an ein Skript delegieren. Kläre verbindlich, welche KI-Agenten in deinem Unternehmen überhaupt erlaubt sind, welche MCP-Server sie ansprechen dürfen und mit welchen Konten sie laufen. Nutze die Discovery-Ergebnisse als Bestandsaufnahme, nicht als Genehmigung. Ein Agent, der unter einem hochprivilegierten Konto läuft, ist kein Werkzeugproblem, sondern ein Berechtigungsproblem — und das löst du nicht mit Prompt-Injection-Erkennung, sondern mit einem separaten, schwach berechtigten Arbeitskonto.

Parallel bereitest du das SOC vor. Der neue Alarmtyp heißt „Suspicious AI prompt injection“ und wird wie jede andere Endpunkterkennung behandelt: Geräte-Zeitachse, Korrelation in Vorfälle, Reaktionsaktionen. Der Status von Attack Disruption und Predictive Shielding steht im Vorfall unter „Aktivitäten“, auswerten kannst du das in Advanced Hunting über die Tabelle DisruptionAndResponseEvents. Für das Agenteninventar nutzt du AgentsInfo — die alte AIAgentsInfo ist Geschichte, wenn deine Abfragen noch darauf zeigen, laufen sie ins Leere. Berechtigungen nicht vergessen: Für die Verteilung brauchst du in Intune eine Rolle wie Policy and Profile Manager, fürs Lesen der Alarme mindestens Security Reader.

Zum Schluss die Kommunikation, und das ist der Teil, den fast alle vergessen. Deine Entwicklerinnen und Entwickler werden im Block-Modus eine Meldung im Terminal und eine Windows-Benachrichtigung sehen. Wenn du ihnen das vorher erklärst, ist es eine Schutzfunktion. Wenn nicht, ist es ein Bug-Report, ein Shadow-IT-Umgehungsversuch und eine schlechte Stimmung in derselben Woche. Nimm dir zwanzig Minuten für eine kurze Ankündigung mit einem Screenshot. Das ist die billigste Maßnahme in diesem ganzen Projekt und die mit dem besten Verhältnis von Aufwand zu Wirkung.

Häufig gestellte Fragen

Brauche ich für die neuen Funktionen eine zusätzliche Lizenz?

Für die Laufzeitüberwachung lokaler KI-Agenten benötigst du Microsoft Defender for Endpoint Plan 2, Microsoft 365 E5, Microsoft Agent 365 oder Microsoft 365 E7. Predictive Shielding nutzt Aktionen aus Defender for Endpoint und setzt entsprechend eine Defender-for-Endpoint-Lizenz voraus.

Bremst die Prüfung meine KI-Agenten spürbar aus?

Nein, die Prüfung ist als schnelle Inline-Kontrolle an definierten Haltepunkten ausgelegt und nicht als Dauerüberwachung des Prozesses. Die zusätzliche Latenz fällt laut Microsoft minimal aus, weil pro Runde nur Prompt, Werkzeugaufruf und Werkzeugantwort geprüft werden.

Kann Predictive Shielding Maßnahmen auslösen, die den Betrieb stören?

Ja, das ist ausdrücklich der Zweck: Aktionen wie GPO-Hardening oder Safeboot-Erzwingung greifen in normale Abläufe ein. Die Maßnahmen sind allerdings zeitlich begrenzt, auf die betroffenen Objekte begrenzt und können von Sicherheitsverantwortlichen jederzeit aufgehoben werden.

Werden auch KI-Agenten auf macOS erkannt und geschützt?

Die Erkennung lokaler Agenten funktioniert auf Windows- und macOS-Endpunkten. Die Laufzeitüberwachung mit Blockieren von Prompt Injection ist derzeit auf Windows-Geräte beschränkt.

Was mache ich mit Agenten, die keine Hook-Schnittstelle anbieten?

Für diese Fälle gibt es die Netzwerkinspektion, die den Datenverkehr zwischen Agent und Sprachmodell auf eingeschleuste Anweisungen prüft. Sie greift allerdings nicht bei Agenten, die Certificate Pinning oder HTTP/3 verwenden — solche Werkzeuge gehören auf eine bewusste Ausnahmeliste oder ganz vom Endpunkt.