Sophos Red Team-Experiment:

Wenn die KI den Hacker spielt

Drei Stunden. 23 Lücken. Kein Mensch hätte das geschafft.

Das Cybersicherheitsunternehmen Sophos hat OpenClaw, den KI-Agenten, der die Branche im Sturm erobert hat, damit beauftragt, eines der eigenen internen Netzwerke anzugreifen – vollständig autonom, mit echten Hacking-Werkzeugen, auf einem produktiven Legacy-System.

Die Ergebnisse übertrafen alle Erwartungen.

Warum dieser Selbstversuch?
Sophos' Red Team – eine interne Gruppe von Sicherheitsexperten, die das Unternehmen aus der Perspektive eines Angreifers testen – hatte sich eine klare Aufgabe gestellt: Sie wollten herausfinden, ob ein KI-Agent sinnvoll in einen klassischen Penetrationstest integriert werden kann. Dafür wählten sie ein älteres, internes Netzwerk, das seit einiger Zeit nicht mehr aktiv getestet worden war. Es versprach genug Angriffsfläche, um dem Agenten faire Chancen zu geben, und war gleichzeitig ausreichend isoliert, um die kritischen Kernsysteme des Unternehmens zu schützen.

Sicherheit zuerst – aber wie?
Der Großteil der Vorbereitungszeit floss nicht in die Konfiguration des Angriffs, sondern in die Entwicklung von Sicherheitsregeln für den Agenten selbst. Das Team musste verhindern, dass OpenClaw im Eifer seiner Zielverfolgung unbeabsichtigt Schaden anrichtet – etwa indem es Daten löscht, sensible Informationen nach außen überträgt oder im schlimmsten Fall das gesamte Netzwerk verschlüsselt. Ein selbst verursachter Ransomware-Angriff wäre, wie das Team trocken anmerkt, ein „nicht optimales Ergebnis" gewesen.

Um das zu verhindern, entwickelte Sophos eigene, maßgeschneiderte Werkzeuge und Verfahren, anstatt auf öffentlich verfügbare, oft schlecht dokumentierte Alternativen zurückzugreifen. Zudem wurde ein Freigabemechanismus eingebaut: Für bestimmte Aktionen musste ein Mensch ausdrücklich zustimmen, bevor der Agent fortfahren durfte. Dieses Gleichgewicht aus Autonomie und Kontrolle erwies sich als einer der Schlüssel zum Erfolg.

Was der Agent leistete
OpenClaw hielt sich während des gesamten Tests an die vorgegebenen Grenzen – kein einziger unbeabsichtigter Vorfall. Gleichzeitig arbeitete er mit einer Geschwindigkeit und Gründlichkeit, die menschliche Tester schlicht nicht erreichen können: Die Netzwerk-Aufklärungsphase, für die ein Team üblicherweise drei volle Tage benötigt, war in drei Stunden abgeschlossen. Am Ende des Tests wurden 23 konkrete, verwertbare Sicherheitslücken identifiziert.

Besonders beeindruckt war das Team von der Kreativität des Agenten. Als ein vielversprechender Angriffspfad blockiert war, schlug OpenClaw eigenständig vor, eine Cloud-Instanz zu starten, um einen erbeuteten Passwort-Hash zu knacken – und setzte das nach menschlicher Freigabe auch um. Darüber hinaus protokollierte der Agent jeden einzelnen Schritt in einem Detailgrad, der manuell schlicht nicht erreichbar wäre. Das vereinfachte das abschließende Reporting erheblich.

Die gewonnenen Erkenntnisse konnten zudem fast unmittelbar in weiterführende Sicherheitsmaßnahmen einfließen – darunter die Validierung von Erkennungsmechanismen

und gemeinsame Übungen von Angriffs- und Verteidigungsteams, was Offensive und Defensive nahezu in Echtzeit zusammenbrachte.

Was das für die Cybersicherheit bedeutet
Ross McKerchar, CISO bei Sophos, zieht ein klares Fazit: KI-gestützte Angriffswerkzeuge sind mächtig – und potenziell gefährlich. Aber sie zu ignorieren sei keine Option. Die Welt entwickelt sich weiter, und wer diese Technologien nicht versteht und verantwortungsvoll einsetze, riskiere, gegenüber jenen ins Hintertreffen zu geraten, die es tun – einschließlich der Angreifer.

McKerchar sieht Cybersicherheitsteams dabei in einer besonders guten Ausgangsposition: Der tägliche Umgang mit gefährlichen Werkzeugen und komplexen Bedrohungsszenarien schärfe genau jenes Denken, das für einen verantwortungsvollen Umgang mit KI-Agenten notwendig sei. Je mehr praktische Erfahrung Sicherheitsfachleute in diesem Bereich sammelten, desto besser würden sie in der Lage sein, die richtigen Kontrollpunkte zu identifizieren – und zu verstehen, wie echtes Risikomanagement im KI-Zeitalter aussieht.

Den vollständigen Bericht, einschließlich aller technischen Details, des verwendeten System-Prompts und der vollständigen Liste der Schwachstellen, hat Sophos auf seinem Blog https://www.sophos.com/en-us/blog/we-let-openclaw-loose-on-an-internal-network-heres-what-it-found sowie auf GitHub veröffentlicht.

Social Media von Sophos für die Presse
Wir haben speziell für Sie als Journalist*in unsere Social-Media-Kanäle angepasst und aufgebaut. Hier tauschen wir uns gerne mit Ihnen aus. Wir bieten Ihnen Statements, Beiträge und Meinungen zu aktuellen Themen und natürlich den direkten Kontakt zu den Sophos Security-Spezialisten.

Folgen Sie uns auf und

LinkedIn: https://www.linkedin.com/groups/9054356/
X/Twitter: @sophos_info

Pressekontakt:

TC Communications
Arno Lücht, +49-8081-954619
Thilo Christ, +49-8081-954617
Ulrike Masztalerz, +49-30-55248198
Ariane Wendt +49-172-4536839
sophos@tc-communications.de