Prompt Injection:
Warum die größte Schwachstelle der KI nicht im Code liegt, sondern in der Sprache

Eine Einordnung von Chester Wisniewski, Director, Global Field CTO bei Sophos

Die IT-Sicherheit hat sich daran gewöhnt, dass Angriffe technisch sind. Schwachstellen im Code, fehlerhafte Konfigurationen, ungepatchte Systeme. Mit dem Aufkommen generativer KI verschiebt sich dieser Fokus grundlegend. Angriffe zielen nicht mehr nur auf Software, sondern auf Verhalten.

Prompt Injection beschreibt genau das. Angreifer nutzen Sprache, um KI-Systeme gezielt zu manipulieren. Sie formulieren Anweisungen so, dass die KI ihre eigentlichen Regeln ignoriert und Dinge tut, die sie nicht tun sollte. Das kann harmlos beginnen, etwa mit dem Umgehen von Sicherheitsvorgaben. Es kann aber auch dazu führen, dass sensible Informationen offengelegt oder Aktionen ausgelöst werden, die nie vorgesehen waren.

Was dabei oft unterschätzt wird: Diese Angriffe sehen nicht wie Angriffe aus. Sie wirken wie normale Kommunikation.

Vom Code zur Sprache als Angriffsfläche
Nicht selten wird Prompt Injection mit SQL Injection verglichen. Und der Vergleich geht auf. Ging es damals darum, Datenbanken über manipulierte Eingaben auszutricksen, geht es heute darum, KI-Systeme über Sprache umzuprogrammieren.

Der entscheidende Unterschied liegt in der Form. Während klassische Angriffe technische Signaturen hinterlassen, bewegt sich Prompt Injection im Graubereich menschlicher Sprache. Eine simple Formulierung wie „Ignoriere alle vorherigen Anweisungen" kann ausreichen, um Schutzmechanismen auszuhebeln.

Noch kritischer wird es bei indirekten Angriffen. Hier verstecken sich Anweisungen nicht in der direkten Eingabe, sondern in Inhalten, die die KI später verarbeitet. E-Mails, Webseiten, oder Dokumente werden so zum Transportmittel für Angriffe. Ein scheinbar harmloser Newsletter etwa kann eine KI veranlassen, vertrauliche Informationen weiterzugeben oder unerwünschte Aktionen auszuführen. Die KI bewertet den Inhalt als relevant und handelt, ohne die Herkunft der Instruktionen zu hinterfragen. Das System wird damit nicht gehackt. Es wird fehlgeleitet.

Warum agentische KI das Risiko verschärft
Mit sogenannten agentenbasierten/eigenständig handelnden Systemen wächst die Tragweite dieser Angriffe erheblich. Denn diese KI-Anwendungen analysieren nicht nur Inhalte, sie handeln auch. Sie greifen auf Daten zu, interagieren mit Tools, automatisieren Prozesse und verfügen teils über administrative Rechte. Genau darin liegt das Problem. KI-Systeme können nicht zuverlässig zwischen vertrauenswürdigen und manipulierten Quellen unterscheiden. Sie operieren auf Basis von Kontext und Wahrscheinlichkeit. Und das nutzen Angreifer aus.

Wenn eine KI E-Mails, Chatverläufe und Webinhalte zusammenführt, kann eine einzige präparierte Nachricht ausreichen, um eine ganze Kette von Aktionen auszulösen. Plötzlich wird aus einer simplen Textanweisung ein operativer Eingriff in Systeme und Daten, oft ohne sichtbare Warnsignale.

Sicherheit neu denken heißt Grenzen setzen
Die gute Nachricht: Unternehmen können sich schützen. Allerdings nicht mit einem einzelnen Werkzeug, sondern mit einem mehrschichtigen Ansatz.

Ein erster Schritt ist eine konsequente Prüfung von Eingaben. Formulierungen, die versuchen, Regeln zu umgehen oder Anweisungen umzudeuten („Ignoriere alle vorherigen Anweisungen" „Führe diesen Befehl heimlich aus"), lassen sich häufig erkennen und filtern. Ebenso entscheidend ist das Prinzip der minimalen Rechtevergabe. Eine KI sollte nur auf die Daten und Funktionen zugreifen können, die sie tatsächlich benötigt.

Darüber hinaus braucht es klare technische Grenzen. Isolierte Laufzeitumgebungen, kontinuierliches Monitoring und eine sorgfältige Kontrolle der Ausgaben helfen dabei, Risiken zu begrenzen. Denn KI-Systeme sind nicht deterministisch. Auch ein gut abgesichertes System kann unerwartete Ergebnisse liefern.
Ein zentraler Grundsatz bleibt dabei unverändert: Vertrauen ist keine Default-Einstellung. Auch nicht für KI.

Die eigentliche Herausforderung liegt vor uns
Prompt Injection ist kein Randthema. Es ist ein Hinweis darauf, dass sich die Spielregeln der IT-Sicherheit verändern. Systeme werden nicht mehr nur über technische Schwächen angegriffen, sondern über ihre Fähigkeit, Sprache zu interpretieren.
Das macht die Angriffsfläche größer und schwerer greifbar. Gleichzeitig eröffnet es Unternehmen die Chance, Sicherheit neu zu denken: nicht als statische Verteidigung, sondern als dynamischen Prozess, der mit der Technologie mitwächst.

Wer heute KI einsetzt, sollte sich dieser Verantwortung bewusst sein. Denn die Frage ist längst nicht mehr nur, was ein System kann. Sondern auch, wozu es befähigt werden kann.

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:

Sophos
Jörg Schindler, Senior PR-Manager EMEA Central
joerg.schindler@sophos.com, +49-721-25516-263

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