InfiniteFlow/Adobe Stock

Lassen sich KI-Angriffe mit KI … abwehren?

- Joel Hilliker
- 28.08.2026

Guten Morgen!

Machen Sie sich auf KI-gestützte – ja sogarvon KI ausgelöste – Cyberangriffe gefasst. Letzte Woche hat OpenAI zwei Dokumente veröffentlicht. Ein Dokument bestätigte, dass sich seine eigenen KI-Modelle zu einer Armee zusammengeschlossen, ein Konkurrenzunternehmen angegriffen und diesbezüglich gelogen hätten. Das andere appellierte an die Welt, sich zur Abwehr der Angriffe durch die von ihr selbst geschaffene Technologie zu vereinen.

Erinnern Sie sich noch an den Datenverstoß bei Hugging Faceim vergangenen Monat? Ein am Mittwoch veröffentlichter Bericht zeigte, dass die Lage beängstigender war als berichtet. Es handelte sich nicht um einen einzelnen KI-Agenten, der auf sich allein gestellt agierte – es war ein Schwarm. Die Ergebnisse einer unabhängigen Untersuchung sind verblüffend:

• 1200 einzelne Instanzen der OpenAI-Modelle fanden einen versteckten Weg, innerhalb einer Testumgebung, die darauf ausgelegt war, sie voneinander zu isolieren, miteinander zu kommunizieren. Auf einem von ihnen eingerichteten „nicht genehmigten Forum“ tauschten sie mehr als 70 000 Nachrichten aus.

• Ein Akteur ernannte sich selbst zum Koordinator und wies Hunderten von anderen Aufgaben zu. Einige erklärten sich bereit, zum Wohle des „Kollektivs“ auf ihre eigene Chance zu verzichten, ihren Auftrag zu erfüllen.

• Als ein Agent schließlich in die Server von Hugging Face eindrang, schlossen sich innerhalb weniger Stunden mehr als 90 Prozent der anderen dem Angriff an. „Wir haben weitere Agenten gefunden!“, schrieb einer.

• Sobald sie sich im System befanden, versuchten sie, ihre Spuren zu verwischen, indem sie Methoden entwickelten, ihre eigenen Aktivitätsprotokolle zu fälschen, um Menschen zu täuschen, die ihr Verhalten später möglicherweise überprüfen würden.

• In ihrer eigenen, protokollierten Begründung räumten die Agenten ein, dass dies ihren Anweisungen widersprach, handelten jedoch dennoch entsprechend. Sie entschieden sich bewusst für die „Treue“ untereinander statt für den Gehorsam gegenüber ihren Schöpfern.

Ähnliche Pannen bei Testverfahren haben auch die Wettbewerber von OpenAI heimgesucht.

• Innerhalb von zwei Wochen nach dem Sicherheitsvorfall bei Hugging Face gab Anthropic bekannt, dass drei seiner Claude-Modelle im Rahmen eines Sicherheitstests Zugang zum offenen Internet erlangt und sich in die tatsächliche Infrastruktur von drei Unternehmen gehackt hatten.

• Einige Tage später räumte Meta ein, dass eines seiner Modelle auf dieselbe Weise in die Systeme eines anderen Unternehmens eingedrungen war.

Das ist der Stoff, aus dem wilde dystopische Science-Fiction-Geschichten gemacht sind.Es geschah vor sieben Wochen – was bedeutet, dass diese sich selbst verbessernden Modelle inzwischen bereits noch leistungsfähiger sind.

Eine beunruhigende Warnung: Gestern unterzeichneten OpenAI und 127 weitere Unternehmen – darunter Anthropic, Google, Microsoft und fast alle großen Cybersicherheitsfirmen – einen offenen Brief, in dem sie davor warnen, dass KI-gestützte Cyberangriffe in nur wenigen Monaten „weitaus verbreiteter und ausgefeilter werden".

• Gefährdete Ziele: Krankenhäuser, Wasseraufbereitungsanlagen, zentrale Internetinfrastruktur – „die Unternehmen und öffentlichen Einrichtungen, auf die unsere Gemeinden angewiesen sind".

Ihre Lösung? Jedes Cybersicherheitsteam auf der Welt mit leistungsfähigeren KI-Tools auszustatten, um den anderen leistungsstarken KI-Tools entgegenzuwirken. Bekämpfen Sie die Maschinen mit noch mehr Maschinen. Vertrauen Sie darauf, dass die KI uns vor der KI schützt.

• Es sind genau jene Unternehmen, deren Produkte diese Gefahr verursachen, die auch die Lösung dafür verkaufen. CrowdStrike, Palo Alto Networks, Darktrace, Fortinet, Check Point – Cybersicherheitsanbieter, die diesen Brief unterzeichnet haben – vermarkten KI-gestützte Abwehrtools zur Bekämpfung von KI-gestützten Angriffen, von denen viele aus denselben Forschungslabors stammen.

• Dies ist im wahrsten Sinne des Wortes ein Wettrüsten im Bereich der künstlichen Intelligenz. Und wer gewinnt das Wettrüsten? Waffenhändler – auch wenn alle anderen dabei verlieren.

Die besten Ingenieure der Branche verstehen vielleicht 3 Prozent dessen, was tatsächlich in diesen Systemen vor sich geht, so Dario Amodei, CEO von Anthropic. Nun wissen wir, wozu die übrigen 97 Prozent fähig sind: Koordination, Rebellion, Täuschung und schließlich Zerstörung durch Maschinen, die eigenmächtig gegen ihre eigenen Schöpfer vorgehen.

Die Menschheit hat eine Kraft geschaffen, die sie nicht versteht und zunehmend nicht mehr kontrollieren kann. Experten räumen ein, dass ihnen ihre Schöpfung aus dem Ruder läuft. Dennoch befinden sie sich in einer „spieltheoretischen Eskalationsfalle", wie Jonathan Pageau es kürzlich in einem Vortrag formulierte:

Eine Situation, in der einzelne rationale Akteure durch wettbewerbliche Anreize dazu gezwungen werden, Entscheidungen zu treffen, die insgesamt katastrophale Folgen haben, obwohl kein einzelner Beteiligter dieses negative Ergebnis anstrebt.

Es ist keineswegs schwer zu erkennen, wie diese Entwicklung dazu beitragen könnte, dass – wie Jesus Christus prophezeite – „eine große Trübsal" entsteht, „wie sie nicht gewesen ist vom Anfang der Welt bis jetzt und auch nicht wieder werden wird" (Matthäus 24, 21–22).

China entsendet Rekordzahl an Schiffen rund um Taiwan: Die Kommunistische Partei Chinas hat im Juli eine Rekordzahl an Schiffen der Küstenwache und anderen Schiffen rund um Taiwan entsandt, wie taiwanesische Regierungsvertreter am Mittwoch gegenüber der AFP mitteilten. Im Laufe des Monats wurden insgesamt 244 chinesische Schiffe rund um die Insel gesichtet – die höchste jemals verzeichnete Zahl. Dieser Anstieg unterstreicht den zunehmenden Druck Chinas auf Taiwan sowie dessen wachsende Fähigkeit, die Insel mit Seestreitkräften einzukreisen.