# Der Narzisst auf meinem Smartphone

*Sophos Security-Experte Chester Wisniewski im Interview zu Fluch und Segen des neuen ChatGPT 4*

Es wurde mit großer Spannung erwartet: Das Update für ChatGPT. Seit letzter Woche gibt es nun die neue Version ChatGPT 4, die laut Hersteller Open AI die „fortschrittlichste“ KI-Technologie darstelle, noch kreativer, dafür „weniger voreingenommen“ sei und im Rahmen eines Tests sogar eine Anwaltsprüfung mit Bestnote bestanden haben soll. Zudem soll die Wahrscheinlichkeit sachlicher Antworten in der neuen Version gegenüber dem Vorgänger ChatGPT 3.5 um 40 Prozent gestiegen und die KI nun in der Lage sein, achtmal mehr Text zu erstellen.

Wie sieht ein Vertreter der IT-Security-Industrie die rasante Entwicklung bei der Chat-KI und wie ordnet er die daraus resultierenden Chancen sowie mögliche Nachteile ein? Chester Wisniewski, Principal Researcher bei Sophos hat hierzu einige Fragen beantwortet. Sein Fazit: Der Nutzen überwiegt vermutlich den Schaden.

*Was sind die Unterschiede zwischen GPT 3 und dem neuen GPT 4?*

Es hat sich eine ganze Menge geändert und verbessert, aber nicht alles ist bekannt. OpenAI hält seine Karten mit dieser Version etwas enger zusammen als in der Vergangenheit, daher sind nicht alle Verbesserungen publik geworden. Im Großen und Ganzen scheint es intelligenter, genauer und fähiger zu sein als frühere Versionen, was die Reaktionen noch realistischer und kompetenter machen sollte. Es ist dennoch wichtig, daran zu denken, dass es zwar weniger häufig falsche Informationen verbreitet, aber immer noch ein ziemlich guter Lügner ist.

*Ist es ethisch vertretbar, einen solchen generativen, vortrainierten Transformator öffentlich zugänglich zu machen?*

Er ist für die Bösen genauso verfügbar wie für die Guten. Der technologische Fortschritt ist wie ein Flaschengeist. Man kann ihn nicht einfach wieder wegstecken, wenn er unangenehm wird. Die Verfügbarkeit von Tools wie ChatGPT mit den Schutzmaßnahmen, die OpenAI durchzusetzen versucht, wird wahrscheinlich nicht mehr Schaden als Nutzen anrichten. Während krimineller Missbrauch von KI unvermeidlich ist, wird das Gute, das aus einer verantwortungsvollen Nutzung dieser Werkzeuge entstehen kann, wahrscheinlich jeden Missbrauch bei weitem überwiegen.

*Es scheint, als gäbe es bereits zahlreiche Beispiele dafür, dass Prompt-Engineering die Leitplanken des Systems durchbricht, und es gibt mindestens ein Beispiel für eine indirekte Prompt-Injektion. Glauben Sie, dass es jemals möglich sein wird, ein GPT-Modell zu schaffen, das nicht missbraucht werden kann?*

Nein. Vermutlich wäre eine detailliertere Antwort jetzt willkommener, aber jedes System, das Missbrauch verhindern soll, aber eben auch so konzipiert ist, dass das System autonom arbeitet, wird höchstwahrscheinlich immer umgangen werden können. Je mehr OpenAI und andere KI-Forscher darüber erfahren, wie die Menschen ihre Schutzmechanismen umgehen, desto schwieriger wird dies aber. Ich hoffe und erwarte, dass es in Zukunft deutlich mehr Geschicklichkeit aufseiten der Cyberkriminalität erfordern wird. Dennoch: es wird immer noch möglich sein, das System zu kompromittieren.

*Schafft ChatGPT irgendetwas Neues, oder setzt es nur das um, was ihm gesagt wurde? Was ist, wenn das Modell Ungenauigkeiten gelernt oder gelehrt bekommen hat?*

Das einzige, was ChatGPT neu erschaffen kann, sind Lügen. Es ist ein ziemlich überzeugender Lügner, der mit den besten Narzissten, die ich kenne, mithalten kann. Dem Modell wurden mit Sicherheit auch Dinge beigebracht, die ungenau oder unwahr sind, und es ist fast unmöglich vorherzusagen, wann es diese faktischen Ungenauigkeiten in seine Antworten einfließen lassen wird.

*Big AI wird im Besitz von Big Tech sein. Auch wenn die in den Eingabeaufforderungen enthaltenen persönlichen und vertraulichen Unternehmensdaten vielleicht nicht in das jeweilige AI-Modell einfließen, so werden sie doch mit Sicherheit den großen Technologieunternehmen zur Verfügung stehen. Und das Geschäftsmodell von Big-Tech-Eigentümern beruht häufig auf dem Verkauf solcher Daten an den Meistbietenden. Kann man etwas gegen die Datenschutzaspekte der öffentlichen KI im Stil von GPT tun?*

Für die meisten Nutzer von KI-Systemen gibt es ist es heutzutage kaum eine andere Wahl, als den Versprechen und Lizenzvereinbarungen, denen sie zustimmen, um Modelle wie ChatGPT nutzen zu können, Vertrauen zu schenken. Viele dieser Modelle entwickeln sich jedoch unglaublich effizient, so dass es keinen Grund gibt, warum sie in Zukunft nicht auf unseren PCs, Laptops oder sogar Telefonen laufen können. Das Training der Modelle ist zwar rechenintensiv, aber die Ausführung ist es nicht. Möglicherweise werden ältere Modelle als Open Source oder sogar als Crowdsourcing angeboten, damit sie jedem zur Verfügung stehen, der sie nutzen möchte. Die modernsten Modelle werden wahrscheinlich in den Händen der großen Technologiekonzerne bleiben, aber Modelle, die mehr als "gut genug" sind, könnten auf einem iPhone in Ihrer Tasche laufen, wenn Sie es wünschen.

**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/
Twitter: @sophos_info

**Pressekontakt:**

Sophos
Jörg Schindler, PR-Manager Central & Eastern Europe
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