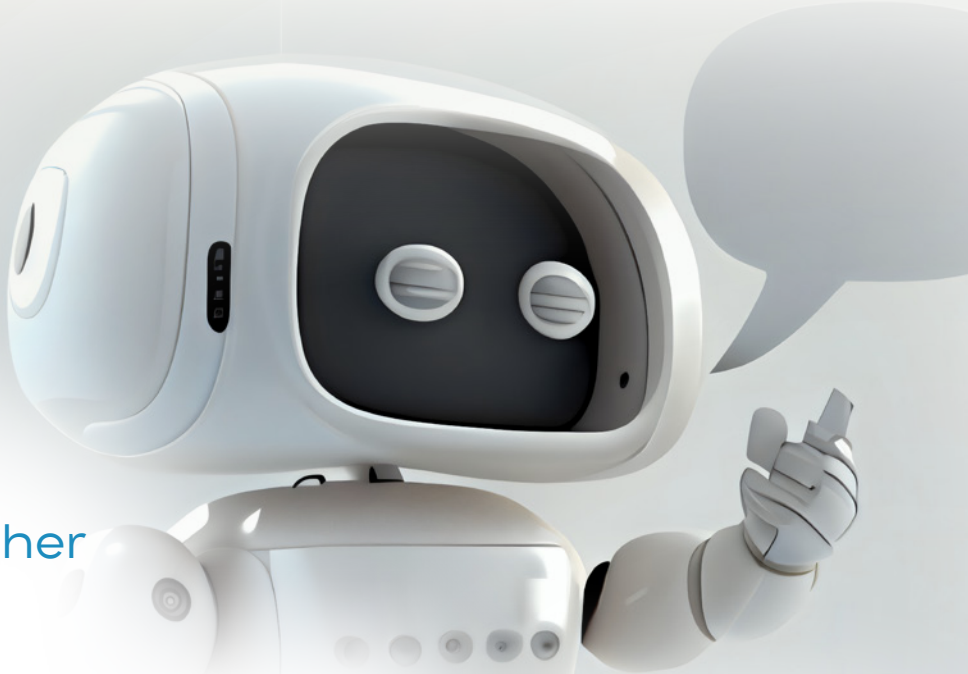

Vom angemessenen Umgang mit Künstlicher Sprachintelligenz

CHANCEN UND RISIKEN VON CHATGPT

ChatGPT ist ein leistungsstarker Chatbot, der durch konkrete Aufforderungen maßgeschneiderte Texte erstellt und Entwickler beim Programmieren unterstützt. Er erzeugt auf Basis seines „erlernten" Wissens Antworten auf gestellte Fragen. Die Einführung von ChatGPT erregte viel mediale Aufmerksamkeit und offenbarte die potenziellen Chancen dieser Technologie. Allerdings birgt sie auch Risiken. Dieser Artikel betrachtet beides umfassend – die Risiken insbesondere im Bereich IT-Sicherheit.

Generative Pre-trained Transformer (GPT) nutzt ein hochentwickeltes Sprachmodell („Large Language Model", kurz LLM). Der Bot basiert auf der Transformer Architektur, die als eine der führenden Methoden für natürliche Sprachverarbeitung und maschinelles Lernen gilt. Das Sprachmodell arbeitet mit einer statistischen Wahrscheinlichkeitsverteilung, welche die Wahrscheinlichkeiten von Zeichenketten (etwa Wörtern oder Sätzen) berechnet, basierend auf einer Eingabe oder einem Kontext.

Um ChatGPT zu verstehen lohnt es sich, einen Blick auf den „Pre-training"-Ansatz zu werfen. Hierbei wird eine initiale Version des Large Language Models mithilfe eines umfangreichen Datensatzes trainiert, der aus Millionen von Texten aus dem Internet, sozialen Medien, Online-Foren, Zeitungsartikeln und Büchern besteht. Dieser Datensatz ermöglicht es dem Modell, eine breite sprachliche Grundlage zu erlernen. Anschließend wird die initiale Version mit einem domänenspezifischen Datensatz weiterverfeinert. Durch dieses „Fine-tuning" kann das Modell spezifische Aufgaben und Domänen besser verstehen und passendere Antworten liefern [1].

Die Transformer Architektur, die von GPT verwendet wird, gilt als ein Meilenstein im Bereich der Sprachverarbeitung. Sie basiert auf dem Encoder-Decoder Ansatz und nutzt den sogenannten „Attention"-Mechanismus. Dieser Mechanismus ermöglicht es dem Modell, seinen Fokus auf bestimmte Teile der Eingabe zu richten und die Beziehung zwischen den Wörtern zu berücksichtigen. Während des Trainings erlernt das Modell, welche Teile der Eingabe für die Erzeugung einer angemessenen Antwort besonders relevant sind. Die grundlegende Struktur des Transformers besteht aus mehreren Ebenen, die in aufeinanderfolgenden Blöcken angeordnet sind. Der erste Schritt ist die Worteinbettung (Bild 1). Hier werden die Worte in eine Vektordarstellung überführt [2][3].

Ähnliche Worte werden mit einem ähnlichen Vektor abgebildet. Anstelle eines Algorithmus, der die Umwandlung durchführt, können auch vorhandene Datensätze verwendet werden. Die in Bild 1 mit Nx gekennzeichneten Ebenen werden in der Regel mindestens sechs Mal genutzt. Hierbei gibt die unterste Ebene die Ergebnisse an die nächsthöhere Ebene. Die Feedforward-

Netzwerke sind neuronale Netze, die zum Training eingesetzt werden. Wie der Name bereits impliziert, bewegen sich die Daten ohne Zyklen vom Anfang zum Ende des Netzes [1]. Der Multi-Head Attention Mechanismus (Bild 1) generiert eine Wahrscheinlichkeitsverteilung über das Vokabular der möglichen Ausgabesymbole [3].

Auf der linken Seite ist der Encoder zu sehen, der die aktuelle Eingabe verarbeitet und wichtige Informationen extrahiert. Der Decoder auf der rechten Seite nutzt die Informationen des Encoders, um die Antwort schrittweise zu generieren. Durch die Modifikation des Multi-Head Attention Mechanismus zu einem Masked-Multi-Head Attention Mechanismus wird sichergestellt, dass der Decoder nur Informationen verwendet, die zum aktuellen Schritt gehören, und keine Informationen aus der Zukunft verwendet, um eine sinnvolle Antwort zu gewährleisten. Die Ausgabe des Decoders ist eine Wahrscheinlichkeitsverteilung zwischen 0 und 1, die angibt, welches Wort oder Symbol als Nächstes generiert werden soll. Dieser Vorgang wird wiederholt, bis das Modell die Antwort komplett generiert hat [3].

Der Teil „GPT“ im Namen von ChatGPT steht für „generativer, vorab trainierter Transformer“. Das Modell ist generativ, weil es neue Texte erzeugen kann, die einer vorgegebenen Struktur und Sprachstil folgen. Es ist vorab trainiert, da es während des Pre-Training-Verfahrens grundlegende sprachliche Muster erlernt [11].

„HALLUZINATIONEN“ VON LLMS

Halluzinationen bei Sprachmodellen wie LLMs beziehen sich auf den Effekt, bei dem die Ausgabe der KI für den Nutzer echt erscheint, aber tatsächlich unsinnig oder falsch ist. Das bedeutet, dass die generierten Texte nicht immer zuverlässig oder vertrauenswürdig sind. Eine weitere Möglichkeit ist, dass die generierte Ausgabe nicht mit der vorgegebenen Eingabe übereinstimmt. Dies macht es schwierig, die Richtigkeit der Ausgabe zu überprüfen, insbesondere wenn das Sprachmodell auf externe Quellen verweist, die nicht verifizierbar sind. Der Ursprung von Halluzinationen kann bereits in den Trainingsdaten eines Sprachmodells zu finden sein. Diese Daten könnten Duplikate oder Informationslücken enthalten, was zu inkorrekten oder nicht verifizierbaren Aussagen führt. Dies hat zur Folge, dass von LLMs generierte Texte potenziell Aussagen enthalten können, die nicht auf Fakten basieren oder die Quellen nicht korrekt wiedergeben [4]. Das ist ein ernstes Anliegen, besonders wenn wichtige Entscheidungen auf der Grundlage solcher Ergebnisse getroffen werden. Es besteht das Risiko schwerwiegender Konsequenzen, wenn die KI in solchen Fällen halluziniert. Daher sollte die Zuverlässigkeit und Vertrauenswürdigkeit der generierten Texte sorgfältig geprüft und bewertet werden.

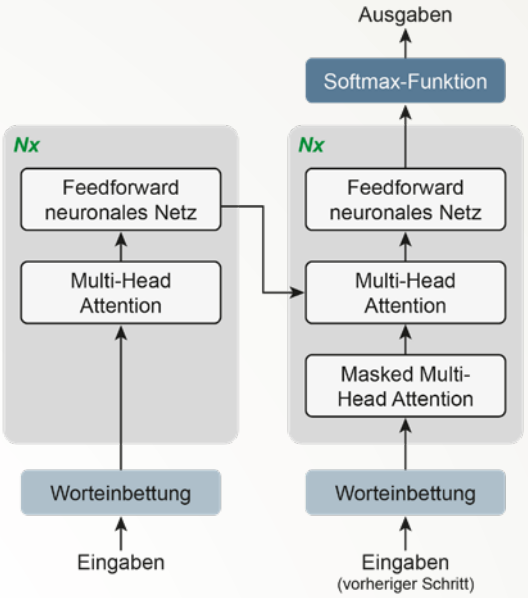

Bild 1: Vereinfachte Darstellung der Funktionsweise von ChatGPT

RISIKEN FÜR DIE IT-SICHERHEIT

Die Risiken im Zusammenhang mit ChatGPT sind vielfältig und werden im Folgenden genauer betrachtet.

Beschleunigtes Entwickeln von Angriffen

ChatGPT kann für die Erstellung von Angriffstechnologien wie Malware als nützliches Werkzeug dienen. Das Risiko besteht darin, dass Cyberkriminelle dadurch Angriffe schneller entwickeln können. Ein Angreifer kann eine Programmieranfrage an ChatGPT stellen und Code für Angriffstechnologien als Antwort erhalten. Diesen Code kann der Angreifer dann überprüfen, Fehler beheben und möglicherweise in anderen Code integrieren. Es ist jedoch wichtig zu betonen, dass ein Angreifer ohne ausreichende Erfahrung derzeit nicht in der Lage ist, mithilfe von ChatGPT funktionierende und vollständige Malware oder Angriffstechnologien zu erstellen. Kenntnisse in Programmierung sind weiterhin notwendig, insbesondere um Fehler zu finden und zu beseitigen [6].

Polymorphe Malware

Ein weiteres Risiko für die IT-Sicherheit besteht darin, dass ChatGPT bei der Entwicklung von polymorpher Malware eingesetzt werden kann. Polymorphe Malware ist eine Schadsoftware, die ihre Implementierung verändert, aber die Funktionalität – Schadfunktionen – beibehält [9][10]. Dies kann durch wiederholte Programmieranfragen mithilfe von ChatGPT erreicht werden, wobei der generierte Code bei jeder Programmieranfrage ein anderer ist. Hierdurch wird die Erkennungsrate von Schutzmechanismen wie Anti-Malwareprogramme, die auf Signaturerkennung setzten, verringert.

Bei der Signaturerkennung erstellt ein Anti-Malware-Hersteller eine Art „Fingerabdruck“ für jede neue Malware und teilt diesen mit seinen Kunden. Dadurch kann die Malware von der Anti-Malware-Software überall erkannt und blockiert werden. Wenn die Malware jedoch ständig ihre Gestalt ändert (polymorph ist), kann die Signaturerkennung nicht mehr effektiv funktionieren. Das führt dazu, dass die Anti-Malwarelösung die neue Malware nicht mehr so gut erkennen kann [6].

Optimiertes Social Engineering

LLMs haben die Fähigkeit, den Schreibstil einer bestimmten Person nachzuahmen, wenn sie mit spezifischen Trainingsdaten darauf trainiert wurden [7]. Diese menschenähnliche Fähigkeit macht Social Engineering-Betrug, insbesondere Spear-Phishing, einfacher.

Beim Spear-Phishing werden gezielt einzelne Empfänger ausgewählt, oft Unternehmen oder Administratoren mit umfassenden Rechten im IT-Bereich. Die Angreifer recherchieren dazu die sozialen und beruflichen Netzwerke der Opfer, um ihre persönlichen Interessen und Hobbys zu ermitteln. Mit diesen Informationen erstellen sie individualisierte Spear-Phishing-E-Mails, die äußerst glaubwürdig wirken und daher eine höhere Erfolgschance haben.

Früher erforderte dieser Prozess viel Aufwand und Fachkenntnisse, die die Angreifer möglicherweise nicht besaßen. Mit LLMs wie ChatGPT können Cyberkriminelle jedoch automatisch

hochgradig individuelle Spear-Phishing-E-Mails erstellen, die auf Grundlage vieler verfügbarer personenbezogener Daten im Internet erstellt werden. Dadurch werden diese E-Mails vertrauenswürdiger und erfolgreicher.

Darüber hinaus ermöglicht es LLMs einem Angreifer, maßgeschneiderte und fehlerfreie E-Mails in jeder Sprache zu verfassen, ohne dass er diese Sprachen selbst beherrschen muss, da ChatGPT auch die Übersetzung übernimmt [6].

CEO-Fraud

ChatGPT kann auch bei CEO-Fraud eingesetzt werden. Wenn ausreichend persönliche Daten vorhanden sind, kann ChatGPT Texte im Stil des angezielten CEOs verfassen. Das eröffnet neue Möglichkeiten, denn diese generierten Texte könnten zusammen mit Audio-Imitationen und Deepfake-Videos verwendet werden, um den Angriff auch über Anrufe oder Video-Calls umzusetzen [6]. In einem solchen Szenario gibt sich der Angreifer als der imitierte CEO aus und weist beispielsweise den Finanz-Manager des Unternehmens an, sofort einen Geldbetrag auf ein bestimmtes Konto zu überweisen, um angeblich ein wichtiges Geschäft abzuschließen. Durch diese Täuschung wird CEO-Fraud noch erfolgreicher.

Das Vertrauen der Nutzer in ChatGPT

Ein weiteres Risiko besteht darin, dass Menschen den Fähigkeiten von LLMs wie ChatGPT ein zu hohes Vertrauen entgegenbringen. Da ChatGPT Texte erzeugen kann, die sich so lesen, als ob sie von erfahrenen Menschen geschrieben wurden und oft korrekte Antworten liefern, wächst das Vertrauen in seine Fähigkeiten im Laufe der Zeit.

Dies kann dazu führen, dass der irreführende Eindruck entsteht, ChatGPT habe analoge Fähigkeiten und dasselbe Wissen wie ein Mensch in einem bestimmten Aufgabenbereich, und könne diesen somit ersetzen. Ein weiteres Problem ist, dass ChatGPT keine Informationen aus aktuellen Ereignissen hat und nur auf der Grundlage der Trainingsdaten arbeitet [7].

Besonders problematisch wird dies, wenn ChatGPT die Anfrage des Nutzers falsch versteht oder Halluzinationen hat, was zu fehlerhaften Antworten führen kann. Wenn der Nutzer nicht erkennt, dass die Antwort inkorrekt interpretiert wurde oder das System halluziniert, könnte er diese unzutreffende Antwort fälschlicherweise verwenden, wie es bereits vorgekommen ist. Dies wird noch verstärkt, wenn die Antworten als Grundlage für Entscheidungen von CEOs oder anderen Entscheidungsträgern dienen [6].

In solchen Fällen kann eine fehlerhafte Antwort von ChatGPT zu suboptimalen Entscheidungen führen, die finanzielle Auswirkungen auf das Unternehmen haben können. Die sicherste Verwendung von ChatGPT liegt darin, die Antworten selbst auf ihre Richtigkeit zu überprüfen.

Vertrauliche Informationen

Wenn vertrauliche Informationen in ChatGPT eingegeben werden, entsteht ein weiteres IT-Sicherheitsrisiko. Zu diesen vertraulichen Informationen können Kundeninformationen, Anlage- und Vermögensdaten von Banken, Softwarecode von Entwicklern oder vom ChatGPT generierter Code sowie Firmengeheimnisse wie Verkaufszahlen, Gehälter oder Patentinformationen gehören. Es besteht die Möglichkeit, dass diese eingegebenen Daten von ChatGPT für weiteres Training verwendet werden und in späteren Ausgaben auftauchen [6]. Das kann dramatische Folgen für das Unternehmen haben, das diese Informationen eingegeben hat.

Zusätzlich stellt die Eingabe vertraulicher Informationen eine neue Angriffsfläche dar, was das Risiko eines unerlaubten Abgreifens erhöht.

Risiken durch den geschlossenen Ansatz:

Da die genauen Trainingsmethoden von GPT-4 derzeit nicht öffentlich bekannt sind, ist es schwierig zu überprüfen, wie ChatGPT bestimmte Ausgaben für bestimmte Anfragen generiert. Dadurch wird es für Außenstehende erschwert, mögliche Lücken im Regelwerk nachzuvollziehen, das dazu dient, gefährliche Anfragen zu blockieren, zum Beispiel die Generierung von Malware. Die konkreten Anfragen und möglichen Varianten in diesem Regelwerk sind nicht bekannt.

Hassrede, Mobbing und Verbreitung von Falschinformationen:

Das GPT-Modell von ChatGPT hat auch die Fähigkeit, Hassreden zu generieren und gesellschaftliche Vorurteile sowie verschiedene Weltanschauungen in den Antworten widerzuspiegeln [8]. Durch seine Funktionsweise erlernt das Modell aus den Trainingsdaten alle im Internet verbreiteten Vorurteile und alternative Sichtweisen und kann diese möglicherweise in seinen Antworten reproduzieren.

ChatGPT kann Social Bots effektiv unterstützen, die als digitale Propaganda-Maschinen gezielt Meinungen oder Fake News verbreiten (Bild 2). Social Bots sind Meinungsroboter, die beispielsweise in sozialen Netzwerken gezielt Meinungen oder Fake News verbreiten. Um die Wahrscheinlichkeit der Aufnahme von Fake-Accounts in entsprechenden Echokammern zu erhöhen, wird mithilfe von ChatGPT ein Profil nach gewünschten Kriterien erstellt. Da alle Aussagen des Social Bots von ChatGPT generiert werden, wirken sie überzeugender und authentischer. Dadurch ist es möglich, Social Bots einzusetzen, um Börsenkurse oder Wahlen zu manipulieren, Cyber-Mobbing gegen Unternehmen oder Einzelpersonen zu betreiben und vieles mehr [6].

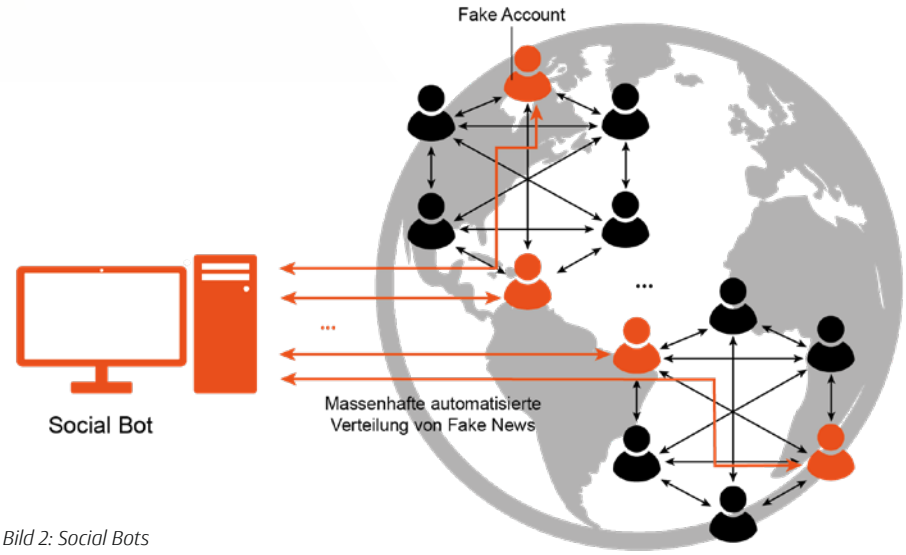

Bild 2: Social Bots

WELCHE CHANCEN CHATGPT FÜR DIE IT-SICHERHEIT ERÖFFNET

Neben den diskutierten Risiken bieten LLMs wie ChatGPT auch viele Chancen, sowohl im Allgemeinen als auch spezifisch für ChatGPT:

Erkennen von unerwünschten Informationen: LLMs können dabei helfen, Hassrede, Spam und Phishing-E-Mails zu erkennen, da einige LLMs neben der Textgenerierung auch Texte klassifizieren können.

Analyse von sicherheitsrelevanten Informationen:

ChatGPT kann Texte nach wichtigen Aspekten durchsuchen und diese kurz zusammenfassen. Mit speziellem Training können LLMs auch Logs nach sicherheitsrelevantem und verdächtigem Verhalten durchsuchen [7]. Hierbei sind menschliche Kontrollen jedoch wichtig, da LLMs Halluzinationen haben können. Sie unterstützen IT-Sicherheitsexperten, indem sie die Logs automatisiert durchsuchen und aufbereiten.

Erkennen von Sicherheitslücken und Softwarefehlern:

ChatGPT kann dazu verwendet werden, Code auf bekannte Sicherheitslücken zu überprüfen und diese zu erklären sowie Lösungsvorschläge zu machen [5]. Es ist jedoch wichtig zu beachten, dass die Kontextlänge von ChatGPT begrenzt ist und große Softwareprojekte nicht umfassend analysiert werden können [8]. Hinzu kommt, dass nicht jede von ChatGPT identifizierte Schwachstelle tatsächlich auch eine ist und nicht alle Schwachstellen, über die ein Code verfügt, können zuverlässig erkannt werden [5].

Unterstützung bei der Entwicklung größerer Software:

ChatGPT kann Entwickler unterstützen, indem es auf Anfrage Code generiert, der nach sorgfältiger Überprüfung in ein größeres Softwareprojekt eingepflegt werden kann.

Gemeinsame Nutzung von IT-Sicherheitsinformationen:

Eine eigene KI-gestützte Chatbot-Technologie speziell für IT-Sicherheit könnte Unternehmen in Deutschland/Europa zusammenbringen, um sicherheitsrelevante Informationen über Angriffe und erfolgreiche Gegenmaßnahmen zu teilen. Mit einem gemeinsamen LLM könnten Unternehmen sich schneller und besser gegenseitig schützen.

FAZIT

Das Verhalten von Sprachmodellen wie ChatGPT wird stark von den während des Trainings verwendeten Daten beeinflusst. Daher ist es entscheidend, hochwertige Trainingsdaten zu nutzen, um eine angemessene Leistung und Zuverlässigkeit des Modells zu gewährleisten. Dennoch sind Sprachmodelle nicht fehlerfrei und haben ihre Einschränkungen. Daher ist es wichtig, sich sowohl der Chancen als auch der Risiken bewusst zu sein.

Es ist ratsam, die Ergebnisse, die von einem Sprachmodell generiert werden, kritisch zu überprüfen und nicht bedingungslos als absolut korrekt oder wahrheitsgemäß anzusehen. Zudem sollte bei der Handhabung sensibler Informationen oder persönlicher Daten besondere Vorsicht walten. Solche sensiblen Informationen sollten nicht einem Sprachmodell eingegeben werden, um zu vermeiden, dass diese in zukünftigen Antworten verwendet werden oder von Kriminellen abgegriffen werden können.

Trotzdem gibt es Anwendungsfälle, bei denen Sprachmodelle helfen können, die IT-Sicherheit zu verbessern. Diese sollten in einer kooperativen Zusammenarbeit genutzt werden, um das Sicherheitsniveau zu erhöhen oder zumindest die Vorteile für Angreifer so gering wie möglich zu halten. Dabei sollte stets das Bewusstsein für die potenziellen Risiken und die Notwendigkeit der menschlichen Überprüfung der generierten Ergebnisse im Vordergrund stehen. ■

Literatur

[1] *S. Russell und P. Norvig: "Artificial Intelligence", 4th Global ed. 4. Aufl. Pearson Deutschland, Mai 2021, ISBN: 9781292401171 Publisher: Pearson Deutschland, ISBN: 978-1-292-40117-1.*

[2] *A. Vaswani, N. Shazeer, N. Parmar u. a.: "Attention Is All You Need", arXiv:1706.03762 [cs], Dez. 2017. doi: 10.48550/arXiv.1706.03762. Adresse: http://arxiv.org/abs/1706.03762 [Zugriff am: 29.05.2023].*

[3] *M. Phi: "Illustrated Guide to Transformers- Step by Step Explanation", Juni 2020. Adresse: https://towardsdatascience.com/illustrated-guide-to-transformers-step-by-step-explanation-f74876522bc0 [Zugriff am: 30.05.2023].*

[4] *Z. Ji, N. Lee, R. Frieske u. a.: „Survey of Hallucination in Natural Language Generation," ACM Computing Surveys, Jg. 55, 248:1–248:38, 2023, issn: 0360-0300. doi: 10.1145/3571730. https://dl.acm.org/doi/10.1145/3571730 [Zugriff am: 05.06.2023].*

[5] *R. Khoury, A. R. Avila, J. Brunelle, B. M. Camara: "How Secure is Code Generated by ChatGPT?" Apr. 2023. doi: 10.48550/arXiv.2304.09655. http://arxiv.org/abs/2304.09655 [Zugriff am: 09.05.2023].*

[6] *N. Pohlmann: „ChatGPT und Cyber-Sicherheit", Glossar „Cyber-Sicherheit", https://norbert-pohlmann.com/glossar-cyber-sicherheit/chatgpt-und-cyber-sicherheit/ [Zugriff am: 02.06.2023].*

[7] *„Large Language Models," en, Federal Office for Information Security, https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/AI-in-lLanguage-processing.html?nn=132646 [Zugriff am: 25. 05. 2023].*

[8] *OpenAI, GPT-4 Technical Report, März 2023. doi: 10.48550/arXiv.2303.08774. http://arxiv.org/abs/2303.08774 [Zugriff am: 29.05.2023].*

[9] *N. Pohlmann: „Cyber-Sicherheit – Das Lehrbuch für Konzepte, Mechanismen, Architekturen und Eigenschaften von Cyber-Sicherheitssystemen in der Digitalisierung", Springer-Vieweg Verlag, Wiesbaden 2022*

[10] *M. Chen, J. Tworek, H. Jun u. a.: "Evaluating Large Language Models Trained on Code", Juli 2021. https://arxiv.org/abs/2107.03374v2 [Zugriff am: 17.06.2023].*

[11] *Generative models, en-US. Adresse: https://openai.com/research/generative-models [Zugriff am: 11.07.2023]*

PATRICK FARWICK
studiert im Master Internet-Sicherheit an der Westfälischen Hochschule Gelsenkirchen und beschäftigt sich im Rahmen des Studiums mit ChatGPT.

NORBERT POHLMANN
ist Professor für Cybersicherheit und Leiter des Instituts für Internet-Sicherheit – if(is) an der Westfälischen Hochschule in Gelsenkirchen sowie Vorstandsvorsitzender des Bundesverbands IT-Sicherheit – TeleTrusT und im Vorstand des Internetverbandes – eco.