

Künstliche Intelligenz als Schlüssel für mehr Cybersicherheit in der Digitalisierung

Für die Cybernation Deutschland müssen wir die Digitalisierung nachhaltig voranbringen. Nur wenn alle Systeme auf dem Stand der Technik sind, können wir die sicherheitsrelevanten Vorzüge vollumfänglich nutzen und mit der Geschwindigkeit der Angreifenden Schritt halten. Notwendige Schlüsseltechnologien müssen dafür entwickelt, erforscht und eingesetzt werden.

Die Schlüsseltechnologie Künstliche Intelligenz (KI) eröffnet uns viele Möglichkeiten bei der Digitalisierung. Dazu zählen Anwendungen wie Chatbots, KI-gestützte Diagnosen oder automatisierte Auswertung großer Datenmengen. KI bringt jedoch auch eine neue Qualität von IT-Sicherheitsrisiken mit sich. Das BSI leistet Grundlagenforschung und entwickelt bedarfsorientierte und praxisnahe Prüfkriterien und Prüfmethodologien, um den Einsatz von KI so risikoarm wie möglich zu gestalten.

Sprachmodelle wie GPT4, Gemini, Mistral oder Luminous gehören der generativen Familie der KI an. Es sind Computerprogramme, die natürliche Sprache verstehen und erzeugen. Sie basieren auf einem statistischen Modell, das Muster in Text- oder Bilddaten erkennt und diese Muster verwendet, um zukünftige Texte vorherzusagen. Dabei liegen den Modellen meist neuronale Netze mit Milliarden oder gar Billionen von Parametern zugrunde, welche die Wahrscheinlichkeitsverteilung während des Trainingsprozesses anhand umfangreicher Textkorpora erlernen.

Sprachmodelle kommen in vielen Bereichen zum Einsatz: automatische Übersetzung von Texten, Spracherkennung, Textgenerierung und Chatbot-Entwicklung. Sie können auch dabei unterstützen, unerwünschte Inhalte wie Hate-speech in sozialen Netzwerken aufzuspüren. Sie ermöglichen es, große Mengen an Daten schnell und effizient zu analysieren und zu interpretieren, was in vielen Bereichen zu erheblichen Fortschritten führt. Gleichzeitig bergen sie neuartige Risiken, die einerseits in der Natur der Technologie selbst und andererseits in der gezielten Ausnutzung durch Angreifende begründet sind.

Einmal können Angreifende Sprachmodelle beispielsweise für Fälschungen missbrauchen. Da sie Texte in verschiedenen Sprachen erzeugen und Schreibstile von Personen oder Organisationen imitieren können, kann man sie verwenden, um Social-Engineering-Inhalte oder Falsch-



Raphael Zimmer

Raphael Zimmer ist promovierter Mathematiker und seit 2017 im Bundesamt für Sicherheit in der Informationstechnik. Im August 2021 startete er am Saarbrücker KI-Standort damit, das Referat „Sicherheit in der Künstlichen Intelligenz“ aufzubauen. Sein Team arbeitet forschungsnah daran, KI-Systemen zu schützen.

informationen zu erstellen. Zum anderen können Sprachmodelle beispielsweise in Form von sogenannten Prompt Injections angegriffen werden. Dabei beeinflussen spezielle Texteingaben das Verhalten des Modells. Filtermechanismen laufen so ins Leere.

In vielen Sprachmodellen sind Maßnahmen implementiert, die schadhafte Ein- und Ausgaben herausfiltern. Diese schützen aber meist nur partiell vor Missbrauchs- und Angriffsszenarien. Nutzende müssen sich daher der Risiken bewusst sein. Das BSI hat dazu eine Publikation „Große KI-Sprachmodelle – Chancen und Risiken für Industrie und Behörden“ (s. QR-Code) herausgegeben, die die Potenziale von Sprachmodellen erläutert und zugleich die aus ihnen resultierenden Probleme und Gefahren mit Fokus auf die IT-Sicherheit vertieft.

Die technologische Dynamik in KI-Anwendungen erfordert auch in Zukunft eine ebenso dynamische Bewertung der IT-Sicherheit beim Einsatz von Künstlicher Intelligenz. Die Rolle des BSI ist und bleibt dabei die des Möglichmakers, Helfers und Mitgestalters – ein starker Partner in der deutschen Sicherheitsarchitektur.



Publikation „Große KI-Sprachmodelle“