BSI

Positionspapier zu KI-Sprachmodellen

[16.05.2023] KI-Anwendungen, die natürliche Sprache verstehen und generieren können, sind auf dem Vormarsch – ChatGPT ist nur das prominenteste Beispiel. Das BSI hat nun ein Positionspapier publiziert, das Unternehmen und Behörden über Nutzen, Gefahren und Sicherheitsmaßnahmen informiert.

Große KI-Sprachmodelle, so genannte Large Language Models (LLMs), sind in der öffentlichen Diskussion omnipräsent. Insbesondere die Ankündigung und Veröffentlichung von Modellen wie ChatGPT haben solche KI-Sprachmodelle schnell bekannt gemacht. Im Rahmen des 19. Deutschen IT-Sicherheitskongresses hat das Bundesamt für Sicherheit in der Informationstechnik (BSI) ein Positionspapier veröffentlicht, in dem es über Stärken und Risiken solcher KI-Sprachmodelle informiert und geeignete Vorsichtsmaßnahmen empfiehlt.
KI-Sprachmodelle könnten sich als nützliche Werkzeuge in Bezug auf die IT-Sicherheit erweisen, sagt BSI-Vizepräsident Gerhard Schabhüser. So könnten sie beim Erkennen von Spam oder Phishing-Mails hilfreich sein oder beim Aufspüren unerwünschter Inhalte wie Fake News oder Hate Speech auf Social-Media-Plattformen. In gleichem Maße bergen KI-Modelle aber auch Risiken. Bereits jetzt werde im Darknet über den geeigneten Einsatz von KI zu Erstellung von Schadcode und Phishing-Mails diskutiert. Zudem eigneten sich KI-gestützte Sprachmodelle leider auch sehr gut zur Erstellung und Verbreitung von Falschinformationen. Es gelte jetzt dagegen aktiv zu werden und die Gesellschaft für den Umgang mit KI zu schulen, betonte Schabhüser.

Risikoanalysen und Aufklärung sind unerlässlich

Unternehmen oder Behörden, die über die Integration von LLMs in ihre Arbeitsabläufe nachdenken, rät das BSI, eine Risikoanalyse für ihren konkreten Anwendungsfall durchzuführen und die im Positionspapier genannten Risiken dahingehend zu evaluieren, ob diese für ihre Arbeitsabläufe eine Gefahr darstellen. Darauf aufbauend sollten existierende Sicherheitsmaßnahmen angepasst werden. Grundsätzlich sollten KI-Sprachmodelle aus Sicht des BSI derzeit als Werkzeuge betrachtet werden, deren Ergebnisse, etwa bei der Erstellung von Programmcode oder Texten, von einer menschlichen Intelligenz zu überprüfen sind. Manipulierte Bilder, Videos und Sprachausgaben sind nach Einschätzung des BSI Risiken, denen mit geeigneten Vorsichtsmaßnahmen begegnet werden sollte. So kann etwa die Authentizität von Texten und Nachrichten nachgewiesen werden, indem die Urheberschaft technisch belegt wird (wir berichteten).
Von besonderer Bedeutung ist nach Einschätzung des BSI die Aufklärung der Nutzenden über die Fähigkeiten Künstlicher Intelligenz. Durch die sprachlich oftmals fehlerfreie Textgenerierung entstehe häufig der Eindruck eines menschenähnlichen Leistungsvermögens und damit ein zu großes Vertrauen in KI-generierte Inhalte. Dafür zu sensibilisieren sei eine wichtige Maßnahme.

(sib)

BSI-Paper: Große KI-Sprachmodelle – Chancen und Risiken für Industrie und Behörden

Stichwörter: IT-Sicherheit, BSI, Künstliche Intelligenz (KI), ChatGPT