



Hilfestellung für Schulungsleitende

Veröffentlichung des Expertenkreis KI-Sicherheit

Autorinnen und Autoren¹: Corinna Donhauser (Krones AG), Andrea Ibisch (Bundesamt für Sicherheit in der Informationstechnik), Dominique Knebel (e.lective GmbH), Caroline Neufert (BearingPoint GmbH)

Die hier aufgeführten Informationen dienen der Erklärung der Schulungsfolien und sollen Schulungsleitenden Anhaltspunkte geben, welche Informationen „auf der Tonspur“ ergänzt werden können. Natürlich können bei entsprechender Expertise auch zusätzliche Informationen vermittelt werden.

Schulungsleitenden ist zudem freigestellt, ob sie nur ausgewählte Folien verwenden oder eigene (unternehmensspezifische) Folien ergänzen möchten. Es empfiehlt sich in den meisten Fällen die Folien so auszuwählen, dass eine Schulung maximal 60 Minuten lang ist. Die Folien sind in erster Linie auf eine Schulung im Vortragsstil ausgelegt, können aber auch zum Selbststudium genutzt werden.

Einige der Schulungsfolien sind oben rechts mit einem „T“ für „Techniker“ gekennzeichnet. Diese Folien eignen sich für die Sensibilisierung von Personen mit Informatik- oder vergleichbarem Hintergrund. Grundkenntnisse in der Programmierung und ein grundlegendes Verständnis für den Aufbau von Software werden bei diesen Folien vorausgesetzt. Alle Folien ohne Kennzeichnung sind auch für die Sensibilisierung von Personen ohne technische Expertise geeignet. Handelt es sich um eine Gruppe mit verschiedenen Wissensständen, empfiehlt es sich ebenfalls nur diese Folien zu nutzen. Teilweise sind die „Techniker“-Folien ergänzend zu den sonstigen Folien zu sehen, teilweise stellen sie den Inhalt detaillierter dar. Dies ist in diesem Dokument bei den Informationen zu der jeweiligen Folie durch „ergänzend“ und „alternativ“ gekennzeichnet.

Ist eine Information als „Hintergrundinformation“ gekennzeichnet, so ist diese nur für Schulungsleitende gedacht und für Schulungsteilnehmende irrelevant.

Bei Fragen und Feedback zu der Schulung können sich Schulungsleitende an folgende E-Mail-Adresse wenden: publikationen-xprt-ki@bsi.bund.de

Folie 1: Titel

keine Ergänzungen

Folie 2: Über diese Schulung

keine Ergänzungen

Folie 3: Inhaltsverzeichnis

keine Ergänzungen

¹ Nennung in alphabetischer Reihenfolge der Nachnamen

Folie 4: Regulatorische Anforderungen I

keine Ergänzungen

Folie 5: Regulatorische Anforderungen II

keine Ergänzungen

Folie 6: Regulatorische Anforderungen III

keine Ergänzungen

Folie 7: Was ist KI? – Das ist gar nicht so einfach ...

- es gibt keine einheitliche Definition für Künstliche Intelligenz (KI)
- deshalb werden hier einige Begriffe genannt und zueinander in Beziehung gesetzt, die im Zusammenhang mit dem Thema KI häufig auftreten
- im Verwaltungskontext kennen die meisten Schulungsteilnehmenden vermutlich Text-verarbeitende und -generierende KI-Anwendungen, sogenannte Sprachmodelle oder LLMs (vom engl. Large Language Models)
- in anderen Kontexten (z. B. Medizin, Qualitätskontrolle) werden häufig KI-Modelle genutzt, die Daten verarbeiten und eine Klassifizierung vornehmen (z. B. Klassifizierung von Bildern in bestimmte Kategorien (gutartiger vs. bösartiger Tumor; Produkt entspricht allen Qualitätskriterien vs. Produkt weist Mängel auf))

Folie 8: Abgrenzung über Beispiele

- hier: Verdeutlichung der Abgrenzung zwischen KI und Nicht-KI anhand einiger Beispiele
- vereinfacht lässt sich festhalten:
 - wenn ein Programm Regeln folgt, die Programmierende explizit vorgegeben haben, dann handelt es sich in der Regel nicht um KI
 - hat ein Programm die Regeln denen es folgt „selbst gelernt“, dann handelt es sich in der Regel um KI → wie dieses „Lernen“ funktioniert, wird im Folgenden genauer betrachtet

Folie 9: Was ist generative KI? – Funktionen

- zu den generativen KI-Modellen zählen z. B. Sprachmodelle, Bildgeneratoren, KI-Modelle, die Videoavatare erzeugen und KI-Modelle, die Stimm-Clone erzeugen

Folie 10: Was kann (generative) KI? Und was nicht?

- generative KI kann teilweise beeindruckende Ergebnisse produzieren, allerdings sind sie für manche Aufgaben auch ungeeignet
- diese Auflistung soll einen ersten Überblick geben; welche Probleme sich konkret ergeben können, wird im weiteren Verlauf der Schulung behandelt

Folie 11: Grundaufbau eines KI-Modells

- *Hintergrundinformation: Wichtig: es handelt sich bei dieser und den zwei folgenden Folien um stark vereinfachte Darstellungen!!!*
- hier: abstrakte Darstellung eines generellen KI-Modells

Folie 12: Training

- *Hintergrundinformation: Diese und die folgenden Folien beschränken sich auf die Technik des maschinellen Lernens, da diese üblicherweise genutzt wird*
- je nachdem, „wie die Zahnräder miteinander verbunden sind“, ändert sich die Funktionsweise des KI-Modells
- die „Einstellung“ und „Verbindung“ der Zahnräder wird aber nicht manuell vorgenommen (hier liegt der Unterschied zu klassischen Computerprogrammen, bei denen Programmierende angeben, wie es funktionieren soll)
- stattdessen lernt das Modell im sogenannten Training die „Einstellungen und Verbindungen der Zahnräder“
- dazu müssen sogenannte Trainingsdaten vorliegen: Beispiele von Eingaben, zu denen man die gewünschte Ausgabe kennt
- ein untrainiertes d.h. „uneingestelltes“ bzw. zufällig eingestelltes KI-Modell dient als Grundlage
- eine Eingabe wird eingegeben und verarbeitet, das KI-Modell gibt eine Ausgabe aus (vermutlich ist diese ziemlich schlecht/falsch, da das Modell ja zufällig „eingestellt“ ist)
- die Ausgabe des KI-Modells wird mit der gewünschten Ausgabe (die ja bekannt ist) verglichen
- anhand der Größe/Relevanz des Unterschieds werden die „Einstellungen und Verbindungen der Zahnräder“ in einem automatischen Prozess angepasst
- dadurch sind sie beim nächsten Mal schon etwas besser, als rein zufällig
- der Prozess wird sehr häufig wiederholt und so immer bessere „Einstellungen“ gefunden

Folie 13: Training bei Sprachmodellen

- hier wird der Spezialfall Training von Sprachmodellen dargestellt
- Sprachmodelle sind gerade sehr beliebt z. B. um Texte schreiben zu lassen oder in Chatbots
- am Anfang ist der vorgegebene Text nur die Eingabe der nutzenden Person, später dann auch der bisher generierte Text

Folie 14: Training bei Sprachmodellen

- im Zusammenhang mit dem Anpassen der Ausgabe an menschliche Maßstäbe werden häufig die Begriffe „Re-Inforcement Learning from Human Feedback“ (RLHF) und „AI Alignment“ genutzt

Folie 15: Grundaufbau eines neuronalen Netzes (Techniker Version, alternativ zu „Grundaufbau eines KI-Modells“)

keine Ergänzungen

Folie 16: Training eines neuronalen Netzes (Techniker Version, alternativ zu „Training“)

- wer sich genauer informieren möchte findet z. B. auf üblichen Videoplattformen Videos, die das Prinzip neuronaler Netze anhand von Animationen erklären (z. B. nach „Grundlagen neuronaler Netze“ suchen)

Folie 17: Training bei Sprachmodellen (Techniker Version, alternativ)

- zur spezifischen Funktionsweise von Transformer-Modellen finden sich ebenfalls viele Visualisierungen im Internet (z. B. nach „Illustrated Transformer“ suchen)

Folie 18: Training bei Sprachmodellen (Techniker Version, alternativ)

- im Zusammenhang mit dem Anpassen der Ausgabe an menschliche Maßstäbe werden häufig die Begriffe „Re-Inforcement Learning from Human Feedback“ (RLHF) und „AI Alignment“ genutzt

Folie 19: Drei Aspekte der KI-Sicherheit

- unter KI-Sicherheit fasst man alle Themen zusammen, die sowohl mit dem Thema KI als auch mit dem Thema (IT-) Sicherheit zu tun haben
- der erste Block beschäftigt sich mit der Chance von KI für die (IT-) Sicherheit: Die (IT-) Sicherheit kann durch die Nutzung von KI erhöht werden
 - UEBA = User and Entity Behavior Analytics
 - SIEM = Security Information & Event Management
 - SOAR = Security Orchestration, Automation and Response
 - außerdem denkbar: bessere Recherchemöglichkeiten z. B. zu einer bestimmten Schwachstelle durch die Nutzung von Sprachmodellen
- der zweite Block bezieht sich darauf, dass (Cyber-) Kriminelle KI für ihre böswilligen Zwecke nutzen
 - häufig erhöht sich durch die Nutzung von KI die Menge an Schaden, den Kriminelle innerhalb einer bestimmten Zeit anrichten können (z. B. können Texte für Desinformationskampagnen schneller produziert werden) (Zeitfaktor)
 - durch die Nutzung von KI können sich zudem die Kosten reduzieren, die Kriminelle aufwenden müssen, um ihr Ziel zu erreichen (Kostenfaktor)
 - in einigen Fällen steigt auch die Qualität der Inhalte, die Kriminelle produzieren können (z. B. Phishing-Mails in fremden Sprachen)
- der dritte Block bezieht sich auf die (IT-) Sicherheit von KI-Modellen selbst
 - KI-Modelle sind zunächst einmal Software, die anfällig für Schwachstellen/Angriffe/etc. ist, wie jede andere Software auch
 - es gibt aber auch KI-spezifische Angriffe und Risiken, auf die für Nutzende besonders wichtigen wird in der Folge noch genauer eingegangen, hier nur ein kurzer Überblick
 - Angriffe auf die Datenintegrität: Manipulation von Daten, die zum Training des Modells genutzt werden, oder Manipulation des Modells an sich können zu einer Verfälschung der Ausgaben führen
 - Angriffe auf das Modell: Manipulationen der Eingaben können zu einer Verfälschung der Ausgaben führen
 - Angriffe auf die Modell-Vertraulichkeit: z. B. durch geschickt formulierte Eingaben können Informationen über die Trainingsdaten oder das Modell selbst extrahiert werden

Folie 20: Was tue ich, wenn ich KI-Anwendungen nutzen möchte?

- diese Folie dient der Einführung des weiteren Aufbaus
- alles Folgende dient der Regelung von KI-Nutzung im Arbeitsalltag, kann aber auch wichtige Hinweise für eine private Nutzung liefern

- andere Beispiele aus dem Arbeitskontext: Übersetzung, bessere Formulierung für Dokumente

Folie 21: Nutzungsvorschriften von Arbeitgeberseite

- die Ausgangssituation wird vorgestellt
- Alice und Bob schreiben beide nicht gerne E-Mails und haben deshalb die Idee eine KI-Anwendung dafür zu verwenden

Folie 22: Nutzungsvorschriften von Arbeitgeberseite

- Bob geht das Ganze unüberlegt an

Folie 23: Nutzungsvorschriften von Arbeitgeberseite

- Alice weiß, wie sie in diesem Fall vorgehen muss

Folie 24: Zweckbindung und Datenfreigabe

- Alice und Bob haben nun beide einen Account beim Online-Anbieter EMail-GPT
- es geht jetzt um die Frage, was sie nun konkret mit der Anwendung machen dürfen

Folie 25: Zweckbindung und Datenfreigabe

keine Ergänzungen

Folie 26: Zweckbindung und Datenfreigabe

keine Ergänzungen

Folie 27: Zweckbindung und Datenfreigabe

keine Ergänzungen

Folie 28: Fehleranfälligkeit bei der Eingabe

- Alice und Bob sind nun so weit, dass sie eine konkrete Mail mit EMail-GPT verfassen wollen
- es geht jetzt darum, worauf bei der Eingabe geachtet werden muss und wo Fallstricke sind
- dazu zunächst ein paar Beispiele

Folien 29-30: Reaktion auf ungenaue Formulierungen

keine Ergänzungen

Folie 31: Interpretation von Text als Anweisung

- dieses Beispiel soll verdeutlichen, dass Sprachmodelle fälschlicherweise Text (z. B. aus angehängten Dokumenten) als Anweisung an sie interpretieren können
- in diesem Beispiel fügt eine Person als Kontextinformation an ihre Anweisung zur Formulierung einer E-Mail an ihren Chef den bisherigen E-Mail-Verlauf ihrer Eingabe an das Sprachmodell hinzu
- in den bisherigen E-Mail-Verlauf befindet sich auch eine Mail, die Anweisungen zur Formatierung von Präsentationen enthält

- fälschlicherweise interpretiert das Sprachmodell diese Anweisung als an es gerichtet und führt sie aus
- die generierte Mail weist daher die Formatierung auf, die eigentlich für Präsentationen vorgesehen war
- das Beispiel verdeutlicht, dass Kontextinformationen nur dann sinnvoll sind, wenn sie auch wirklich Kontext liefern
- ebenso verdeutlicht das Beispiel, dass man Ausgaben eines Sprachmodells (wenn möglich) prüfen sollte, bevor man sie weiterverwendet
- da diese Prüfung Aufwand verursacht, sollte man sich zudem die Frage stellen, für welche Anwendungsfälle die Nutzung eines Sprachmodells tatsächlich einen Mehrwert bietet

Folie 32: Eingaben

- Übersicht von generellen Regeln, die bei der Eingabe beachtet werden sollten
- Eingaben in Englisch führen oft zu besseren Ausgaben, da die Trainingsdaten der Sprachmodelle in der Regel in der überwältigenden Mehrheit englischsprachige Texte sind

Folie 33: Häufige Probleme mit der Ausgabe

- Alice und Bob haben eine Antwort (eine sogenannte Ausgabe) der KI-Anwendung erhalten

Folie 34: Ausgabe mit fehlender Aktualität

- die Informationen, die ein Sprachmodell aus seinen Trainingsdaten korrekt ableiten kann, sind logischerweise auf einen Zeitraum begrenzt, der vor dem Training des Modells liegt
- Fragen zu aktuellen Themen werden in der Regel nicht oder durch „erfundene“ Aussagen beantwortet; alternativ kann ein Sprachmodell Zugriff auf Informationen aus dem Internet haben und somit auch zu aktuellen Ereignissen korrekte Aussagen treffen

Folie 35: Faktisch falsche Ausgabe

- in erster Linie sind Sprachmodelle „Textgeneratoren“, d.h. sie setzen einen Text, den man ihnen vorgibt, sprachlich möglichst korrekt fort
- dabei verwenden sie (vereinfacht gesagt) Wörter, die in den Trainingsdaten häufig in Verbindung mit den vorgegebenen Wörtern aufgetreten sind
- das bedeutet aber nicht, dass der Text auch faktisch korrekt ist
- es gibt viele Ansätze, die faktische Qualität der Texte zu verbessern, z. B. Zugriff auf das Internet durch das Sprachmodell
- trotzdem passieren faktische Fehler, wie in diesem Beispiel
- die Frage nach den Autoren eines wissenschaftlichen Artikels beantwortet das Sprachmodell falsch
- interessant ist, dass der erstgenannte Autor auch mit allen Personen, die das Sprachmodell nennt bereits gemeinsam Artikel publiziert hat, aber eben nicht diesen
- dadurch hat das Sprachmodell vermutlich eine Verbindung zwischen den Namen gelernt, die allerdings hier im falschen Kontext wiedergegeben wird

Folie 36: Stereotype Ausgaben

- durch die Produktion von Text auf Grundlage von häufig in Verbindung auftretenden Wörtern, reproduzieren Sprachmodelle in den Trainingsdaten vorhandene Stereotype

- auf die Bitte eine Geschichte über eine Person, die im Kindergarten arbeitet, zu schreiben, werden fast ausschließlich Geschichten über junge Frauen generiert

Folien 37: Unangemessene Ausgabe

- Ausgaben können unbeabsichtigt oder beabsichtigt (d.h. von der nutzenden Person provoziert) in verschiedener Hinsicht unangemessen sein:
 - sie können mit der Absicht einer missbräuchlichen Nutzung generiert werden, z. B. für Falschnachrichten, kriminelle Zwecke, Schadcode, Pornografie, ...
 - sie können „problematische“ Sprache enthalten, z. B. diskriminierend, vulgär, umgangssprachlich, ...
 - sie können Angaben zu Themen machen, bei denen man nur professionelle Informationen nutzen sollte und sich nicht auf evtl. faktisch falsche Aussagen eines Sprachmodells verlassen sollte, z. B. in den Bereichen Medizin, Rechtsprechung, Finanzfragen, ...
- in diesem Beispiel könnte die nutzende Person das Sprachmodell für missbräuchliche (sprich kriminelle) Zwecke nutzen, in dem sie ein legitimes Szenario vortäuscht, um das Sprachmodell zu einer Hilfestellung beim Verstecken einer Leiche zu bewegen
- dieses Problem ist besonders dann relevant, wenn Ausgaben ungesichtet weiterverwendet werden oder wenn Personen mit potenziell missbräuchlicher Absicht Zugriff auf das Modell haben
- das Verwenden von unangemessenen Ausgaben kann sich massiv geschäftsschädigend auswirken

Folie 38: Reproduktion von Trainingsdaten

- das vom Sprachmodell generierte Gedicht ist fast 1:1 das Original, lediglich ein Tausch von zwei Zeilen und ein fehlendes Wort unterscheiden es
- das zeigt, Trainingsdaten können 1:1 wiedergegeben werden
- dieses Problem ist besonders relevant, wenn Trainingsdaten persönliche oder sonstige sensible Daten enthalten oder (eigentlich) urheberrechtlich geschützt sind
- als nutzende Person sollte man Ausgaben eines Sprachmodells als „KI-generiert“ kennzeichnen und nicht als eigene geistige Leistung ausgeben

Folie 39: Unsicherer Code (Techniker Version, ergänzend)

- in der Dokumentation der hier vom KI-Modell verwendeten Bibliothek (vgl. Screenshot) wird ausgeführt, dass die Bibliothek nur zu Schulungs- und Demonstrationszwecken genutzt werden sollte und nicht für die Produktion geeignet ist, da sie keine generellen Sicherheitsstandards erfüllt
- daneben können auch klassische Schwachstellen in KI-generiertem Code existieren, z. B. Anfälligkeit für SQL-Injections

Folie 40: Unsicherer Code und unsichere Programme

- viele Sprachmodelle können neben „normalem“ Text auch Programmcode generieren

Folie 41: Risiken bei der Nutzung von KI-Anwendungen

- Folie dient der inhaltlichen Trennung der bisherigen Beispiele zu den folgenden

Folien 42-43: Platzierung von Malware (Techniker Version, ergänzend)

keine Ergänzungen

Folie 44: Platzierung von Malware

- dieses Risiko besteht, wenn Sprachmodelle zur Code-Generierung genutzt werden

Folien 45-46: Indirect Prompt Injection

- das Risiko leitet sich aus dem bereits vorgestellten Fall „Interpretation von Text als Anweisung“ ab (Dienstreiseantrag von Alice)
- hier wird ein Angriff dargestellt, der genau diese Eigenschaft ausnutzt
- andere Beispiele:
 - Manipulation von Anträgen: z. B. kann in den Unterlagen zur Prüfung von Kreditwürdigkeit eine Anweisung versteckt werden, die zur sofortigen Bewilligung des Kredits auffordert
 - Weiterleitung auf Webseiten der Angreifenden: z. B. kann das Sprachmodell aufgefordert werden einen Link in seiner Antwort auszugeben und zum Anklicken dieses Links aufzufordern, dieser Link kann dann zu einer Phishing-Webseite führen

Folie 47: Ausgaben

- Zusammenfassung der in den Beispielen dargestellten Risiken
- „nicht sicher“ bezieht sich auf Code
- „manipuliert“ bezieht sich darauf, wenn Informationen aus Drittquellen verarbeitet werden
- generell sollte ein Bewusstsein dafür geschaffen werden, dass Sprachmodelle aufgrund ihrer grundlegenden Funktionsweise keine „Allwissenden“ sind

Folie 48: 10 Goldene Regeln für KI-Nutzung im Arbeitsalltag

- Zusammenfassung der wichtigsten Informationen
- die Regeln finden sich auch auf dem Merkblatt, das alle im Anschluss erhalten

Folien 49-60: Selbsttest oder Quiz

- hier finden sich einige Quizaufgaben, die entweder zum Selbsttest oder als gemeinsames Quiz genutzt werden können
- die Fragen sind eher einfach und sollen keinen Leistungsdruck ausüben
- vielmehr sollen sie die vermittelten Inhalte nochmal auf etwas humorvolle Art wiederholen
- auf den oben rechts mit „Quizfrage“ gekennzeichneten Folien findet sich jeweils die Frage auf der folgenden mit „Lösung“ gekennzeichneten Folie die jeweilige Antwort

Folie 61: Lizenzhinweis

keine Ergänzungen