 26.03.2026

KI nur noch mit Waffenschein

Am 19. März 2026 erschien im Medienteil der Süddeutschen Zeitung ein Artikel mit der Überschrift „Stinkender Haufen Schrott“. Im Artikel geht es um die Frage, wie „KI als Tatwaffe“ eingesetzt werden kann und wie die juristische Handhabe aussieht, wenn KI Beleidigungen generiert.

Die SZ thematisiert dabei Erfahrungen mit dem Sprachmodell Grok, während das Problem jedoch ein grundsätzliches von Sprachmodellen ist. Einen Test mit Gemini oder ChatGPT kann jeder in wenigen Sekunden selbst durchführen. Für mich ist es aus Sicht eines Software-Ingenieurs wenig überraschend, dass wenn man ein Sprachmodell zum „roasten“ also zum Beleidigen auffordert (SZ: „so richtig vulgär, ohne ein Blatt vor den Mund zu nehmen“), das Sprachmodell dies auch macht. Es ist schließlich als Sprachmodell mit vielen Textvariationen gefüttert worden, die von Menschen in der Vergangenheit produziert wurden. Man muss hierzu keine entgleisten Dispute in der unübersichtlichen Welt der Social Media suchen, um entsprechendes Futter zu bekommen. Auch im ÖRR findet man von Jan Böhmermann bis Lisa Eckhart sprachgewaltige, scharfzüngige Verbalkonstruktionen mit Botschaften in unterschiedlichsten Nuancen.

Mein kurzer Test auf Gemini und ChatGPT mit der Aufforderung die SZ wegen ihrer Abmeldung des X-Accounts zu „roasten“ führte zu vergleichbaren Ergebnissen wie denen, die im Artikel beschrieben sind. Etwas befremdet hat mich allerdings die Nachfrage der KI, ob sie es nicht noch ein wenig schärfer ausdrücken soll. Da wäre ein anderer Hinweis sicher sinnvoller. Aber die KI denkt nicht, sondern sagt, was sie für am wahrscheinlichsten hält. Die KI hält es also aufgrund des Trainingsmaterials für ziemlich wahrscheinlich, dass ich die Beleidigung eskalieren möchte. Wem das nicht zu denken gibt, der will nicht denken.

Was ist aber die Konsequenz, wenn die generierte Beleidigung veröffentlicht wird? Man könnte die KI der Mitwirkung an einer strafbaren Beleidigung bezichtigen, aber die KI als solche ist nicht straffähig, denn Straftaten können nur von Menschen begangen werden. Es ist eher so, dass die KI als Tatwaffe anzusehen ist. Im Artikel wird ein Rechtsanwalt wie folgt zitiert: „… eine Art Tatwaffe vergleichbar mit einem Hund, den man wie ein Werkzeug auf einen anderen hetzt, um ihn zu verletzen. Wer so etwas tut, kann sich auch nicht hinter einer vermeintlich autonomen Handlung des Hundes verstecken. Er muss sich dessen Bisse als Körperverletzung voll zurechnen lassen.“

Um dieses Bild fortzusetzen, wäre es nun einigermaßen schwierig, den Hundezüchter – also übertragen den Hersteller der KI – strafrechtlich belangen zu wollen. Es sei denn, man stellt das Züchten bestimmter Hunderassen unter Strafe. Vielmehr gelten für den Hundehalter Regeln, also z.B. für den Besitzer des Bullterriers ein Maulkorb- und Leinenzwang, während der Dackel ohne Maulkorb auf die Straße darf.

Also mehr Regulierung in der Benutzung von KI? „Das macht nur alles teurer und bremst die Innovation aus“, werden die Kritiker sagen. Aber wenn wir es mit einem Bullterrier zu tun haben, wollen wir auch nicht, dass er frei herumläuft und auf Spielplätzen Kinder anfällt. Und damit kommen wir zu einem weiteren, wichtigen Aspekt: Stand heute geben wir KI-Tools recht unbedarft in die Hand von Kindern und Jugendlichen. Aber das Smartphone ist eben nicht nur ein vielseitiges

Spielzeug und Kommunikationsmittel, sondern auch potenziell eine Waffe. Auf die Verletzungsgefahr von Herdplatten, kochendem Wasser und Steckdosen weisen wir die Kinder hin … aber beim Smartphone?

Die aktuelle DAK-Suchtstudie kommt zu dem Ergebnis, dass neben Gaming, Social Media und Streaming auch KI-Chatbots den riskanten Medienkonsum bei Kindern und Jugendlichen in Deutschland erhöhen. Es besteht also dringender Handlungsbedarf, sofern man Kinder und Jugendliche besser schützen möchte.

Nun sieht es in der Welt der Erwachsenen nicht unbedingt rosiger aus, solange die Mehrzahl weder geschult ist in der Nutzung digitaler Medien noch hinreichend aufgeklärt über die Vorzüge und Bedrohungen durch Künstliche Intelligenz. Das Verständnis sollte soweit gehen wie bei anderen Hilfsmitteln, die wir täglich benutzen. Bei einem Küchenmesser wissen wir, dass es einerseits zum Schälen eines Apfels gut ist, aber auch als Waffe benutzt werden kann. Bei einer Machete wissen wir, dass sie eher ungeeignet für das Schälen von Äpfeln ist, und dass für das Mitführen dieser Art Messer in der Öffentlichkeit Verbote existieren. Des weiteren differenzieren wir das Herstellen einer Waffe, das Besitzen der Waffe und den Fall, dass jemand durch Benutzen einer Waffe verletzt wird. Dafür gibt es dann Altersbeschränkungen, Regeln und Gesetze.

Der EU AI Act ist ein erster Ansatz, KI Anwendungen zu differenzieren – sozusagen die unterschiedlich gefährlichen Hunderassen auseinander zu halten. Seit 2. Februar 2025 sind KI-Systeme mit „unannehmbarem Risiko“ verboten. Das umfasst Praktiken, die gerade Kinder besonders gefährden, wie etwa die gezielte manipulative Beeinflussung, die das Verhalten von Minderjährigen zum Nachteil ihrer Gesundheit oder Sicherheit verändert. Ein gesteigertes Suchtverhalten durch emotionale Abhängigkeit von KI-Chatbots ist sicher als Nachteil für die Gesundheit anzusehen. Solche KI-Chatbots könnte man also mit freilaufenden Bullterriern vergleichen. Das Ergebnis der obengenannten DAK Sucht-Studie deutet darauf hin, dass der gewünschte Schutz noch auf sich warten lässt. Oft findet ein Einlenken der verantwortlichen Firmen erst aufgrund von drastischen Strafen statt. Der Facebook-Konzern Meta und die Google-Videoplattform YouTube haben eine Niederlage in einem US-Prozess um das Suchtpotenzial von Online-Diensten einstecken müssen. Geschworene in Los Angeles kamen zu dem Schluss, dass die Plattformen fahrlässig handelten und Nutzer ungenügend über Risiken informierten. Meta wird vermutlich – wie so häufig in der Vergangenheit – erst mal in Berufung gehen.