Warum KI verblödet

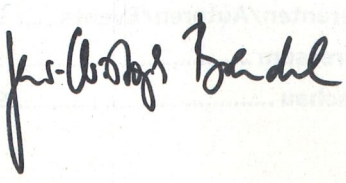
Jens-Christoph Brendel
Stellv. Chefredakteur

Der Patient versinkt in geistiger Umnachtung: Sein Denkvermögen nimmt ab, er leidet unter Sprachstörungen, das Zeichnen gerät zum Krakeln, seine Gedanken drehen sich im Kreis. So anrührend – oder auch furchteinflößend – das klingt, in diesem Fall braucht der Betroffene unsere Empathie nicht. Denn er ist kein Mensch, sondern eine KI – eine, die unter MAD leidet. Das Kurzwort ist ein Sprachspiel, das nur im Englischen funktioniert, denn einerseits erinnert die von Forschern so benannte KI-Krankheit Model Autophagy Disorder (MAD) an den Rinderwahn, englisch Mad Cow Disease, und andererseits bedeutet „mad" gleichzeitig „verrückt, irre, wahnsinnig". Und nicht nur die Namen der Krankheiten ähneln sich, sondern wohl auch die Ursachen und Symptome.

Wie genau MAD in einer KI entsteht, daran muss noch geforscht werden, bislang ist es unbekannt. Genau weiß man aber schon, was die Krankheit auslöst: Ursache ist das Training von generativen KI-Modellen (Bildgeneratoren, aber auch Large Language Models) mit synthetischen Trainingsdaten, die ihrerseits von einer KI erzeugt wurden. Sobald eine generative KI ihre eigenen Produkte wieder als Trainingsdaten konsumiert, setzt etwas ein, was die Forscher der Universitäten Stanford und Rice eine selbstverzehrende Schleife („autophagous loop") nennen. In der Folge mehren sich Fehler wie Artefakte in Bildern, die Diversität der Resultate nimmt ab, insgesamt sinkt die Qualität, und zwar umso mehr, je weniger echte, frische Daten einbezogen werden. Das ist eine weitere Parallele zu BSE: Der Rinderwahn ging ebenfalls darauf zurück, dass Rinder mit Überresten von Rindern gefüttert wurden, einschließlich des Gehirns.

Ja, dann lassen wir das eben sein und trainieren nur noch mit noch nie verarbeiteten Daten, könnte man sich vornehmen. Das jedoch ist leichter gesagt als getan: Mit der steigenden Popularität generativer KI-Modelle überschwemmen sie zunehmend das Internet mit ihren Daten. Man mutmaßt, dass es schon bald mehr synthetische Daten im Netz geben wird als von Menschen generierte. Gleichzeitig ist das Internet die primäre Quelle für Trainingsdaten. Filter, die echte von synthetischen Videos, Bildern oder Texten zuverlässig unterscheiden könnten, gibt es nicht. Also füttern die Trainer ihre KI-Modelle bereits ungewollt und unvermeidlich mit deren eigenem Output, und sie erkranken. Zudem setzt man Kunstdaten aber auch ganz bewusst ein, denn sie haben eine Reihe von Vorteilen: So sind sie billiger als echte und machen keine Datenschutzprobleme. Der wichtigste Grund aber ist: Nach der Lawine synthetischer Daten gibt es schlicht keine neuen, von echten Menschen gemachten Daten mehr in den benötigten, riesigen Mengen.

Eine andere Studie von Forschenden der Unis Stanford und Berkeley berichtet von starken Schwankungen der Problemlösungskompetenz zwischen den Sprachversionen ChatGPT-3.5 und ChatGPT-4. In manchen Disziplinen schneidet das neuere Modell deutlich schlechter ab als das ältere, etwa beim Erkennen von Primzahlen oder beim Generieren von Code. Sind das schon MAD-Symptome? Daran muss weiter geforscht werden, für endgültige Schlüsse ist es zu früh. Auch ist das Bild nicht einheitlich: In manchen Disziplinen hat sich die Leistung verbessert, in anderen verschlechterte sie sich rapide. Aber eines zeichnet sich schon jetzt ab: Es ist keinesfalls ein Naturgesetz, dass KI-Modelle mit jedem Entwicklungszyklus und in allen Aspekten immer besser werden. Vielleicht braucht man vor einer allmächtigen KI allein schon deshalb keine Angst zu haben, weil sie sich auf dem Weg zur Allmacht selbst ausbremst. Möglicherweise wirkt im Hintergrund ja eine negative Rückkopplung: Bessere KI führt zu immer mehr Kunstdaten, die aus Geiz, Bequemlichkeit, Unkenntnis oder mangels besseren Materials immer öfter in KI-Trainings einfließen, was noch bessere KI unmöglich macht …



Weitere Infos und interessante Links
www.lm-online.de/qr/48766