Wenn aus Intelligenz Unsinn wird

Johannes Wirz

Künstliche Intelligenz (KI) ist nicht mehr aus der Welt wegzudenken. Sobald grosse Datenmengen auf Muster – Ähnlichkeiten und Unterschiede – geprüft werden müssen, ist die Maschinenintelligenz unverzichtbar.

Als ich vor mehr als 40 Jahren meine Promotion gemacht habe, galt es als «Gesetz», dass die räumliche Struktur eines Proteins aus der sog. Primärsequenz, d.i. die Reihenfolge der Aminosäuren (Bausteine), nie abgeleitet werden kann. 2020 präsentierte AlphaFold die dreidimensionale Struktur aller Proteine des Menschen und darüber hinaus Strukturen von Millionen Proteinen aus nicht menschlichen Lebewesen, von Bakterien, Pflanzen und Tieren – abgeleitet aus der Primärstruktur und trainiert mit Strukturen, die durch die experimentelle Röntgenkristallografie ermittelt worden waren. Das Aufregende dabei war, dass AlphaFold die Strukturen in derselben Auflösung darstellen konnte, wie sie im Experiment erscheinen: 6-9 Angström ($6x10^{-10}$ m).

Heute modellieren vergleichbare KI-Verfahren die Wetterprognosen, die Klimaveränderung und Kursverläufe von Aktien an der Börse, sie diagnostizieren Krebserkrankungen, bevor der Histologe sie unter dem Mikroskop entdeckt, und sie steuern Autos ohne Fahrer durch den Verkehr.

Zukunft und Zweifel

Glaubt man dem Google Unternehmen AlphaFold, so werden mit KI bald neue innovative Heilmittel entwickelt. 3D-Drucker lassen «neue» Bilder der grössten Maler entstehen, die auch von ausgewiesenen Experten nicht mehr als Fälschung erkannt werden können. Und auch die Schreibkünste werden nicht mehr aus der Feder von AutorInnen fliessen, sondern von Programmen wie ChatGPT geschrieben.

Die schöne neue Welt schafft jedoch auch grosse Probleme. Ein erstes betrifft die Prognosen für die Zukunft, die selbstverständlich nur eine Projektion aus akkumulierten Daten der Vergangenheit sein können. Stefan Brotbeck, der Basler Philosoph, hat diese Art der Projektion an einer landwirtschaftlichen Tagung ohne Bezugnahme auf KI als spirituellen Mähdrescher bezeichnet: Er wirft die Ernte stets vor das Schneidewerk und sammelt sie wieder ein. Diese Art der Zukunftsgestaltung bezeichnete er als Futurum – im Gegensatz zum Adventum, mit dem die Zukunft uns entgegenkommt.

Das zweite Problem ist technischer Art und hängt damit zusammen, dass die verfügbaren Daten nicht nur real erhoben werden, sondern selbst zunehmend aus KI-Modellierungen stammen.

KI produziert Unsinn

Dieses Problem wurde kürzlich in der Zeitschrift Nature eindrücklich analysiert (*Shumailov et al.* 2024). Die Autoren zeigen, unter welchen Bedingungen Large Language Models (LLMs) wie ChatGPT oder Llama ohne Korrektur durch einen menschlichen Editor in wenigen Zyklen («Generationen») aus sinnvoller Information Unsinn produzieren. Der erste Fehler entsteht durch die Begrenzung der zugänglichen Daten, also durch statistische Approximation. Der zweite Fehler, die Autoren bezeichnen ihn als funktionale Expressivität, tritt auf, wenn ein LLM unterschiedliche Datensätze vereinigt. Zur Veranschaulichung verwenden die Autoren ein einfaches Beispiel. Wenn zwei Gauss'sche Normal-Verteilungen zu einer vereinigt werden, werden die äusseren Wahrscheinlichkeiten der Verteilungen in der einen zusammengeführt, die inneren dagegen verschwinden.

Die dritte Bedingung wird als Fehler der funktionalen Approximation bezeichnet. Er hängt zusammen mit der begrenzten maschinellen «Lernfähigkeit» der LLMs.

So abstrakt die Sache klingt, und offen gestanden habe ich die Fehlerbedingungen nicht im Detail verstanden, so konkret sind die Beispiele, die sich aus der Modellierung ergeben haben.

Wenn die LLMs nur noch ihre eigenen, selbst generierten Informationen weiterverarbeiten – eine Situation, die sich mit der zunehmenden Anzahl der maschinell erzeugten Daten einstellen kann – kollabiert das Modell nach wenigen Zyklen der Datenverarbeitung. Der Vorgang ähnelt einer Inzuchtdepression, wie wir sie bei Pflanzen und Tieren kennen (*Wenger* 2024).

In einem anderen Beispiel erzählt *Elizabeth Gibney* (2024), was geschieht, wenn LLMs nur die von ihnen generierten Texte verwenden. Es mutet grotesk an, dass aus Wikipedia-Artikeln über Kirchtürme in England nach neun Zyklen (Generationen) eine Abhandlung über die Färbung der Schwänze von Hasen entsteht, wenn die LLMs lediglich mit den von ihnen selbst generierten synthetischen Daten gefüttert wurden.

Das Fazit dieser Untersuchungen liegt auf der Hand: Wir sollten die Modellierung der Zukunft nicht ausschliesslich den Maschinen überlassen, sondern auf unsere intuitiven Fähigkeiten der Zukunftsschau vertrauen und dafür sorgen, dass KI «lernt», die immer grösser werdende Menge synthetischer Daten von den realen zu unterscheiden. Dafür braucht es jetzt und in Zukunft immer mehr unsere wache Urteilsfähigkeit.