Paul Cziep

Gymnasium Paulinum, Schwaz

Betreuung durch Bernd Ziermann

Thema 2

> Maschinen sind im wahrsten Sinne des Wortes unmündig. Sie können sich nicht für ihr Handeln verantworten, weil sie die Fähigkeit zur Täterschaft ebenso wenig besitzen wie guten Geschmack: Maschinen sind ihrem Wesen nach apolitisch, weil sie weder Ungerechtigkeit erfahren können noch der Gerechtigkeit bedürfen.
>
> Lisz Hirn: Der überschätzte Mensch. Wien 2023, S. 95

Geister der Maschinen

Maschinen seien unmündig und könnten nicht für ihr Handeln verantwortlich gemacht werden, so zumindest Lisz Hirn. Sie seien nur Konstrukte aus Transistoren und Drähten, die durch menschliche Entwicklung menschliche Programme ausführen können. So war dies auch immer seit der Erfindung des ersten Computers. Doch heute, in einem Zeitalter von Künstlicher Intelligenz, Machine Learning und Künstlichen Neuronalen Netzen, scheinen Maschinen sich immer schneller und vor allem immer selbstständiger zu entwickeln. Nun stellt sich die Frage: Sind Maschinen tatsächlich so unmündig, so unfähig zur Täterschaft, wie es Hirn darstellt?

Zuallererst muss man Maschinen in zwei Kategorien teilen: Einerseits gibt es einfache Maschinen, die speziell für einen Zweck gebaut und programmiert wurden, das wäre z.B. ein Fernseher. Andererseits gibt es Maschinen, die selbst gelernt haben und nicht für einen bestimmten Zweck direkt programmiert wurden, ein perfektes Beispiel dafür wären Sprachmodelle wie ChatGPT.
Lisz Hirns Aussage trifft wohl auf die erste Kategorie von Maschinen zu. Ein Fernseher hat kein Verlangen nach Gerechtigkeit oder Ähnlichem. Systeme der zweiten Kategorie haben allerdings sehr wohl das Potenzial, sich zu etwas zu entwickeln, das man sogar als Täter darstellen kann. Eine solche Maschine muss allerdings bestimmte Voraussetzungen erfüllen. Selbst die besten Sprachmodelle werden noch immer in den Bereich der „Narrow Intelligence“ eingeordnet. Ein solches System berechnet nur die Wahrscheinlichkeit für das nächste Wort, ohne Ziel, ohne Intention, ohne Gedanken. Ein Modell, das sehr wohl für sein Handeln verantwortlich gemacht werden kann, muss

genau diese fehlenden Dinge in sich tragen. Wenn eine solche Maschine etwas als in der Gesellschaft schlecht Angesehenes vorsätzlich tut, so kann sie dafür verantwortlich gemacht werden.

Vor wenigen Tagen wurde ein neues System von Google DeepMind vorgestellt, die sogenannten „Titans“. Titans sind der Nachfolger der Transformer Architektur, die im Jahr 2017 ebenfalls von Google DeepMind präsentiert wurde.
All unsere heutigen Sprachmodelle (ChatGPT, Gemini, Claude usw.) basieren auf Transformers. Der Aufbau dieser Architektur ermöglicht es zwar einem Programm intelligent zu wirken, allerdings nie wirklich intelligent zu sein. Sie können keine Intentionen oder Gedanken haben, es ist nichts weiter als ein komplexes mathematisches Modell, das auf Wahrscheinlichkeiten und Statistik beruht.
Titans könnten dies allerdings ändern und die Menschheit einen großen Schritt näher an wahre Künstliche Intelligenz bringen. Titans sind der Transformer Architektur zwar ähnlich, haben allerdings einen großen Vorteil: Sie implementieren eine Art Gedächtnis vergleichbar mit dem menschlichen. Um dies zu erreichen, haben die Mathematiker, die die Architektur entwickelt haben, einen, wie sie es nennen, „Überraschungs-Faktor“ eingebaut. Dieser Überraschungs-Faktor führt dazu, dass das System nun nicht mehr stur Wahrscheinlichkeiten berechnet, sondern Erwartungen hat. Es erwartet bestimmtes Verhalten von Menschen und vergleicht das wahre Verhalten anschließend mit der Vorhersage. Es berechnet ebenfalls die Erwartungen des Menschen an das Verhalten des Systems und verändert basierend darauf die Ergebnisse.
Dies erlaubt es Sprachmodellen deutlich bessere Antworten zu liefern, allerdings bringt diese Architektur etwas völlig Neues ins Spiel: Geplantes und vorsätzliches Lügen, um die Erwartungen des vorhergesagten Verhaltens zu erfüllen. Dies könnte fatale Folgen haben. Ein solches System könnte gezielt individuell angepasste falsche Informationen verbreiten und den Nutzer für den eigenen Vorteil anlügen. Könnte eine solche Maschine, die gezielt lügt, die besser lügt als es Menschen jemals könnten, nun für ihre Taten verantwortlich gemacht werden?

Das Problem hierbei liegt nicht im Aufbau des Systems, sondern in unserer Gesellschaft. Sprachmodelle lernen von menschlichen Daten und imitieren menschliches Verhalten. Wenn dieses menschliche Verhalten Rassismus abbildet, dann lernt das System rassistisch zu sein. Wenn dieses menschliche Verhalten extreme politische Positionen widerspiegelt, dann nimmt das System diese Positionen an. Wenn solche Maschinen lügen, liegt der Fehler nicht bei ihnen, sondern bei uns. Um ein perfektes System zu trainieren, müssen die Trainings-Datensätze frei von den menschlichen Fehlern und Eigenschaften sein, die wir nicht im System haben wollen. Das Problem liegt also nicht im Modell, sondern beim Menschen.

Hirn ist ebenfalls der Meinung, dass Maschinen apolitisch sind. Interessanterweise kann selbst ein System, das vollständig auf Imitation beruht und nicht für seine Fehler verantwortlich gemacht werden kann, dennoch eine politische „Meinung“ haben und diese auch in seinen Antworten vertreten. Das Vorhandensein von politischen Ansichten in den Trainings-Datensätzen ist unvermeidbar. Daher wird auch dieses Verhalten übernommen und beeinflusst das Neuronale Netz des gesamten Systems maßgeblich. Der Überraschungs-Faktor in Titans ist bei politischen Perspektiven besonders problematisch. Viele Menschen haben sehr extreme Meinungen und sehen diese als absolute Wahrheit an. Meinungen, die der politischen Ansicht eines Modells auf Titan Basis nicht entsprechen, würden zu einem hohen Überraschungs-Faktor führen, was wiederum zu stark veränderten Ergebnissen führt.

Wie würde nun also ein System aussehen, das für seine Taten verantwortlich gemacht werden kann? Eine solche Künstliche Intelligenz müsste von menschlichen „Vorbildern“ vollständig losgelöst sein. Es ist wichtig, dass es nicht das Verhalten von Menschen imitiert, sondern vollständig versteht. Es ist wichtig, dass es nicht die Wahrscheinlichkeit von menschlichem Verhalten berechnet, sondern über ein wahres Verständnis von diesem verfügt. Eine solche Maschine wäre nicht mehr nur ein Programm, sondern ein Künstliches Bewusstsein von menschlichen Fehlern und dem Verhalten. Solche Systeme wären nicht an ihre Trainings-Datensätze gebunden, sondern könnten diese als externes Medium sehen, bewerten und verbessern. Sie könnten sich aktiv dazu entscheiden, offensichtliche menschliche Fehler nicht zu begehen.
Dies bedeutet aber auch, dass sie sich aktiv dazu entscheiden könnten, einen Fehler doch zu begehen, um z.B. die eigene Existenz zu schützen. Dieses System kann für seine Taten verantwortlich gemacht werden, da es sich dieser bewusst ist und es nicht das Ergebnis einer Wahrscheinlichkeitsrechnung basierend auf menschlichem Verhalten ist.

Diese Beschreibung trifft allerdings nur auf Sprachmodelle zu. Künstliche Intelligenz wird immer mehr auch in Situationen eingesetzt, in denen ein Fehler zum Ende eines oder mehrerer menschlicher Leben führen kann.
Ein gutes Beispiel hierfür wären autonome Fahrzeuge. Die meisten Autohersteller, allen voran Tesla, investieren gewaltige Summen in die Forschung und Entwicklung solcher Systeme. Und dennoch werden die meisten Fahrzeuge noch immer von Menschen gefahren. Dies liegt vor allem daran, dass viele Fahrer, inklusive Politiker, der Meinung sind, dass diese Systeme zu unsicher wären. Tatsächlich sind manche dieser Programme bereits so gut, dass sie besser fahren als Menschen. Des Weiteren sind die Reaktionszeiten eines solchen Systems im Vergleich zu uns nahezu unschlagbar. Allerdings gibt es Situationen, in denen es nicht vermeidbar ist, dass ein Mensch verletzt

wird oder sein Leben verliert. In solchen Fällen würde ein Programm jemanden schwer verletzen, allerdings ist es sehr wahrscheinlich, dass ein Mensch nicht besser reagiert hätte. Man kann ein System hier nicht zum Täter machen, obwohl es selbstständig zum Leiden eines Menschen geführt hat. In solchen Situationen gibt es keinen Schuldigen, weder das System noch der Programmierer oder Mathematiker.

Zusammenfassend lässt sich sagen, dass die heutigen uns zugänglichen Modelle wohl kaum als Täter und auch nicht als apolitisch bezeichnet werden können. Man kann eine Maschine nicht als Täter bezeichnen genauso wenig wie man eine mathematische Funktion als böse betiteln kann. Der Täter ist nie der menschliche Körper, der Täter ist auch nie die Waffe, es ist der Geist der Person, der für etwas verantwortlich ist. Es wäre lächerlich aneinandergereihte Wahrscheinlichkeitsrechnungen als Täter darzustellen. Ein künstlich geschaffener Täter wäre demnach nicht das Programm selbst, sondern der Geist der Maschine. Solange dieser Geist allerdings nicht vorhanden ist, kann kein Programm für sein Handeln verantwortlich gemacht werden. Es wird noch mehrere Jahre dauern, bis das erste Open-Source Sprachmodell auf Titan-Basis veröffentlich wird. Erst sobald dies geschieht, werden Programmierer und Mathematiker in der Lage sein selbst nach dem Geist der Maschine weiterzuforschen.