Moritz Schmid

Gymnasium Lustenau

Betreuung durch Stephan Schmid

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*

Maschinen in jeglicher Form sollen unmündig sein. Doch was ist Unmündigkeit überhaupt? Und wer oder was erklärt uns Menschen für mündig? Das Gesetz. Laut diesem ist man in Österreich ab 18 mündig. Man kann somit alle Entscheidungen für und über sich selbst treffen und muss auch die Verantwortung für seine Taten tragen. Die wohl am „mündigsten" wirkende unmündige Maschine ist ChatGPT. Sie schafft es, Texte zu generieren, die in ihrer Qualität die vieler über 18-jähriger Menschen übertreffen und die auch auf Anhieb intelligent, erfahren und vor allem mündig wirken. Was macht diese Maschine nun unmündig, wenn sie mündigen Menschen zum Teil deutlich überlegen scheint? Ist es ebenfalls das Gesetz? ChatGPT ist noch ein Kleinkind im Alter von drei Jahren. Wäre es in fünfzehn Jahren mündiger als jetzt? Wohl kaum.

Wer mündig ist, muss sich für seine eigenen Taten verantworten, doch das setzt voraus, dass man in der Lage ist Handlungen zu setzen. ChatGPT scheint das mit der Sprache zu gelingen. Doch sind das wirklich Taten? Die Voraussetzung für Handlungen und Taten ist das Bewusstsein. Es wird also bewusst etwas an der Welt verändert und die Verantwortung für diese Veränderung wird übernommen. Diese Fähigkeit besitzen bisher nur wir Menschen. Niemand würde bei einem Hund, der ein Kaninchen über die Wiese jagt, von einer Tat sprechen. Dem Hund fehlt nämlich die bewusste Entscheidung, das Kaninchen jagen zu wollen. Viel mehr folgt er einfach seinem Jagdinstinkt. Hat ChatGPT diesen Vorsprung des Menschen in nur drei Jahren bereits aufgeholt?

Dazu muss man den Chatbot auf ein Bewusstsein untersuchen? Wittgenstein zufolge erschafft Sprache Wirklichkeit und wer Wirklichkeit wahrnimmt, muss wohl ein Bewusstsein besitzen. Ist die Tatsache, dass ChatGPT intelligente Sätze formulieren kann, also Beweis genug? Wenn dem so wäre, würde sich die Frage stellen, ob die Maschine Meinungen besäße. Um das herauszufinden, muss man einen Blick auf die Funktionsweise werfen. Wie funktioniert das Sprachmodell nun im Innersten? Eigentlich so, wie es wir Menschen tun: mit Neuronen. Die „Künstliche Intelligenz" besteht im Kern nämlich aus einem Netzwerk von Neuronen, einem sogenannten „künstlichen neuronalen Netzwerk". Ein solches KNN besitzt eine Inputschicht, in die ein Wert, in ChatGPTs Fall eben die Eingabe des Nutzers, eingegeben wird. Diese Inputschicht ist mit vielen weiteren Schichten von Neuronen verbunden und alle dieser Verbindungen besitzen sogenannte Gewichte. Diese Gewichte sind eine Art Multiplikator, mit der der vorherige Wert multipliziert wird. So kann man das Signal des Inputs

verstärken und verändern. Nach vielen solcher verborgenen Schichten gelangt man zur Outputschicht. Dort wird der Endwert, der alle Neuronen zuvor durchlaufen hat, ausgegeben. In gewisser Weise handelt es sich bei ChatGPT also um eine riesige, sehr komplexe Funktion. Diese Gewichte beziehungsweise Parameter der Funktion können nun durch große Datenmengen trainiert werden. Sie werden fortlaufend verändert und mit realen Ergebnissen verglichen, um den Output zu verbessern. ChatGPT besteht folglich aus einer Art mathematischer Funktion, die die Wahrscheinlichkeit der Wörter berechnet, die an einen Text angehängt werden. Die Tatsache, dass das so gut funktioniert ist zwar sehr beeindruckend, doch kann man hier von Bewusstsein sprechen?

Dafür müsste dieses System die Sprache, die es verarbeitet, verstehen und dafür würde es folgenderweise über Begriffe verfügen müssen. Begriffe sind Lautfolgen, die wir Menschen gewissen Referenzen gegeben haben und die wir so identifizieren. Diese Begriffe kann man in drei Unterscheidungsmöglichkeiten aufspalten: die Vorstellung, den Sinn und die Bedeutung. Die Vorstellung ist dabei die Reizverarbeitung, also die Verarbeitung der Schallwellen, und die Interpretation. Der Sinn ist die Assoziation des Begriffs in unseren Gedanken und die Bedeutung ist die weltliche Referenz. Verfügt ChatGPT nun über Begriffe? Abgesehen davon, dass der Chatbot keine Wahrnehmung besitzt und somit die Vorstellung sowieso wegfällt, kann er natürlich auch keine Bedeutung verstehen. Wie auch soll er, ohne je einen Apfel gesehen, gespürt, knacken gehört oder gerochen zu haben, wissen was ein Apfel ist. Man könnte zwar davon sprechen, dass er den begrifflichen Sinn besitzt, da er Wörter in vieldimensionale Vektoren umwandelt und ähnliche Wörter im Raum nahe beisammen liegen, was der Idee der Assoziation ähnelt, doch das ist hinfällig, da ohne Bedeutung auch kein Sinn möglich ist. Wie soll man etwas mit etwas anderem assoziieren, wenn man nicht weiß, um was es sich handelt. Einen Apfel mit einem Baum in Verbindung zu bringen, ohne zu wissen, was ein Apfel ist, wird schwierig. Die Vektoren sind also nur Zahlen, die am Schluss, in der Textausgabe wieder in Wörter verwandelt werden. Kant sagt, man müsse ein Individuelles und ein Allgemeines bringen, doch ChatGPT bringt weder das Allgemeine, noch nimmt es das Individuelle wahr. Somit schreibt eine der wohl am fortschrittlichsten entwickelten Sprachgenerierungsmaschinen ihre Texte nicht bewusst und kann folglich, im sprachlichen Rahmen weder Meinungen formulieren, noch Taten oder Handlungen setzen. Stattdessen wählt die Maschine schlicht und einfach das wahrscheinlichste Wort für den bereits vorhandenen Text.

Der zweite Schritt zur Mündigkeit, neben den Handlungen ist die Verantwortung: Um Verantwortung zu übernehmen ist ein Begriff von Wahrheit eine hinreichende Bedingung, denn wer keine Wahrheit besitzt, kann sich nicht für eine solche verantworten. Laut Kant ist ein Begriff die kleinste Einheit, bei dem das Bewusstsein Verantwortung übernimmt. Denn wer einen Begriff auf eine weltliche Referenz anwendet, bezeichnet diese mit ihm und muss folglich die Verantwortung für seine Aussage tragen. Er fällt also ein Urteil. Nun kann man aber irren und etwas irrtümlich mit einem falschen Begriff bezeichnen und das ist, was uns Menschen von den anderen Tieren und - zumindest bisher - von Maschinen abhebt: unsere Fähigkeit zum Irrtum. Diese ist essentiell für einen Wahrheitsbegriff, wie es Donald Davidson mit seiner sprachlichen Triangulation beweist: Davidson geht dafür von einem Dreieck aus. Dabei besteht dieses aus zwei Menschen, die ein Objekt beobachten. Sie reagieren

sprachlich auf das Objekt und die jeweilige Reaktion des anderen. Die Menschen treffen also Aussagen über das Objekt und solange diese sich nicht unterscheiden, ist alles gut. Doch irgendwann kommt es vor, dass Person A eine andere Aussage als Person B tätigt. Die Personen müssen folglich ihre Aussagen mit der weltlichen Referenz abgleichen und kommen zur Erkenntnis, das Person A irrt. Wer ein Konzept des Irrtums besitzt, besitzt auch eines der objektiven Wahrheit und kann deshalb Verantwortung übernehmen oder sich für seine Handlungen verantworten. Für dieses Konzept der Objektivität ist allerdings Sprache nötig. Da ChatGPT aber der Sprache mächtig scheint, stellt sich die Frage, ob es Teil eines solchen Dreiecks sein könnte.

Wie ich aber bereits festgestellt habe, besitzt ChatGPT keine Begriffe. Außerdem fehlt ihm die Fähigkeit, die Welt wahrzunehmen, was die Teilnahme an der Triangulation unmöglich macht. Deshalb kann es zusätzlich seine Aussagen nicht mit der Welt vergleichen und somit niemals an Wahrheit gelangen. Wenn es also nicht über Wahrheit verfügt, kann es sich nicht für die Aussagen über die Realität verantworten, da es diese weder als wahr noch falsch erkennt. Wer nicht versteht, dass er zwei Begriffe in einem Satz verwechselt hat und nicht erkennt, dass er sich irrt, kann die objektive Realität, die darin besteht, dass zwei Begriffe verwechselt wurden, nicht wahrnehmen und sich folglich nicht für diese verantworten und somit nicht mündig sein. Man könnte jetzt argumentieren, dass ChatGPT, wenn es erst auf den Fehler hingewiesen wird, diesen erkennt, doch dem ist nicht so. Stattdessen wird wieder der gesamte Chatverlauf ins KNN eingegeben und anhand der Wahrscheinlichkeit eine Antwort generiert, die so wirkt, als hätte es das Problem verstanden. Doch ohne Bewusstsein und Begriffe ist das im Kern, dem KNN nicht so. Deshalb kann das Sprachmodell weder Taten setzen, noch theoretisch die Verantwortung für Taten übernehmen.

Die letzte Frage, die es zu beantworten gilt, ist, inwiefern die Wahrnahme von Gerechtigkeit für ChatGPT möglich ist. Die Antwort ist einfach: Unmöglich. Wer kein Bewusstsein besitzt, nicht in der Wirklichkeit, sondern nur als Programm existiert und wessen einzige Möglichkeit, etwas wahrzunehmen, die Sprache ist, kann, ohne Begriffe zu verstehen, wohl kaum das Gefühl entwickeln, ungerecht behandelt zu werden. Bereits im Tierreich gibt es keine Gerechtigkeit und Tiere sind sowohl bewusst als auch reale Geschöpfe, die die Umwelt wahrnehmen. Nichtsdestotrotz gilt im Tierreich nur ein einziges Gesetz: das des Stärkeren. Niemand würde einen Löwen als unfair verurteilen, wenn er eine Gazelle reißt. Die Gazelle würde sich wohl ebenfalls nicht als ungerecht behandelt empfinden, da das Konzept der Gerechtigkeit ein rein menschliches ist und wenn doch, wäre sie nicht in der Lage, dies zu äußern. Wer dieses Konzept nicht besitzt, kann somit weder Ungerechtigkeit erfahren noch der Gerechtigkeit bedürfen. Im Allgemeinen ist Gerechtigkeit also ein Gefühl und nur wir Menschen sind in der Lage, unsere Gefühle sprachlich zu äußern.

ChatGPT, wohl unsere „intelligenteste“ Maschine, ist also in dem Sinne völlig unmündig, dass sie weder in der Lage ist, Handlungen zu setzen, noch sich für diese zu verantworten oder sich ungerecht behandelt zu fühlen und folglich ist sie nicht einmal mit einem Baby zu vergleichen, denn selbst dieses nimmt die Umwelt zumindest wahr. ChatGPT hingegen besitzt keine Umwelt, da das Sprachmodell weder an einen bestimmten Ort im Raum gebunden ist, noch über irgendeine Form der Wahrnehmung verfügt.