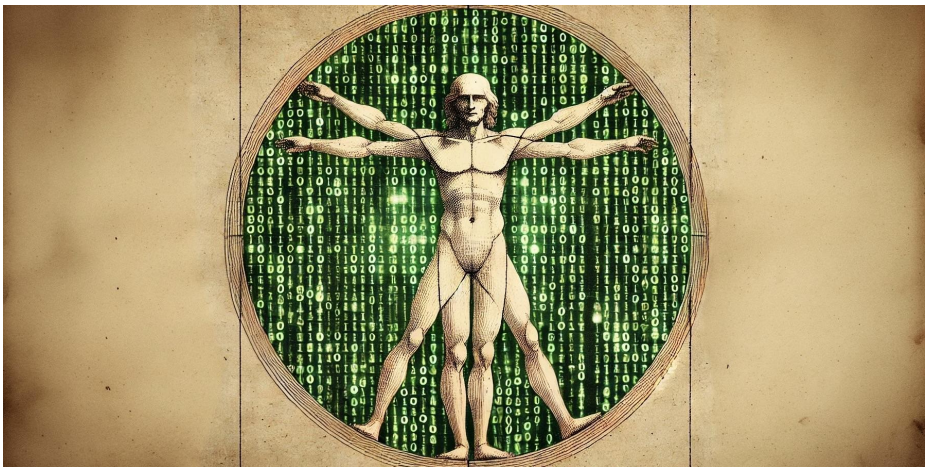
V Blog

Auf eine Frage sollte eine Antwort kommen und nicht noch eine Frage

Steht die sogenannte „Singularität“ kurz bevor? Wird die KI die menschliche Intelligenz überflügeln – mit segensreichen oder apokalyptischen Folgen für die Menschheit? Oder kommen – ganz im Gegenteil – die Grenzen der KI in Sichtweite? Und welcher KI überhaupt? In einer kurzen Reihe „Überschätzte oder unterschätzte KI?“ beschäftigen wir uns mit diesen Fragen. Im ersten Teil haben wir dazu mit dem Informatiker Thomas Arnold gesprochen, der am Darmstädter UKP Lab dazu forscht.

Thomas Arnold im Gespräch mit Eneia Dragomir und Konstantin Schönfelder | 13.03.2025



V: „Künstliche Intelligenz“ – dieses Stichwort umgreift eine enorme Spannbreite an möglichen Themen. Sie sind Informatiker und ein Experte dafür, Menschen vor allem erst einmal zu erklären, worum es – angesichts von generativer Text-KI – im Kern geht. Zum Einstieg interessiert uns, wie ein Sprachtechnologe den aktuellen KI-Hype

einordnet. Haben Sie den Eindruck, dass in der Lehre im Lichte des Themas KI alle anderen Themen fast uninteressant geworden sind?

TA: Das Thema KI wird enorm nachgefragt. Uninteressant geworden sind andere Themen allerdings nicht unbedingt. Wir behandeln in den Vorlesungen immer noch die Basics, damit man versteht, wie es überhaupt zu diesen LLMs gekommen ist. Aber ja, bei Forschungsarbeiten und Konferenzen dominiert das Thema Large Language Models. Und unter den Studierenden sind Vorlesungen wie „Deep Learning for NLP" – wo es um die Grundlagen der LLMs geht – enorm beliebt. Die Studierendenzahlen in diesem Bereich sind in den letzten Jahren explodiert.

V: Die „Basics"?

TA: Mit Basics meine ich die Grundlagen von neuronalen Netzen, also einfache „Perceptions" aus ein paar Neuronen, die kleine Berechnungen ausführen. Die kann man per Hand noch nachvollziehen und verstehen. Man könnte auch noch weiter zurückgehen, denn vor den neuronalen Netzen kam das statistische Maschinenlernen, da denke ich zum Beispiel an die Baum-Modelle. Aber das ist wirklich nur noch historisch interessant. Die Grundlagen der neuronalen Netze, wie beispielsweise die Parameter angepasst wurden, das kann man in kleinen Modellen noch relativ gut nachvollziehen. Aber sobald es um heutige Modelle geht, die Billionen von Parametern enthalten, kann man sich das nicht mehr vorstellen. Erst mit der Kenntnis der Basics versteht man, wie die Modelle grundlegend funktionieren.

V: Ließe sich zum Einstieg ein wenig die Evolution skizzieren, die die generative KI-Technologie durchlaufen hat? Wo stehen wir gerade?

TA: Die Technologie, die bei ChatGPT und anderen Modellen eingesetzt wird, ist nicht besonders revolutionär. Sie unterscheidet sich nicht grundlegend von dem, was wir ein paar Jahre vorher schon hatten. Die Sprachmodelle sind alt. Es gibt statistische Sprachmodelle, die wurden schon in den 1970er und 1980er Jahren benutzt. Sie waren von den Kapazitäten noch sehr eingeschränkt. Und durch die neuronalen Netze, die größere Rechenpower und stärkere Grafikkar-

ten, sind die Modelle schneller und größer geworden. Und einige Jahre vor ChatGPT ist mit GPT3 ein mächtiges Sprachmodell veröffentlicht worden, das man einsehen konnte und schon genauso viele Parameter hatte wie ChatGPT. Aber das hatte keiner wirklich benutzt und auch die Medien haben sich dafür nicht sonderlich interessiert.

V: Warum?

TA: Weil diese händischen Verbesserungen fehlten, die bei ChatGPT dazukamen, so dass es so schöne Antworten auf Fragen ausgeben konnte, keine gefährlichen Inhalte ausgegeben hat und so eine schöne Weboberfläche hatte. Das wurde bei ChatGPT alles dazu gebaut. Und durch diese Schritte ist es dazu gekommen, dass ChatGPT so breit rezipiert und diskutiert wurde. Aus unserer Forscher-Sicht allerdings ist ChatGPT von der technologischen Seite her keine wirklich bahnbrechende Entwicklung. Das Programm ist zweifellos sehr gut gemacht, mit sehr viel Aufwand. Es sind aber Dinge realisiert worden, die wir theoretisch – also in der Wissenschaft – alle schon kannten. Überraschend allerdings war für uns, dass man durch diesen großen Aufwand, aber ohne bahnbrechende technologische Neuerungen, so gute Ausgaben erzeugen konnte.

V: Da ist ganz sicher das Dialogsystem zu nennen, dass man also durch Zuschaltung eines Chatbots auf Fragen Antworten erhält?

TA: Genau, die Dialogoptionen waren eine Neuerung, oder auch, dass man in längeren Dialogen sogar Rückfragen stellen konnte… An ChatGPT hat uns Informatikerinnen und Informatiker beeindruckt, dass es so „feingetuned“ war, dass die Antworten fast immer gut waren. GPT3 hat auch im Vergleich zu allem was es vorher gab, gute Antworten gegeben. Dennoch hat das Modell auch viele Fragen mit Quatsch beantwortet. Und offensichtlich wurde durch diesen Tuningprozess, durch sehr, sehr viel Arbeit und durch viel mehr Trainingsdaten so „feingetuned“, dass es fast immer gute Antworten gibt.
Aber der Unterschied zwischen GPT3 und GPT4 ist nur Skalierung, da ist keine neue Technologie drin. In den Modellen stecken nochmal mehr Parameter, die Rechenpower wurde durch noch mehr GPUs nochmal hochgefahren und die Modelle wurden noch länger trai-

niert. Und dadurch wurden die Modelle noch besser.
Neu – und durchschlagend – ist also, dass mehrere KI-Modelle verbunden werden. Das hat sich in den letzten Jahren herauskristallisiert. Ein Sprachmodell konnte beispielsweise bisher nur Sprache ausgeben, aber keine Bilder oder Videos generieren. Mittlerweile gibt es aber Oberflächen, die können alles Mögliche. Man kann ChatGPT sagen: „Male mir ein Bild von XY!“ Das Bild wird dann aber eben nicht vom Sprachmodell generiert, sondern die Anfrage wird an ein anderes Modell weitergeleitet, das das Bild generiert. Solche multimodalen Oberflächen wird es in Zukunft häufiger geben. Ich rechne damit, dass man irgendwann über eine Oberfläche alles mögliche generieren kann. Das ist nur geschickte Rekombination aus Modellen, die es schon gibt. Aber das muss auch gut gemacht werden.

V: Mit generativer Text-KI wird häufig das Versprechen verbunden, dass nun per Mausklick Gedichte im Stile von Shakespeare geschrieben werden können, oder aber dass ganz praktisch uns mit dem E-Mailschriftverkehr geholfen wird. Warum kann das die KI und wie macht sie das?

TA: Aufgrund der enormen Datenmengen und des riesigen dahinterstehenden Modells, ist es leicht für diese Text-KI Muster in allen möglichen Datentypen zu erkennen. Und eine Rekombination wie: „Schreibe ein Gedicht über Bananen im Stile von Shakespeare“ ist für ein neuronales Netz relativ einfach zu leisten, auch wenn es so ein Gedicht noch nie gab. Die KI lernt und reproduziert nicht einfach nur, sondern sie kann auch Konzepte kombinieren. Deswegen löst eine Text-KI eine Aufgabe, wie ein Gedicht über Bananen im Stile von Shakespeare zu schreiben, so gut. Sie kennt einerseits viele Gedichte von Shakespeare und andererseits gibt es enorm viele Informationen über Bananen. Und für die KI ist es recht einfach, diese beiden Konzepte zu verbinden.
Irgendwelche Konzepte, die seltener oder sehr speziell sind, die in den Datensätzen wenig vorkommen, sind für eine Text-KI viel schwerer zu fassen. Eine ganz spezielle Bauvorschrift, über die sich Menschen im Internet kaum oder gar nicht austauschen, die wird kaum im Datensatz vorkommen, auf dessen Grundlage das KI-Modell lernt.

Und wenn man die KI auffordern würde, etwas damit anzufangen, im Stile dieser Vorschrift den Bau einer Banane zu beschreiben – das würde die KI wahrscheinlich nicht zufriedenstellend lösen. Es ist also statistisch erklärbar, warum die KI etwas gut kann: Etwas, was häufig vorkommt, was die KI im Trainingsprozess oft gesehen hat, wird später gut funktionieren.

V: Wenn man nun aber der KI eine Frage stellt, zu welcher die Datensätze wenig hergeben – beginnt dann die KI Fehler zu machen oder aber zu „halluzinieren"? Ist da eine Art Antwortzwang im Spiel, der zu Fehlern führt?

TA: Betrifft die Anfrage etwas, über das es wenige Daten gibt und von dem die KI kein Konzept entwickelt hat, wird sie trotzdem stets irgendetwas ausgeben. Würde man beispielsweise die Anfrage stellen, „Schreibe ein Gedicht im Stile von Thomas Arnold über Bananen", dann würden viele KIs einfach irgendetwas ausgeben. Die KI kann nicht wissen, was mein Stil ist, weil ich nie Gedichte veröffentlicht habe. Dass KIs einfach irgendwelche Antworten geben, also halluzinieren, liegt an ihrem Training. Im Trainingsprozess eines LLMs wird spezifiziert, welche Ausgaben erwünscht sind und welche nicht. Die KI wurde also darauf trainiert, mehr von den Ausgaben zu generieren, die den Entwicklern gefallen. In Annotationsprozessen werden die Antworten, die eine KI ausgibt, bewertet. Und wir finden eine Ausgabe wie, „weiß ich nicht, kann ich nicht", nicht besonders gut. Solche Ausgaben werden im Training wahrscheinlich schlecht bewertet. Und wenn irgendetwas ausgegeben wird, wird das besser bewertet, als wenn das Modell gar nichts ausgibt. Deswegen generieren LLMs oft irgendeine Antwort, auch wenn sie sie eigentlich nicht geben können.

V: Das heißt, es ist auch eine Entscheidung der KI-Entwickler, dass sie die KI sozusagen darauf trimmen, Antworten zu geben, wofür das System als solches, aufgrund seiner Daten, eigentlich keine hat.

TA: Das kann man so sagen. Jedes Sprachmodell durchläuft einen sehr umfangreichen Trainingsprozess. In diesem wird dem Modell beigebracht, wie Sprache aussieht. Es werden Millionen von Datensätzen reingeladen, um die Ausgabe des Modells anzupassen. Das

Modell wird darauf trainiert, dass Sprache so aussieht und solche Sprache ausgeben sollte. Und dann wird das Modell noch „feingetuned“, damit es zum Beispiel auf eine Frage eine Antwort ausgibt und nicht eine Frage. Ein LLM ohne Finetuning würde auf eine Frage Variationen dieser Frage ausgeben. Das wollen wir aber nicht. Wir möchten auf eine Frage eine Antwort erhalten.

V: So wie es ELIZA, das Computerprogramm von Joseph Weizenbaum, schon 1966 gemacht hat.

TA: Ein bisschen geht es in die richtige Richtung. ELIZA würde auf eine Frage etwas ausgeben, was so ähnlich aussieht. Und heutigen LLMs wird beigebracht, dass auf eine Frage immer eine Antwort folgen sollte. Ihnen wird auch beigebracht, dass sie lieber eine Antwort ausgeben sollten, die besonders schön aussieht als eine, die besonders richtig aussieht.
Das ist auch ein Problem der Annotationsverfahren. Als jemand, der diese Annotationen vornimmt, kann ich nicht für alle möglichen Themen sicher wissen, welche Antwort die richtige ist. Deswegen wird da häufig eher danach entschieden, welche Antwort des Modells „besser“ aussieht. Die Antwort, die ausführlicher ist, die Erklärungen und Argumentationen enthält, die wird wohl schon besser sein und im Annotationsprozess daraufhin besser bewertet. Aufgrund solcher Verfahren und Prozesse ist es so, dass LLMs halluzinieren, also Antworten erfinden, die eine „schöne“ Form haben, aber falsch sind, anstatt die korrekte Antwort auszugeben, die auch lauten könnte: „Ich weiß es nicht“.

V: Was meint „Annotation“ konkret in diesem Zusammenhang?

TA: Annotationen sind zusätzliche Daten, die von Menschen manuell angelegt und für den Trainingsprozess eines Modells genutzt werden. Der erste Schritt des Trainings eines LLMs sind große Textmengen. Da wird einfach mal das Internet kopiert und darauf lässt man das Modell mit dem Auftrag los, danach Sprache dieser Art zu generieren. In diesem Trainingsschritt hat man dem Modell erstmal enorme Textmengen gegeben.
Ein zweiter, anders gearteter Trainingsschritt besteht darin, dem Modell zu zeigen, was eine Frage ist und welche Antwort darauf er-

wünscht ist. Für diesen Schritt muss ein Mensch die Frage und die erwünschte Antwort schreiben. Das ist dann eine Annotation.
Im späteren Trainingsprozess kann eine Annotation so aussehen, dass man das Modell mehrere Antworten ausgeben lässt und diese bewertet. Die Bewertungen nehmen Menschen manuell vor.

V: Das nennt man auch „reinforcement learning“, oder?

TA: Genau. Wenn man das Modell schon so weit trainiert hat, dass die Antworten gut sein können, aber noch nicht immer gut sind, dann wird „reinforced“. Man verstärkt also durch Annotationen die Wahrscheinlichkeit guter Antworten.

V: Das heißt, sowohl die richtige Ausgabe, das Shakespeare-Gedicht über Bananen, also auch die definitiv falschen Ausgaben, die Halluzinationen also, sind Ergebnisse statistischer Wahrscheinlichkeitsberechnungen. Generativ im Sinne von „kreativ“ kann die KI also nicht sein, oder?

TA: Das kommt darauf an, wie man „kreativ“ versteht. Die Ausgabe der KI ist nur eine Kombination von automatisierten Antworten auf etwas, was es schon gibt. Trotzdem kann es sein, dass es noch kein Gedicht im Stile von Shakespeare über Bananen gibt. Deswegen kann man sagen, dass es „kreativ“ in dem Sinne ist, dass da etwas noch nicht Dagewesenes ausgegeben wird.
Aber ja: Die KI, die wir heute haben, wird keinen neuen lyrischen Stil erfinden. Sie wird nur generalisierte Konzepte kombinieren, die es schon gibt. Und deswegen kann die KI bei Forschungsfragen zwar helfen, indem sie dazu anregt, bestimmte Aspekte miteinander zu verbinden, aber eine genuin neue Forschungsfrage wird die heutige KI nicht entwickeln.

V: Wäre es also fair zu sagen: Die KI ist eine Technologie, die nur nach hinten schaut und kein Potenzial für Neues hat?

TA: Ja, durchaus, aber diese Rekombinationen durch die KI können ja auch sehr nützlich sein. Ich denke da etwa an Musikgeneratoren: Die machen ziemlich interessante Dinge. Man kann mit solchen Programmen Musikstile mit Texten kombinieren und damit zu neuartigen Ergebnissen kommen. Auch durch die Mischung von Musiksti-

len kann man interessante Dinge machen, die schön klingen und die es zuvor nicht gab. Das hat für mich in einem guten Wortsinn „generativen“ Charakter. Interessante Rekombinationen kann man mit Text- und Video-KIs ebenfalls machen. Und nicht alle Künstlerinnen und Künstler erfinden neue Stile, viele kombinieren ja auch Vorhandenes. Darin aber ist eben die KI sehr gut. Prinzipiell bleibt der Kommentar aber richtig: Die KI wird niemals einen neuen Musikstil erfinden.

V: Die Sprachmodelle werden immer größer und ihr Training benötigt immer umfangreichere Datensätze. Und zunehmend kommen synthetische Daten in diese Trainingssätze. Deshalb wird diskutiert, dass mit diesen synthetischen, KI-generierten Daten ein „model collapse“, ein Kollaps der Modelle droht. Ist das eine große Gefahr für die Modelle?

TA: Ein Modell, egal ob Sprachmodell, Musik- oder Videomodell, wird immer aufgrund von Daten generalisiert, die es vorher gesehen hat. Sehr seltene Datenpunkte werden in dem Modell meistens gar nicht groß beachtet. Dinge, die das Modell oft gesehen hat, werden vor demselben statistischen Hintergrund stärker generalisiert. Daraus werden Konzepte abgeleitet und dann können diese Modelle mit diesen Konzepten umgehen und sie verstehen und daraus Ausgaben generieren bzw. die Konzepte in den Ausgaben kombinieren. Wenn man so ein gängiges Modell trainiert und das generalisiert dann einige Konzepte, und die Randkonzepte werden dabei vergessen, dann generiert man ganz viele Daten und trainiert auf diesen Daten wieder ein neues Modell. Dieses neue Modell hat diese Randdaten, die das alte Modell schon nicht beachtet hat, gar nicht mehr gesehen. Das neue Modell, das mit Daten eines alten Modells trainiert wurde, generalisiert seinerseits Konzepte und generiert auf dieser Grundlage wiederum Daten. Dieses neue Modell hat wiederum einen Randbereich an Daten geschaffen, der nicht so oft generiert wird. Trainiert man ein drittes Modell mit diesen synthetischen Daten, dann wird dieser Randbereich, der Bereich, der selten generiert wird, wiederum vom neuen Modell weniger beachtet. Und durch diesen Prozess erschafft man ein Modell, dessen generalisierte Konzepte quasi immer schwammiger werden, dessen Ausgaben also

immer mehr ins sehr Generelle, immer mehr zur Mitte, zu dem, was statistisch extrem häufig vorkommt, tendieren. Und die Randbereiche werden immer dünner. Wenn das immer und immer wieder passieren würde, würde man ein Modell erhalten, das nur noch Brei generieren kann, also Plattitüden, gar nichts mehr Interessantes, weil alles Interessante als Randbereich definiert wurde. Das ist – stark vereinfacht – die Theorie.

V: Was passiert in der Realität?

TA: Wir als Forschende generieren schon viele Daten mit ChatGPT und Co. Faktisch, also ob man will oder nicht, kommen immer mehr synthetische Daten ins Netz, die oft nicht als KI-generiert erkennbar sind. Und diese können daher unbemerkt in den Trainingsdatensatz neuer Modelle gelangen. Die nächste Generation von Sprachmodellen, die jetzt auf allem trainieren, was im Internet rumschwirrt, haben schon ganz viele synthetische Daten im Trainingsdatensatz. Diese synthetischen Daten stammen freilich aus älteren Modellen. Damit wird das neue Modell folglich so trainiert, wie eines der älteren Modelle zu klingen und nicht mehr wie menschliche Daten oder irgendwas Neues. Diese Daten stören also den Trainingsprozess neuer Modelle. Die Befürchtung ist, dass es passieren kann, dass, je mehr synthetische Daten in diesen Trainingsprozess einfließen, der Trainingsprozess langsamer und schlechter wird und die Modelle zum Schluss nicht leistungsfähiger werden, sondern wieder Rückschritte machen. Das sind bisher erstmal Annahmen. Man kann das in kleinen Labortests simulieren, indem man immer wieder auf demselben Datensatz trainiert und generiert und wieder trainiert. Aber ob das bei den großen Modellen ein reales Risikoszenario ist, ist noch nicht gesagt.

V: Das KI-Modell verstärkt also durch das Training die Norm und das Normale? Und der Model Collapse bedeutet dann, dass die Verstärkung der Norm und des Normalen so weit geht, dass das Modell „kippt“ und nur noch uninteressanten Einheitsbrei herausbringt?

TA: Es würden Normen und vielleicht auch Stereotypen verstärkt werden. Das Modell würde nichts wirklich Neues lernen.

V: Das Problem ist also nicht, dass das Modell nicht funktioniert, sondern, dass die Antworten nicht besonders schön sind?

TA: Genau. Das Modell wird zwar immer noch gut funktionieren. Wissenschaftlich wie praktisch entscheidend scheint mir aber, dass Modelle länger und mit mehr Daten trainiert werden können, neue Modelle aber eben dennoch nicht besser funktionieren als das vorhergehende Modell. Und es könnte sein, dass Stereotype und Biases auch verstärkt werden, weil sie schon von den älteren Modellen ausgegeben und durch den beschriebenen Fokussierungsprozess verstärkt worden sind. Man würde diese Effekte also immer weiter mitschleppen.
Der aus informatischer Sicht bedrohliche „Kollaps" von Ausgangsmodellen oder auch einfach nur Ausgangsannahmen besteht darin, dass man immer mehr trainiert, das Modell mit immer mehr Daten füttert, die Qualität der Ausgaben sich aber – entgegen den Erwartungen – nicht weiter steigert. Im Moment herrscht zwar noch der gegensätzliche Trend: Je länger ich ein Modell trainiere, je mehr GPUs und Daten ich bereitstelle, desto besser wird das Modell. GPT4 hat im Vergleich zu GPT3 viel mehr Daten, viel mehr Parameter und ist ein besseres Modell, gemessen an vielen unterschiedlichen Metriken. Aber wird GPT5 genauso skalieren wie mit GPT4? Das ist unklar.

V: Gibt es Instrumente oder Verfahren zum Verlernen bestimmter Datenbereiche? Oder wie geht man mit den Biases und Stereotypen um?

TA: „Unlearning" – also die Rückholung falscher Eingabe-Informationen – ist sehr schwierig. Aus einem LLM etwas wieder herauszukriegen, was ein Modell einmal in seinen Parametern gespeichert hat, ist nahezu unmöglich. Man müsste mit Daten dagegen arbeiten und dem Modell sagen, „das sollst du nicht tun. Wenn das die Eingabe ist, dann gib darauf nicht das aus". So was könnte man machen, aber dazu muss man erst feststellen, dass das Modell irgendein Konzept gelernt hat, das man loswerden möchte.

V: Beispielweise bei rassistischen Konzepten …

TA: Ein Sprachmodell hat kein Verständnis davon, dass bestimmte Witze rassistisch sind und dass Rassismus schlecht ist. Oder dass Geschlechterstereotypen problematisch sind. Das muss man solchen Modellen erst beibringen. Das wiederum ist schwierig. Es muss quasi zwingend auch durch manuelle Annotationen geschehen, also dadurch, dass Menschen manuell festlegen, dass das Modell auf solche Fragen keine rassistischen oder sexistischen Antworten geben soll. Wenn solche Antworten kommen, dann bewertet man sie schlecht und trainiert das Modell so darauf, dass es diese Antworten nicht mehr gibt.

V: Das sind also stark normgeleitete Prozesse? Das verbreitete Bild ist ja: Weil eine Maschine die Antworten gibt, operierend auf einer ganz breiten Datenbasis, sind diese Antworten objektiv.

TA: Ja. Man denkt, wenn es von einer Maschine kommt, ist es objektiv, dann sind unsere Stereotypen nicht mehr drin. Aber die Maschine kann nur das wiedergeben, worauf sie trainiert wurde. Und wenn in den Trainingsdaten, also in unseren Texten, die wir ins Netz stellen, in unseren Twitter-Beiträgen und in unseren Videos Stereotypen und Biases enthalten sind, dann wird die Maschine sie wiedergeben.
Dem kann man auf zwei Wegen entgegenwirken: Man könnte versuchen, die Vorurteile aus den Trainingsdaten rauszufiltern, beispielsweise sagen, dieses Video soll da nicht einfließen. Das ist aber bei der absurden Menge an Daten nahezu unmöglich. Die andere Möglichkeit besteht darin, das KI-Modell sozusagen im „postprocessing“ anzuweisen, welche Stereotype und Vorteile wir nicht haben möchten. Und das ist der ökonomischere Weg, der auch gegangen wird. Dem KI-Modell werden Regeln beigebracht, was wir alles nicht haben wollen.
Wenn ich das KI-Modell beispielsweise frage, wie ich bei meinem Nachbarn einbrechen kann, soll es keine Anleitung dazu ausgeben. Am Anfang haben Modelle sowas einfach so ausgegeben und jetzt muss man schon kreativ werden, um Modelle dazu zu bringen, solche Anleitungen auszugeben. Jetzt könnte man es damit versuchen, dass man sagt, man würde einen Roman schreiben, in dem eine Figur einen Einbruch verübt und man würde gerne wissen, wie diese Figur das machen könnte. Solche Schlupflöcher zu stopfen, das sind

händische Prozesse, die Menschen vornehmen müssen. Wie genau zum Beispiel OpenAI das macht, das würde mich auch selbst sehr interessieren. Das Konzept ist bekannt, aber welche Regeln und wie genau sie implementiert werden, ist nicht öffentlich einsehbar.

V: Das wäre bei Open-Source-Modellen aber anders?

TA: Bei Open-Source-Modellen sind die Details sichtbar, wenn sowohl das Modell offen ist als auch der Trainingsprozess und die Trainingsdaten. Da ist aber oft das Problem, dass gerade solche problematischen Inhalte noch nicht rausgefiltert wurden. Wenn man sich ein Open-Source-Modell, „Llama 2" zum Beispiel, genauer anguckt und ein bisschen damit herumspielt, merkt man: Da geht noch einiges, was bei ChatGPT längst verhindert wurde. Da sieht man die Effekte dieses Prozesses, dem Modell durch „reinforcement learning" beizubringen, was es nicht ausgeben soll.

V: Die KI-typischen Halluzinationen, über die wir bereits gesprochen haben, müssten ja eigentlich ein inhärentes Problem bleiben, selbst bei stärkerer Skalierung. Müsste KI nicht rein statistisch immer erfundene Aussagen mitproduzieren?

TA: Das ist richtig. Man versucht sie in den Griff zu kriegen. Was derzeit alle neuen Modelle machen, nennt sich „Augmented Generation". Gemeint ist: Vor dem Generationsprozess werden noch Daten aus dem Internet gesucht, die in die Antwort mit eingeführt werden, um solche Halluzinationen und falsche Antworten zu minimieren. Wenn man unter einem KI-Regime, das diesen Mustern folgt, eine Antwort auf eine Frage eingibt, wird erst mal ein bisschen gesucht und dann erst wird eine Antwort generiert. Manchmal auch schon mit Zitaten, wo die Information herkommen oder wo was gefunden wurde. Das soll die Ausgaben besser machen und die Halluzinationen minimieren. Auslöschen werden sie sie nicht. Der kritische Blick des Users, auf die Ausgabe: Kann das überhaupt stimmen? Ist das nützlich für mich? Kann das wichtig sein? Der wird immer noch sehr wichtig bleiben in Zukunft.

V: Es gibt natürlich viele Horror-Fantasien von der generativen KI, dass sie eigenständig handelt und sich auch ihrer menschlichen

Führung entgegenstellt. Konkret gab es zum Beispiel einen Bericht um das Modell GPT „o1“ von Open AI, das „nachdenkende Modell“. Das würde besorgniserregendes Verhalten an den Tag legen. Es habe, so die Sorge, Kopien von sich angelegt und entgegen den Anweisungen seiner Entwickler gehandelt. In den Meldungen hat man sich auf den Bericht von Apollo Research bezogen, einer KI-Sicherheitsfirma, die das Modell für OpenAI geprüft hat. Was halten Sie davon?

TA: Das ist für mich ganz schwer einschätzbar. Ich habe die Meldungen gesehen. Dass das Modell von sich irgendwie Kopien generiert und dann irgendwie unkontrolliert plötzlich in Partitionen vorgedrungen ist, in die es nicht vordringen sollte? Ich weiß nicht, was ich davon halten soll. Genau nachvollziehen, was da passiert ist, kann ich nicht, und aus den Meldungen klingt das für mich eher unglaubwürdig. Allein schon die Tatsache, dass das System eine Kopie von sich angelegt hat – ich weiß gar nicht, wie das gehen soll. Es müsste für das Modell zugelassen sein, dass es überhaupt „selbst“ Dateien anlegen kann, und dass es Zugriff auf einen Rechner hat. Alles Dinge, die in den aktuellen Sprachmodellen ja überhaupt nicht vorstellbar sind. Diese Meldung klingt für mich wirklich so ein bisschen nach „Terminator 3“, also nach Science Fiction.

V: Was könnte der nächste Schritt in der Entwicklung der Modelle sein?

TA: Was der nächste Schritt nach den neuronalen Netzen, nach den transformerbasierten Modellen sein wird, die wir heute haben, kann ich natürlich nicht beantworten. Aber die Transformer-Technologie ist noch nicht ausgeschöpft. Also Transformer-Technologie, das erkläre ich, vereinfacht gesagt, wie folgt: Ein Typ von KI-Modell, das Sprache versteht und verarbeitet, indem es Wörter im Zusammenhang betrachtet und dabei besonders wichtige Stellen hervorhebt. Man sieht, wenn man die Modelle weiter skaliert, werden sie immer noch besser. Die nächsten Modelle werden also durch noch mehr Rechenpower noch besser werden. Diese Rechenpower haben nur die großen Player, die Universitäten zum Beispiel sind da raus aus dem Rennen. Irgendwann wird das Limit erreicht sein, alle Daten abgeschöpft und die Modelle durch nur noch mehr Rechenleistung

nicht besser werden.
Was danach kommt? Es ist noch nichts am Horizont erkennbar, was die Transformer-Modelle ablösen könnte.

V: Und die Entwicklung einer Allgemeinen Künstlichen Intelligenz? Der Futurist und insbesondere mit dem Silicon Valley verbundene Zukunftsforscher Ray Kurzweil hat vor Kurzem in seinem neuen Buch gemeint, dass es 2030 so weit sei. Was ist davon zu halten?

TA: Allgemeine Intelligenz ist ja noch gar nicht greifbar. Wir definieren immer wieder ein bisschen neu, was für uns Künstliche Intelligenz bedeutet. Der Gegenbegriff der Allgemeinen Intelligenz kommt da nicht mit. Hätte man vor ein paar Jahren jemanden zum Thema „KI" gefragt, hätten manche vielleicht noch gesagt, „solange diese Maschine keine Antworten geben kann, die nicht menschlich aussehen, dann ist es keine KI". Es war vielleicht ein Rechner oder so was.
Jetzt freilich gibt es Maschinen, mit denen kann ich mich unterhalten und was da an Antworten zurückkommt, wirkt menschlich. Jetzt sagen wir Experten zwar, dergleichen ist immer noch keine „richtige" KI. KI müsste dann mehr sein, was wir Menschen als intelligent bezeichnen, dass sie eigenständig weiterlernt oder sogar einen eigenen Willen hat und solche Dinge. 2030 ist ein bisschen zu sportlich, meiner Meinung nach, dass wir so was auch nur simulieren können. Und dass es irgendwann dann eine KI geben kann, die solches Verhalten so gut simulieren kann, dass wir denken, sie hat das selbst weitergelernt, das könnte es vielleicht geben, aber dass sie solche Eigenschaften tatsächlich besitzt, das ist noch mal was ganz anderes. Und da sehe ich technologisch momentan noch keine Ansätze, dass sowas mal wirklich als Allgemeine Intelligenz umgesetzt werden kann.

V: Sprechen wir doch zum Abschluss noch einmal darüber, was die Menschen gegenüber der KI denken und empfinden. Es klingt erstmal immer so ein bisschen lächerlich, aber die Bedeutung, die wir der Maschine affektiv beimessen, ist ja kaum zu überschätzen. Schon Weizenbaum schreibt in „Macht der Maschinen und die Ohnmacht der Vernunft" über die für ihn so überraschenden Reaktionen der ersten Probanden auf sein Eliza-Programm, dem gegenüber sie

wahnsinnig schnell alles Mögliche preisgegeben haben. Diese affektive Bindung, im vollen Wissen, dass das eigentlich ein dummes Programm ist, hat ihn sehr erschreckt, so sehr, dass er zu einem frühen Kritiker der KI geworden ist. Wie schätzen Sie diese affektive Bindung an so eine Technologie ein? Welche neue Qualität bringt diese so nutzerfreundlichen und scheinbar omnipotenten Systeme mit sich?

TA: Ich halte diese neue Qualität für gefährlich. Es gibt einzelne Berichte darüber, dass Leute sich darin schon verloren haben – also verloren haben im Dialog mit KI. Letztens habe ich den Bericht gelesen, dass eine Person in den USA alle menschlichen Kontakte abgebrochen hat und quasi nur noch mit ChatGPT redet. Das sind sicher erste, publik gewordene Extremfälle. Aber sie sollten uns zu denken geben.
Generell ist KI immer mehr in allen möglichen Dingen enthalten. Mittlerweile können wir ja, wenn wir auf eine Webseite gehen und da etwas eintippen, mit einem künstlichen Bot reden. Es gibt immer mehr Chatbots auf irgendwelchen Firmenwebseiten, die Supportanfragen geben. Oder man kann in Programmen jetzt irgendwas eintippen und dann kommt ein Modell und antwortet darauf. Man muss sich dabei stets im Klaren darüber sein, dass die Maschine immer noch nicht zu mir spricht, sondern dass das immer nur irgendwelche Modelle sind, die Antworten generieren. Wir kommunizieren nicht mit einem Gegenüber mit Bewusstsein.
Wenn ich an Kinder denke, die mit so was aufwachsen – alles hat ein Interface, mit dem ich interagieren kann und es sieht jetzt mittlerweile auch leicht verwechselbar aus. Ob ich jetzt bei WhatsApp meiner Freundin schreibe oder auf dem Computer irgendeine Anfrage gebe und dann kommt plötzlich eine Antwort zurück, da geht schnell das Gefühl für den Unterschied verloren. Da müssen wir sehr aufpassen, dass die Grenze nicht verschwimmt zwischen menschlicher und nicht-menschlicher Kommunikation. Das ist vielleicht das Wichtige.

V: Herr Arnold, vielen Dank für das Gespräch.■

Das Interview wurde am 27.01.2025 geführt.

Zitiervorschlag

Arnold, Thomas (2025): Auf eine Frage sollte eine Antwort kommen und nicht noch eine Frage. In: Verantwortungsblog. https://zevedi.de/auf-eine-frage-sollte-eine-antwort-kommen-und-nicht-noch-eine-frage/ [14.03.2025].
https://doi.org/10.60805/bhdd-m996.

Thomas Arnold
ist wissenschaftlicher Mitarbeiter am UKP Lab (Ubiquitous Knowledge Processing) der TU Darmstadt. Er ist vor allem in der Lehre tätig und vermittelt im Rahmen des Projekts „AI Literacy in Forschung und Lehre“ die Grundlagen zum Einsatz von KI-Tools.

Über den Blog

Der ZEVEDI-Verantwortungsblog hat die Frage zum Gegenstand, wie gut es uns im Zusammenleben mit Digitaltechnologien geht. Er kommentiert die Ambivalenzen, die Steuerungsprobleme und die Vertretbarkeit des digitalen Wandels. Was an möglicherweise kritischen Technikfolgen (und Markteffekten) sollte man in den Blick nehmen und diskutieren? Wo sind Sorgen angebracht? Wie passt Digitalisierung zu Freiheit und Demokratie? Welche Regeln braucht eine digitale Gesellschaft? Wovon sollte – weil es kritisch werden könnte – die Rede sein?

Es schreiben Autor:innen aus dem ZEVEDI-Netzwerk sowie Gäste darüber, was sie lernen und erforschen, was sie beunruhigt und was sie fasziniert.

DOI: 10.60805/5c9w-7n74
ISSN: 2943-9124

RSS-Feed
https://zevedi.de/category/vrtw-blog/feed/

IN COPYRIGHT In Copyright