Auf Worte folgen Taten

Unter dem Schlagwort „Agentic AI“ sollen auf Basis großer Sprachmodelle autonom handelnde KI-Systeme entstehen. Doch was steckt wirklich hinter den vollmundigen Ankündigungen der Techkonzerne?

Kaum jemand hat damit gerechnet, dass der Turing-Test so schnell geknackt sein würde. Dabei markieren die großen Sprachmodelle mit ihrer nahezu perfekten Simulation menschlicher Sprache wohl nur den Anfang der Entwicklung einer neuen Art künstlicher Intelligenz. Das Wissen über die Welt, das sie während ihres Trainings verinnerlicht haben, prädestiniert sie nämlich für Höheres: Sie sollen die „Gehirne“ autonom handelnder KI-Agenten werden, die sich in der digitalen Welt frei bewegen, um ihren Nutzern dienlich zu sein. Für einfache Konsumenten soll das mit Einkaufen, dem Buchen von Reisen oder dem Ausfüllen von Formularen beginnen.

Der aktuelle Hype um den Begriff „Agentic AI“ dient dabei vor allem als Narrativ, um Risikokapital anzuziehen. Dabei wurden Programme, die eigenständig handeln, Ziele verfolgen und zumindest rudimentär auf Wahrnehmungen reagieren, in der KI-Forschung eigentlich schon immer als „Agenten“ bezeichnet. Neu ist lediglich, dass die aktuellen Entwicklungen auf den Fähigkeiten großer Sprachmodellen aufbauen können. Doch wie wird aus einem Algorithmus, der eigentlich immer nur auf einen Prompt hin das wahrscheinlichste nächste Wort ausspuckt, ein autonomer Akteur?

Spielende Sprachmodelle

„Ausgangspunkt für die Entwicklung komplexer KI-Agenten ist die Erkenntnis, dass sich natürliche Sprache sehr gut als Austauschformat zwischen den einzelnen Komponenten eines Agent eignet“, sagt David Schlangen, Professor für Grundlagen der Computerlinguistik an der Universität Potsdam. „Die Zukunft der Sprachmodelle liegt also nicht unbedingt darin, mit Menschen zu sprechen, sondern Aufgaben zu erledigen.“ Schlangen selbst forscht gewissermaßen an der einfachsten Form von sprachmodellbasierten Agenten: reine Sprachmodelle, die Wortspiele wie Taboo oder 20 Questions spielen. Anders als bei ihrem ursprünglichen Einsatz als reine Antwortmaschinen ist es bei solchen Spielen nämlich bereits erforderlich, Strategien zu entwickeln und ein Ziel zu verfolgen. Die Modelle bekommen dafür auch kein spezielles Training sondern lediglich einen Anfangsprompt, der die Aufgabe schildert.

Um messbar zu machen, wie sich die Sprachmodelle bei solchen Aktivitäten schlagen, haben Schlangen und sein Team mit „clembench“ eine eigene Benchmark entwickelt. Dabei hat sich herausgestellt, dass vor allem das sogenannte Reasoning, bei dem die Modelle eine Art inneren Dialog mit sich selbst führen, bevor sie ihre finale Antwort geben beziehungsweise ihren Spielzug setzen, die Leistungsfähigkeit deutlich gesteigert hat. Doch auch wenn sie sich langsam dem menschlichen Niveau annähern, machen sie immer noch Fehler, die Menschen bei solchen Aufgaben eher selten unterlaufen. Vor allem wenn es darum geht, über mehrere Spielzüge hinweg Informationen zusammenzuführen und sich strategisch zu verhalten, zeigen sie noch deutliche Schwächen. So scheitern selbst die besten der getesteten Modelle bei Taboo oft noch daran, auch die vorangegangenen Hinweise zu berücksichtigen und sich nicht nur die letzte Beschreibung des gesuchten Begriffes zu merken. Und bei 20 Questions erinnern die Fragen der Modelle die Forscher oft an die von Kindern. So verzichten sie etwa darauf, den Suchraum mit geschickten Fragen vorsichtig einzuschränken und wollen stattdessen mit Fragen wie „Ist es ein Elefant?“ sofort zur Sache kommen.

Dennoch schreibt Schlangen den großen Sprachmodellen für Steuerung komplexer Agentensysteme, bei denen sie auch Zugang zu externen Werkzeugen haben, großes Potenzial zu.

„Da diese Modell bereits ein gewisses Weltwissen kodiert haben, kann man sie auf erstaunlich flexible Weise auf Aufgaben ansetzten, die mit früheren Architekturen unlösbar gewesen wären“, sagt der Forscher. Eine einfache und mittlerweile in vielen gängigen Sprachmodellen bereits standardmäßig integrierte Form eines solchen externen Werkzeugs, ist die Internetsuche: Sobald die Chatbots feststellen, dass sie die Frage eines Users nicht direkt aus ihrem neuronalen Netz heraus beantworten können, aktivieren sie eine Suchmaschine, um an die nötigen Informationen zu kommen.

Da Sprachmodelle von ihrem grundlegenden Aufbau her nichts anderes können, als auf Prompts hin neuen Text zu produzieren, verlaufen auch die Interaktionen mit ihren externen Werkzeugen auf diese Weise. „Für das Modell selbst spielt es keine Rolle, ob es einen Text für Menschen produziert oder Programmcode erzeugt, der dann extern ausgeführt wird“, erklärt Schlangen. Schließlich wurden die meisten Chatbots neben menschlicher Sprache auch auf Unmengen an Programmiercode trainiert. Dass sie dabei auch gleich noch von menschlichen Programmierern gelernt haben, Code mit besonderen Zeichen wie Backticks zu markieren, um ihn von ihren in natürlicher Sprache verfassten Kommentaren abzugrenzen, spielt der Entwicklung von KI-Agenten zusätzlich in die Hände. Denn so ist es für ein übergeordnetes Programm ein Leichtes, die Codebausteine in der Ausgabe eines Sprachmodells abzufangen und extern ausführen zu lassen. Das passiert etwa immer dann, wenn ein Chatbot von einem Nutzer eine mathematische Aufgabe gestellt bekommt, die er selbst nicht lösen kann. Das Ergebnis der externen Berechnung wird anschließend einfach wieder in Form eines neuen, internen Prompts an den Chatbot zurückgespielt, der ihn dann mit gewohnter Eloquenz in seine finale Antwort einbaut.

„Im Prinzip lassen sich so beliebig lange Konversationen aufrechterhalten, bei denen das Sprachmodell nicht mit einem Menschen, sondern mit Programmen kommuniziert und je nach Rückgabewert vielleicht weitere Programme aufruft“, sagt Schlangen. Allerdings müssen dafür die Modelle darauf getrimmt werden, immer das richtige Tool für die jeweilige Aufgabe auszuwählen und dieses dann auch korrekt anzusprechen. Denn oft handelt es sich dabei um herkömmliche Computerprogramme, die genau definierte Eingaben erfordern und jede Abweichung würde zu Fehlern führen. Um das zu schaffen bekommen die Sprachmodelle in der Regel zu jedem Prompt eines Nutzers automatisch auch noch einen versteckten Systemprompt mitgeschickt, in dem die ihm zur Verfügung stehenden, externen Werkzeuge aufgelistet sind.

Besonders weit fortgeschritten sind solche Agentensysteme bereits im Bereich des Software-Engineerings, wo das Sprachmodell mit einer lokalen Computerumgebung kommunizieren kann, um etwa Dateien zu erzeugen, Programme auszuführen oder Fehlermeldungen zu analysieren. Aber auch Anwendungen, wie sie etwa Salesforce mit seiner Plattform Agentforce anbietet, scheinen sich auf dem Vormarsch zu befinden. Sie ermöglicht es Unternehmen im E-Commerce, maßgeschneiderte KI-Agenten einzurichten, die Aufgaben wie die Bearbeitung von Produktanfragen oder Garantiefällen übernehmen und dabei auf externe Dienste wie etwa Zahlungsabwicklungen zugreifen. Und auch Microsoft zeichnet das Bild einer Zukunft, in der Unternehmen mit Copilot Studio ihre ganz eigenen KI-Agenten entwickeln können – Systeme, die selbstständig handeln, Entscheidungen treffen und sogar ganze Gruppen weiterer Agenten koordinieren. „Ob man Sprachmodelle, die ja kein echtes Verständnis haben, wirklich auf Werkzeuge loslassen will, die in der digitalen oder echten Welt potenziell auch Schaden anrichten könnten, ist eine spannende Frage, die sich in den nächsten Monaten und Jahren klären muss“, mein Schlangen.

Wer übernimmt die Verantwortung?

Diese Problematik beschäftigt auch Judith Simon, Professorin für Ethik in der Informationstechnologie an der Universität Hamburg und stellvertretende Vorsitzende des

Deutschen Ethikrates. Sie forscht zu ethischen und politischen Fragen, die sich im Zusammenhang mit Künstlicher Intelligenz ergeben und. „Ich bin zwar überzeugt, dass das sehr schnell vermarktet werden wird. Gleichzeitig wird es aber mit großen Problemen verbunden sein“, sagt die Wissenschaftlerin. Wie schnell sich eine Technologie, die mit ihren Nutzerinnen und Nutzern auf einfache Weise in natürlicher Sprache interagiert, auch bei Alltagsanwendungen durchsetzen kann, haben Alexa, Siri und schließlich ChatGPT bereits gezeigt. Zudem suggeriert die intuitive Art der Kommunikation Vertrauenswürdigkeit, nicht zuletzt weil die Systeme in der Regel auch darauf getrimmt sind, ihre User zu bestärken und ihnen etwa ständig zurückspiegeln, wie toll und klug ihre Gedanken sind. „Dieses kommunikative, menschlich wirkende Element führt dazu, dass Menschen den Systemen glauben“, sagt Simon. „Das könnte sich genauso auch auf Agentenmodelle übertragen.“

Ob die Systeme dieses Vertrauen auch wirklich verdienen, steht freilich auf einem anderen Blatt. Denn hinter jedem Agentensystem stecken Anbieter wie Google oder Meta, die in den üblichen Plattformlogiken denken: Nutzerinnen und Nutzer auf der einen Seite, Werbung und Unternehmen auf der anderen. So bieten beispielsweise schon einfache Hotelbuchungsplattformen zwar unbestreitbare Vorteile, etwa weil sie Preisvergleiche und damit bessere Entscheidungen ermöglichen. Allerdings werden sie eben nicht nur von den Nutzern bezahlt, sondern auch von den Anbietern, die sich damit Sichtbarkeit erkaufen. Dieselbe diese Logik könnte es auch bei KI-Agenten geben, wenn ihnen Kaufentscheidungen überlassen werden. „Womöglich werden wir auch hier bald einen systematischen Interessenkonflikt zwischen Nutzern und Anbietern haben“, warnt Simon.

Ein weiteres, tiefer liegendes Problem steckt allerdings in der Funktionsweise der Sprachmodelle selbst. Im Gegensatz zu herkömmlichen Computerprogrammen, die ihren Code exakt nach Punkt und Strich abarbeiten, funktionieren sie nämlich auf der Basis von Wahrscheinlichkeiten. Sie haben während ihres Trainings gelernt, menschliche Texte statistisch auszuwerten und das wahrscheinlichste nächste Wort auf einen Prompt zu finden. Natürlich sind Wahrscheinlichkeiten aber immer mit Unsicherheit verbunden. „Spricht man im Zusammenhang mit Sprachmodellen von „Halluzinationen“ oder „Fehlleistungen“, täuscht das darüber hinweg, dass diese Systeme eigentlich genau das tun, für den sie gebaut worden sind“, sagt Simon. „Nur werden sie leider allzu oft für Aufgaben eingesetzt, für die sie eigentlich nicht geeignet sind.“ Und da Sprachmodelle die Grundlage für die neuen Agenten bilden, überträgt sich dieses Problem zwangsläufig auch auf sie.

Solange KI-Agenten nur triviale Aufgaben wie Restaurant- oder Friseur-Buchungen erledigen, sind die Konsequenzen überschaubar. Außerdem kommt bei aktuellen Systemen in der Regel spätestens bei der Zahlung ohnehin wieder der Nutzer ins Spiel. „Oft werden diese Systeme aber als Zukunft der Suche, als Analysewerkzeuge oder in Wirtschaftskontexten vermarktet, wo es auf Wahrheitsgehalt und Genauigkeit ankommt“, gibt Simon zu bedenken. Und sollten sie in einem nächsten Schritt tatsächlich auch Finanztransaktionen durchführen, ist damit zu rechen, dass es irgendwann zu schwerwiegenden Fehlern kommt. Dann wird die Frage entscheidend sein, ob und wie weit sich die Anbieter von der Haftung für negativen Konsequenzen der Nutzung ihrer KI-Agenten freimachen können.

Bisher sind die großen Techfirmen jedenfalls eher für ihre Tendenzen bekannt, Technologien nach dem Motto „move fast, break things“ zu entwickeln – also neue Technologien schnell auf den Markt zu werfen und Kosten sowie Konsequenzen auf die Allgemeinheit zu verlagern. Bei vielen Sprachmodellen ist in den kleingedruckten Nutzungsbedingungen zu lesen, dass die Anbieter keine Verantwortung für Fehler übernehmen. „Entstehen Schäden, wird die Schuld an die Nutzerinnen und Nutzer zurückgegeben“, sagt Simon. „Bei KI-Agenten bin ich mir unsicher, ob das auch so bleiben wird – und was passiert, wenn die ersten gravierenden Fehler auftreten“,

Doch auch die Gefahr, dass das Delegieren von Entscheidungen an KI-Agenten unehrliches Verhalten bei ihren Nutzer fördert, sollte nicht unterschätzt werden. „Je stärker man Dinge auf Distanz erledigt, desto leichter fällt Betrug“, sagt Simon. Zwar fiel es Menschen immer schon leichter, Regeln zu beugen oder zu brechen, wenn jemand anderes die Handlung ausführt. Eine unter der Leitung des Max-Planck-Instituts für Bildungsforschung durchgeführte und erst kürzlich in Nature veröffentlichte Studie, zeigt allerdings, dass diese Tendenzen durch den Einsatz von Künstlicher Intelligenz noch verstärkt werden können. Das gilt insbesondere dann, wenn die User ihren Agenten lediglich übergeordnete Ziele wie „Maximiere meinen Profit“ zu geben brauchen und sie nicht konkret zu unehrlichen Handeln auffordern müssen. Die in der Studie festgestellte Bereitschaft der KI-Agenten, im Dienste ihrer Anwender zu betrügen, zeigt aber auch, dass es dringend technischer Schutzmechanismen und regulatorischer Rahmenbedingungen braucht, bevor die Technologie ich der Breite bei den Menschen ankommt. Und wie die Forscher betonen, werden wir uns als Gesellschaft wohl auch damit auseinandersetzen müssen, was es eigentlich bedeutet, moralische Verantwortung mit Maschinen zu teilen.

Mit Worten zu besseren Strategien

Trotz aller Risiken und ungeklärter Fragen scheint es der natürliche Lauf der Forschung zu sein, große Sprachmodelle nach ihrem Durchbruch als Antwortmaschinen nun als Grundlage für die Entwicklung einer neuen Generation von KI-Agenten zu nutzen. „Ich glaube, dass das Denken im Sinne von verschiedenen Agenten und deren Zusammenspiel die Informatik langfristig verändern wird“, sagt Professor Kristian Kersting, der an der Technischen Universität Darmstadt das Fachgebiet für Maschinelles Lernen leitet. Anstatt voreilig Produkte zu entwickeln betreibt er mit seinem Team allerdings Grundlagenforschung zur Frage, wie Sprachmodelle die Leistungsfähigkeit von KI-Agenten erhöhen können. Dafür greift er auf eine bereits lange in der Forschung etablierte Testplattform zurück: Atari-Spiele aus den 80er Jahren.

Bereits 2015 hat Google DeepMind in diesem Bereich einen Meilenstein gesetzt, als sein deep Q-network (DQN) bewies, dass es eine ganze Reihe dieser Spiele-Klassiker von Tennis, über Pong bis zu Space Invaders auf oder über menschlichem Niveau spielen konnte. Die DQN-Agenten brauchten dafür keinerlei Hintergrundinformationen zum jeweiligen Spiel, sondern lernten ohne weiteres menschliches Zutun lediglich aus den Pixeln des Bildschirms, sich in ihrer digitalen Umgebung zurechtzufinden und den Punktestand in die Höhe zu treiben. Einige Spiele wie etwa Seaquest, bei dem sie ein U-Boot steuern sollten, das regelmäßig auftauchen muss, um Sauerstoff zu tanken, während es gleichzeitig Gegner bekämpft und Taucher rettet, überforderten DQN jedoch.

„Es ist ziemlich schwierig, aus den einzelnen Pixeln des Bildschirms zu verstehen, dass ein Balken, der immer kürzer wird, einen schwindenden Sauerstoffvorrat darstellt“, sagt Kersting. „Wenn ich möchte, dass ein Agent seine Spielumgebung wirklich versteht, reichen diese Informationen nicht aus.“ Um den von ihm und seinem Team um Hikaru Shindo unter der Bezeichnung „Blend RL“ entwickelten KI-Agenten diesbezüglich auf die Sprünge zu helfen, setzten die Forscher deshalb auf Sprachmodelle. Denn erst wenn ein System Begriffe wie „U-Boot“, „Taucher oder „Sauerstoff“ nutzen kann und ein gewisses „Verständnis“ dafür hat, was sie aussagen, ist es auch in der Lage, die Szenen am Bildschirm einzuordnen, Zusammenhänge zu erkennen und langfristige Strategien zu entwickeln. „Schließlich ist es eine der faszinierendsten Fähigkeiten von Sprachmodellen, dass sie zumindest einfache logische Schlussfolgerungen mittlerweile meistens recht gut hinbekommen“, sagt Kersting.

Allerdings lässt sich die Komplexität eines Atari-Spiels kaum mit der Realität vergleichen. Bei vielem von dem, was die Techfirmen in Bezug auf KI-Agenten versprechen, dürfte es sich deshalb wohl auch noch eher um Marketing handeln - vor allem wenn suggeriert wird, dass sie in einem allgemeinen Kontext einsetzbar sind und in jeder Situation richtig funktionieren. „Ich bin froh, dass

diese Forschung beginnt“, sagt Kersting. „Man sollte allerdings lieber keine voreiligen Fantasien von Allmacht oder vollständiger Problemlösung entwickeln.“