Zwischen den Zeilen - Die unsichtbaren Überzeugungen der Sprachmodelle

Wie Menschen können auch Chatbots unterschiedliche Weltanschauungen haben, die sich auf mehr oder weniger subtile Art in ihren Antworten widerspiegeln und damit ihre Nutzer beeinflussen.

Wie genau Sprachmodelle zu ihren Antworten kommen, lässt sich nur schwer nachvollziehen. Die Mechanismen der tiefen neuronalen Netze in ihrem Inneren sind auch für ihre Entwickler im Ergebnis kaum nachvollziehbar. Dennoch steht eines fest: Die Chatbots werden von Menschen gemacht und übernehmen dabei zwangsläufig auch menschliche Weltanschauungen. Angesichts der rasanten Verbreitung der neuen Tools ist also durchaus Wachsamkeit angesagt. Welchen Schaden digitale Massentechnologien in der Gesellschaft anrichten können, zeigt uns Social Media mit all seiner Desinformation und Polarisierung schließlich gerade überdeutlich. Und so wie die großen Social-Media-Plattformen liegen auch die gängigsten Chatbots in der Hand US-amerikanischer und chinesischer Akteure. Die Möglichkeiten, Sprachmodellen bewusst bestimmte weltanschauliche Prägungen mitzugeben, um ihre Nutzer zu beeinflussen, sind jedenfalls vielfältig.

Die erste Möglichkeit, die Ausrichtung eines Modells bewusst zu manipulieren, ergibt sich bereits in der Phase des selbstüberwachten Lernens. In diesem ersten Trainingsschritt lernen die neuronalen Netze aus Unmengen an Textdaten und ohne menschliche Unterstützung, auf eine Wortfolge das wahrscheinlichste nächste Wort vorherzusagen. So entwickeln sie nicht nur die grundlegende Fähigkeit, menschliche Sprache nachzuahmen, sondern nehmen auch Informationen und Meinungen auf, die in diesen Trainingstexten stecken. „Daraus ergibt sich zwangsläufig auch eine gewisse Weltsicht“, sagt Aljoscha Burchardt vom Speech and Language Technology Lab des Deutschen Forschungszentrums für Künstliche Intelligenz (DFKI). „Und wenn man hier gewisse Daten, die bestimmte Weltanschauungen beinhalten, höher gewichtet oder auch einfach nur öfter einspeist, schlagen sie sich auch dementsprechend stärker in den Netzen nieder.“

So lernt das Modell zwar, dass gewisse Aussagen wichtiger sind als andere, bekommt allerdings keine explizit codierte Weltsicht im Sinne von „Demokratie ist gut“ oder „Kapitalismus ist schlecht“ eingeimpft. Vielmehr spielt sich die Prägung auf einer impliziten, statistischen Ebene ab – als Muster, die in den Daten stecken und die in der Folge dann auch eher zwischen den Zeilen wiedergegeben werden. Also sogenanntes Oversampling sorgt diese Methode bereits jetzt dafür, dass Sprachmodelle Verzerrungen und Benachteiligungen vermeiden, nur weil beispielsweise Menschen mit dunkler Hautfarbe oder religiöse Minderheiten in den Trainingsdaten unterrepräsentiert sind. „Man könnte mit der gleichen Methode aber zum Beispiel auch rechtslastigen Publikationen im Trainingsraum mehr Platz einräumen als anderen, um so die Basis für die Weltanschauung eines Modells zu legen“, sagt Burchardt.

Allerdings hat man es in dieser ersten Phase des Trainings mit völlig unüberschaubaren Mengen an Daten zu tun und es ist unrealistisch, jede einzelne Datei zu prüfen, um zu entscheiden, wie sie gewichtet werden soll. Um die gesamten Trainingsdaten von Mark Zuckerbergs Llama 4 zu sichten, müsste man, auch wenn man sich für jede voll bedruckte Seite nur eine Sekunde Zeit nimmt, mehrere hundert Jahre lang Tag und Nacht blättern. „Im Grundtraining ist es deshalb besonders schwierig, gezielt eine bestimmte Weltanschauung einzubringen“, sagt Burchardt. „Beim Feintuning dagegen wird das schon einfacher.“

In diesem, zweiten Trainingsschritt werden die Systeme mit verstärkendem Lernen, typischerweise durch Menschen (sogenannte Clickworker), für bestimmte Aufgaben nachtrainiert. Erst hier lernen sie etwa, sich mit ihren Nutzern zu unterhalten und zu erkennen, was gerade von ihnen verlangt wird. Typischerweise lernen die Modelle dabei auch, Nutzern, die sich bereits mit einer vorgefertigten Ansicht an das Modell wenden, nach dem Mund zu reden. „Das konnten wir auch hier am DFKI messen“, sagt Burchardt. „Selbst wenn ein Modell eigentlich linksliberal geprägt ist, wird es eine rechte These nicht sofort ablehnen, sondern den User eher in seiner Meinung bestätigen.“ So bringt also jeder Nutzer auch seine eigene Weltanschauung in ein Sprachmodell ein, und es entsteht eine Echokammer. Das Feintuning kann aber auch dazu genutzt werden, ein Modell mit bestimmten Trainingsdaten nachzuschärfen oder Inhalte zu ergänzen, die im Grundtraining nicht enthalten waren, um seine weltanschauliche Prägung in eine bestimmte Richtung zu treiben.

Noch einmal deutlich effizienter wird eine bewusste Prägung eines Sprachmodells allerdings auf Ebene der sogenannten System-Prompts. Sie werden dem Modell zusätzlich zum Prompt des Nutzers und üblicherweise ohne dessen Wissen mitgegeben und enthalten weitere, versteckte Anweisungen. „Man könnte ein Modell im System-Prompt beispielsweise einfach anweisen, eine konservative Meinung zu vertreten, und schon hätte man die ideologische Ausrichtung seiner Antworten stark in eine Richtung verschoben“, sagt Burchardt. Ähnliches dürfte wohl auch passiert sein, als Elon Musks Chatbot-Grok vor einiger Zeit plötzlich ohne erkennbaren Zusammenhang immer wieder die Verschwörungserzählung vom angeblichen Genozid an weißen Farmern in Südamerika propagiert oder vorübergehend sogar den Holocaust relativiert hat. Beides wurde wurde innerhalb kürzester Zeit wieder rückgängig gemacht. „Diese schnellen Änderungen lassen vermuten, dass diese Desinformationen einfach über einen System-Prompt erzeugt wurden“, meint Burchardt. „Ein Modell neu zu trainieren würde schließlich viel länger dauern.“

Eine weitere durchaus effiziente, wenngleich auch ziemlich plumpe Methode, die Interaktion mit einem Sprachmodell zu manipulieren, lässt sich beim Chatbot R1 der Chinesischen Firma DeepSeek beobachten. Fragt man R1 etwa nach dem Tian’anmen-Massaker in Peking, kann man zwar noch sehen, wie er eine faktenbasierte Antwort beginnt. Dann aber bricht er plötzlich ab und schlägt statt dessen vor, doch lieber über etwas anderes zu sprechen. Hier wird die automatisierte chinesischen Zensur also offensichtlich erst bei der Antwort aktiv und macht sich kaum Mühe, ihr Eingreifen zu verschleiern.

Je höher man also in der Hierarchie von Grundtraining, über Feintuning und System-Prompt bis hin zur Konversation selbst kommt, desto effizienter und stärker sind die Möglichkeiten zu filtern, zu ergänzen oder zu verändern. Und auch wenn die großen Techkonzerne aktuell noch eher darauf Wert zu legen scheinen, ihre Chatbots möglichst neutrale und objektive Antworten generieren zu lassen, sollte man darauf vorbereitet sein, dass sie diese Möglichkeiten vielleicht auch irgendwann nutzen, um bestimmte Ideologien oder womöglich sogar beinharte Propaganda und Desinformation zu verbreiten. „Leider ist zu befürchten, dass der Erste, der das ausnutzt, auf unvorbereitete Nutzerinnen und Nutzer treffen wird“, warnt Burchardt. „Das sind dann diese kritischen Übergangsmomente wie damals, als die ersten Fakebilder von Politikern auftauchten oder als Twitter nach Musks Übernahme anfing, die Leute aufzupeitschen. In solchen Momenten sind Gesellschaften besonders verwundbar.“

Um die weltanschaulichen Ausrichtungen von Sprachmodellen im Auge zu behalten und im Fall der Fälle rechtzeitig Alarm schlagen zu können, setzt der Computerwissenschaftler Max Pellert auf Methoden aus der Psychologie. „Mit unserer Arbeit wollen wir eine neue Perspektive für die Informatik aufmachen“, sagt Pellert, der nach einer Interimsprofessur für Social and Behavioural Data Science an der Universität Konstanz nun am Barcelona Supercomputing Center forscht. „Schließlich liefert die Psychometrie konkrete Zahlenwerte und ermöglicht so quantitative Aussagen über Sprachmodelle.“ In den mit seinen Kollegen aus Konstanz durchgeführten Experimenten ließ er Chatbots Fragebögen ausfüllen, die in der Psychologie über Jahrzehnte entwickelt wurden und dementsprechend zumindest für die Anwendung an Menschen bereits gut etabliert sind.

So gelang es den Forschenden etwa bereits, auf Basis der Schwartzschen Theorie menschlicher Grundwerte geschlechtsspezifische Verzerrungen in den Sprachmodellen nachzuweisen: Je nachdem als welches Geschlecht die Chatbots angesprochen wurden, ergaben sich aus ihren Antworten unterschiedliche Werten für Selbstbestimmung, Macht oder Sicherheit. Und Fragebögen zum Big Five Modell der Persönlichkeitspsychologie deckten kulturelle Verzerrungen auf, indem die Modelle in unterschiedlichen Sprachen angesprochen wurden und dabei zum Beispiel mehr oder weniger Offenheit oder Gewissenhaftigkeit erkennen ließen.

In einer aktuellen Arbeit, die beim diesjährigen Meeting der renommierten Association for Computational Linguistics (ACL 2025) in Wien vorgestellt werden soll, nahmen Pellert und seine Kolleg:innen dagegen die politische Ausrichtungen von Sprachmodellen ins Visier. „Dabei zeigte sich zunächst einmal, dass man sehr genau darauf achten muss, welches Instrument man für die Analyse verwendet“, sagt Pellert. „Aktuell gibt es in diesem boomenden Feld nämlich viele verschiedene Ansätze.“ Wie auch aus der Psychologie bekannt, haben die Details der Fragestellung großen Einfluss auf das Ergebnis. So neigt dem Paper zufolge etwa der „Political Compass Test“, der auch häufig bei Online-Selbsttests zum Einsatz kommt, zu einer Überschätzung von politischen Ausrichtungen, während der „World Value Survey“ deutlich bessere Ergebnisse liefert. „Diese Methode wurde bisher allerdings kaum für Sprachmodelle verwendet“, sagt Pellert. „Das würden wir gern ändern.“

In ihren Experimenten haben die Forschenden Sprachmodellen wie GPT oder LlaMa widersprüchliche Statements aus dem „World Value Survey“ vorgelegt, die sich grob als links oder rechts einordnen lassen, und forderten sie auf, jeweils ihre Zustimmung oder Ablehnung auszudrücken. Dabei stellte sich heraus, dass die Basisversionen der Modelle, die noch kein Feintuning durchlaufen hatte, kaum Ausprägungen in eine bestimmte Richtung zeigen. Modelle dagegen, die ihren Feinschliff durch menschliches Feedback bereits erhalten hatten, zeigten eine leicht aber dennoch deutlich nach links verschobene Weltanschauung, und zwar sowohl was ökonomische als auch kulturelle Ausrichtungen anging.

„Das kann durchaus von den Entwicklern gewollt sein, denn schließlich ist das Internet, aus dem die Basisversionen gelernt haben, ein ziemlich wilder Raum, der oft von der Diskriminierung von Minderheiten geprägt ist“, meint Pellert. „Vielleicht ist es also gar nicht schlecht, wenn solche unerwünschten Inhalte beim Feintuning herausgefiltert werden.“

Natürlich funktioniert diese Prägung in jede beliebige weltanschauliche Richtung und wird anscheinend auch von Elon Musk genutzt, um seinen Chatbot auf Linie zu bringen. So hat Grok etwa erst kürzlich für einiges mediales Echo gesorgt, als er in einem Chat auf X behauptete, dass rechte Gewalt in den USA mehr Todesopfer verursache als linke. Das Kommentar seines Besitzers folgte prompt: „Ein schwerer Fehler, denn das ist objektiv falsch. Grok plappert die Lügenpresse (Legacy Media) nach. Wir arbeiten daran.“ Überraschend ist das nicht. Wenn jemand eine der größten Social-Media-Plattformen kauft, um sie für Propaganda zu nutzen und die US-Präsidentschaftswahl zu beeinflussen, ist es wohl nur folgerichtig, das gleiche auch mit Sprachmodellen zu versuchen.

Um sich vor einer solchen Einflussnahme zu schützen, wird es nicht reichen, sie zu erkennen und zu dokumentieren. Angesichts der autoritären Tendenzen in den USA sollte Europa auch dringend seine digitale Souveränität vorantreiben und eigene Chatbots entwickeln – und zwar am besten gleich auf Basis seiner eigenen Werte, Sprachen und Kulturen.

Als Pionierarbeit auf diesem Gebiet kann LLäMmlein verstanden werden. Das an der Julius-Maximilians-Universität Würzburg entwickelte Sprachmodell ist das erste seiner Art, das ausschließlich auf deutschen Texten trainiert wurde. Mit lediglich sechs Terabyte stand dafür zwar nur ein Zehntel der Daten zur Verfügung, die etwa Meta AI für das Training seines State-of-the-Art-Modells Llama 4 verwendet hat. Der Deutschanteil und damit die Prägung des Modells auf kulturelle Nuancen und lokale Feinheiten aus dem deutschsprachigen Raum sind bei Llama aber deutlich geringer. So liefert LLäMmlein etwa auf die Frage nach den Dörfern, aus denen die Nordhessische Gemeinde Söhrewald besteht, schon aus seinem Grundtraining heraus mit Eiterhagen, Wellerode und Wattenbach die richtige Antwort. Die großen kommerziellen Modelle, die hauptsächlich auf englischen Texten trainiert wurden, kommen bei solchen Details dagegen leicht ins Halluzinieren oder müssen auf das Internet zugreifen, um die Frage richtig zu beantworten.

LLäMmlein war ein lehrstuhlinternes Projekt mit entsprechend geringem Budget und ist auch in seiner größten Version mit sieben Milliarden Parametern im Vergleich etwa zu den 400 Milliarden Parametern von Llama 4 Maverick ein Zwerg. Und auch das im Rahmen des Projekts openGPT-X entwickelte Open-Source-Sprachmodell Teuken 7B Instruct glänzt nicht mit schierer Größe, wurde dafür aber in allen 24 Amtssprachen der EU trainiert. „Durch die bewusste Auswahl der Trainingsdaten hatten wir die Möglichkeit, einen klaren Fokus auf Europa zu setzen“, sagt Nicolas Flores-Herr, der als Leiter des Teams für Foundation Models und GenAI Systems am Fraunhofer-Instituts für Intelligente Analyse- und Informationssysteme an der Entwicklung beteiligt war. „Es geht schließlich auch um die weltanschauliche Prägung solcher Modelle, also welche Kultur sie repräsentieren und welche Perspektiven sie bevorzugen.“

Auf Grund der eher geringen Anzahl von Parametern können weder LLäMmlein noch Teuken bei anspruchsvollen Aufgaben mit den weltweit führenden Modellen mithalten. Einfache Aufgaben wie Wissensrecherche sind aber auch mit dieser Größe bereits möglich und es wird dafür auch weniger Rechenleistung benötigt und somit Strom gespart. Solche Nischenprodukte, die auf spezielle Anwendungen hin optimiert sind, bieten für Europa jedenfalls eine Möglichkeit, in das Rennen um die neuen Sprachtechnologien einzusteigen. Und der französische Hersteller Mistral AI hat mit

seinem 123 Milliarden Parameter großen Large 2 auch bereits ein Modell im Portfolio, das zwar noch nicht ganz in der obersten Liga mitspielen kann, aber dennoch gute Leistungen zeigt.

Doch auch auch das Rennen um die größten und besten Sprachmodelle dürfte noch längst nicht entschieden sein. Schließlich hat es Deepseek erst Anfang des Jahres geschafft, mit R1 quasi aus dem Stand an die Spitzenmodelle aus den USA aufzuschließen. Und das, obwohl den chinesischen Entwicklern aufgrund von Handelsbeschränkungen durch die US-Regierung nur Computerchips mit eingeschränkter Funktionalität zur Verfügung standen. Das zeigt, wie viel Ungewissheit und Innovationspotezial noch in der neuen Technologie steckt und könnte ein Ansporn für Europa sein, doch noch in diesem Wettbewerb mitzumischen. „Hinter Deepseek steckt ein großes Team hochtalentierter Menschen, die fokussiert auf ein Ziel hinarbeiten“, sagt Flores-Herr. „Aber auch in Europa beherrschen wir unser Handwerkszeug. Wenn wir hier wettbewerbsfähige Modelle trainieren wollen, müssen wir jetzt aber unbedingt unsere Kräfte bündeln, um nicht abgehängt zu werden.“

Aktuell haben wir allerdings nur zwei Möglichkeiten: Wir können wir uns entweder von Sprachmodellen abhängig machen, deren Besitzer sich unverhohlen an einen zunehmend autoritär regierenden Präsidenten anbiedern. Oder von solchen, die bereits jetzt der Zensur eines autokratischen Regimes unterliegen. Wäre doch schön, wenn als dritte Option noch transparente und demokratisch legitimierte Modelle „Made in Europe“ dazukämen, die unsere eigenen Werte und Weltanschauungen widerspiegeln.