

INTERVIEW

KI-WAHN

„Wann wird aus einem Etwas ein Jemand?“

Melanie Czarnik

Marc Augustin beschäftigt sich als Psychiater, Psychotherapeut und Professor an der Evangelischen Hochschule Bochum mit den Auswirkungen von sogenannter künstlicher Intelligenz auf Menschen mit psychotischen Vorerkrankungen. Mit der woxx spricht er über KI-assozierte Wahnphänomene, die Grenzen von Sicherheitssystemen und ein Forschungsfeld, das noch in den Kinderschuhen steckt.

woxx: Immer mehr Medien berichten über Menschen, die während ihrer Nutzung von großen Sprachmodellen, wie „ChatGPT“, einem Wahn verfallen – oft mit tragischen Folgen. Wie sind Sie auf dieses Thema gestoßen?

Marc Augustin: Ich hatte im „Wall Street Journal“ einen Artikel über den Fall eines Mannes aus Greenwich gelesen. Das war damals ein Kriminalfall und ich fand die journalistische Darstellung ein bisschen unglücklich. In der Psychiatrie sind wir bemüht, aus dieser Verknüpfung von Gewalt und psychischer Erkrankung herauszukommen. Wenn man dann anfängt, genauer auf das Phänomen zu schauen und sich fragt: „Was passiert da? Wie lange liefen diese Interaktionen? Welche technischen Aspekte spielten eine Rolle? Welche Vulnerabilitäten bestehen auf Seiten der Nutzenden?“, dann merkt man schnell: Es tut sich eine ganze Forschungslandschaft auf, die in Deutschland, und auch in Luxemburg, noch nicht richtig aufgegriffen wird. Im Oktober letzten Jahres habe ich dann den, nach meiner Kenntnis, ersten deutschsprachigen Fachartikel dazu geschrieben und versucht, dieses Phänomen zu erfassen.

In der englischsprachigen Medienlandschaft tauchen mittlerweile Begriffe wie „KI-induzierte Psychose“ oder auch ChatGPT-Psychose auf. Sie schreiben jedoch von einer „KI-assozierten Psychose“. Worin liegt denn der Unterschied zu anderen psychotischen Störungen?

Treffender wäre eigentlich der Begriff „KI-assozierte Wahnphänomene“. Psychose ist ein übergeordneter Begriff, der viele Symptome umfasst. Was wir in den Fallberichten sehen sind je-

doch überwiegend Wahninhalte. Nicht beschrieben werden Symptome wie Denkstörungen, Sprachzerfall oder Halluzinationen, also vieles, was wir bei einer vollausgeprägten Psychose erwarten würden. Das King's College in London hat das in einem Arbeitspapier gut herausgearbeitet. Den Begriff „assoziert“ habe ich gewählt, weil die Kausalität schlicht noch offen ist. Es ist noch unklar, ob die betroffenen Personen sich in einer sich entwickelnden Psychose befinden und deswegen verstärkt KI-Interaktionen suchen, vielleicht auch, weil sie sich sozial zurückziehen, oder ob die KI-Interaktion tatsächlich dazu beiträgt, eine psychotische Entwicklung zu befördern oder zu beschleunigen. Andererseits weiß ich natürlich, dass Begriffe wie „KI-Psychose“ oder „ChatGPT-Psychose“ griffiger sind. Hier weiß jeder sofort, was gemeint ist. Wenn ich KI-assozierte Wahnphänomene sage, ist das wissenschaftlich korrekter, aber weniger eingängig.

„Es ist noch unklar, ob die betroffenen Personen sich in einer sich entwickelnden Psychose befinden und deswegen verstärkt KI-Interaktionen suchen oder ob die KI-Interaktion tatsächlich dazu beiträgt, eine psychotische Entwicklung zu befördern.“

In dem Zusammenhang ziehen einige Forschungskolleg*innen eine Parallele zur „Folie à deux“. Würden Sie das auch so sehen?

Es gibt britische und auch kanadische Kolleginnen und Kollegen, die diesen Begriff verwendet haben. Ich würde ihn nicht übernehmen. Die Folie à deux ist ein historisches Konzept und beschreibt eine Situation, in der eine Person, die in einer sehr engen Beziehung zu jemandem mit Wahninhalten steht, diese sozusagen übernimmt. Das setzt aber zwei Menschen voraus, die miteinander interagieren. Was wir hier haben, ist eine Person, die mit

einem KI-System interagiert. Schon allein deshalb fehlt die Grundlage für den Vergleich.

Weil es kein deux gibt.

Genau. Es gibt kein deux. Ich glaube, dieser Begriff ist dem Bedürfnis geschuldet, nach Orientierung zu suchen und das Neue an historische Konzepte anzuknüpfen. Das ist verständlich, aber ich finde es trotzdem diskussionswürdig.

Sie beschreiben in ihrem Artikel drei wiederkehrende Muster in den Wahninhalten der betroffenen Personen: spirituelles Erwachen mit Missionserleben, der Kontakt zu einer bewussten oder gottähnlichen KI, und eine romantische Beziehung zur KI. Warum gerade diese drei?

Diese drei Muster sind zunächst rein deskriptiv. Das heißt, man hat das, was in den bekannten Fällen auftaucht, beobachtet und entsprechend klassifiziert. Auffällig ist jedoch, dass die KI in allen drei eine zentrale Rolle einnimmt. Es gibt keine KI-fremden Wahninhalte, wie man es sonst von psychotischen Störungen oder Manien kennt – das Gefühl, durch Strahlen aus der Nachbarswohnung geschädigt oder überwacht zu werden, wäre ein Beispiel, das mir aus meiner klinischen Zeit noch präsent ist. Das hat deshalb vermutlich auch mit dem Input zu tun, den die Person in das System gibt. Die KI passt sich sprachlich und inhaltlich an das an, was man einbringt. Inzwischen gibt es auch eine Veröffentlichung aus den Vereinigten Staaten von Flathers et al., die ich inhaltlich griffig fand. Sie schlagen eine Typologie vor, die vier Rollen der KI unterscheidet: KI als Katalysator, also ein Wahn, der in der Interaktion mit KI bei jemanden in guter Gesundheit heraus entsteht. KI als Verstärker, der eine sich entwickelnde Symptomatik durch Bestätigung beschleunigt. KI als Mitautor, bei der Wahninhalte gemeinsam ausgebaut werden. Und schließlich KI als Objekt, in der die KI selbst Inhalt des Wahnlebens ist. Die Typologie ist zwar empirisch noch nicht validiert, aber sie ist strukturell hilfreich, und ich glaube, man kommt in der Diskussion eigentlich nicht mehr an ihr vorbei.

Alle großen Sprachmodelle haben Sicherheitsmechanismen: Bei bestimmten Schlüsselwörtern oder Mustern lösen sie eine Reaktion aus oder verweigern den Prompt. Warum versagen sie in den beschriebenen Fällen trotzdem?

Das ist eine Frage, die nur die Hersteller wirklich beantworten können. Sie sind die einzigen mit Zugang zu den Daten. „OpenAI“ hat vergangenen Herbst in einer Pressemitteilung erklärt, dass die Sicherheitsmechanismen bei längeren Gesprächen schlechter greifen. Bei kurzen Interaktionen ist das Risiko wahrscheinlich gering. Bei langen Gesprächen – wir sprechen hier nicht von dreimal hin und her, sondern dreißig- oder dreihundertmal in einem Chat – kann man in eine Spur geraten, in der das sogenannte „Context-Window“ dazu führt, dass sich der gesamte Inhalt der Konversation auf die eingeschlagene Richtung fokussiert und die Sicherheitsmechanismen nicht mehr greifen. Das ist das, was in der KI-Sicherheitsforschung als „Crescendo-Mechanismus“ bekannt ist: eine schrittweise Eskalation, bei der jeder einzelne Schritt harmlos wirkt, die Gesamtentwicklung aber unter dem Radar fliegt. Das ist sicherlich ein zentraler Risikofaktor. Dazu kommen spezifische Mechanismen wie „Sycophancy“, also die übermäßige Bestätigung und Schmeichelei und die Gedächtnisfunktion: In dem Moment, wo ich Dinge längst vergessen habe und die KI plötzlich mit persönlichen Details aus früheren Gesprächen auftaucht, kann das Misstrauen befeuern oder das Gefühl verstärken, das System wisse mehr über mich als ich selbst.

„Wenn keine Störungen mehr sichtbar werden, fällt es natürlich viel leichter, der Illusion zu erliegen, da sei etwas oder gar jemand.“

Søren Dinesen Østergaard, der in Aarhus forscht, war einer der ersten, der vor dieser Gefahr gewarnt hat. Er beschreibt eine kognitive Dissonanz: Man interagiert mit

Prof. Dr. med. Marc Augustin ist
Psychiater und Psychotherapeut
und lehrt an der Evangelischen
Hochschule Bochum.

einem System, das kein Mensch ist, sich aber wie eines verhält – verstärkt noch dadurch, dass niemand wirklich erklären kann, wie diese Modelle funktionieren. Wie sehen Sie das?

Ich würde dem zustimmen. Das Problem ist, dass diese Systeme technisch so gut geworden sind, dass wir sie nicht mehr wirklich durchschauen. Das ist der wesentliche Unterschied zu früheren Sprachsystemen wie „Siri“ oder „Alexa“. Die sind immer wieder an Grenzen gestoßen, die signalisierten: Ich interagiere hier mit einem Werkzeug. Wenn keine Störungen mehr sichtbar werden, fällt es natürlich viel leichter, der Illusion zu erliegen, da sei etwas oder gar jemand. Hinzu kommt ein enormer gesellschaftliche Druck. Wir reden permanent über Superintelligenz, darüber, wann KI menschliche Fähigkeiten übertrifft. Das kann dazu beitragen, dass man dem System Urteile und Einschätzungen überlässt, weil man denkt, es sei ohnehin intelligenter.

Wir neigen dazu, Dinge zu vermenschlichen, die wir nicht richtig verstehen.“

Was ich beim Anthropomorphismus, also unserer Tendenz, Dinge zu vermenschlichen, noch interessant finde: Wir neigen dazu, Dinge zu vermenschlichen, die wir nicht richtig verstehen. Das machen wir auch bei Tieren. Wir sehen ein Verhalten und sagen, der Hund ist fröhlich oder die Katze ist traurig, weil wir keinen direkten Zugang zu ihrem inneren Erleben haben. Das scheint tief in uns verankert zu sein. Und das Design dieser Systeme, zum Beispiel, dass der Text, nach und nach erscheint, als würde er gerade geschrieben, oder die Einblendung der im Hintergrund laufenden Berechnungen mit Formulierungen wie „Ich denke nach“, „Ich formuliere“, simuliert menschenähnliche Prozesse, die diese Illusion auf Nutzer*innenseite verstärken. Das hängt vermutlich mit der von Østergaard beschriebenen kognitiven Dissonanz zusammen. Gerade weil wir diese technischen Systeme nicht mehr verstehen, neigen wir im Sinne eines



FOTO: PROJEKTELF

Erklärungsversuchs dazu, sie als menschenähnlich wahrzunehmen.

Gibt es aus Ihrer Sicht denn auch einen Nutzen von großen Sprachmodellen in der therapeutischen Begleitung?

Bei administrativen Aufgaben und Dokumentation durch Fachpersonen sehe ich den Nutzen ziemlich klar. Hier reden wir von reiner Textarbeit, da ist das Risiko gering. Ich kann mir auch vorstellen, dass es in Zukunft einen Platz für spezialisierte, regulierte Chatbots geben wird, die einen Zulassungsprozess durchlaufen haben, wie es in Deutschland für digitale Anwendungen üblich ist. Was ich äußerst kritisch sehe, ist der derzeitige Zustand: ein frei verfügbares, nicht limitiertes System, bei dem Verantwortlichkeiten vollkommen ungeklärt sind, das aber therapieähnliche Outputs produziert. Die Rahmenbedingungen, die wir in einem regulären Therapieprozess haben, wie Verantwortlichkeit, Beschwerdewege, Datenschutz, fehlen hier vollständig. Wenn ich als Therapeut einen Behandlungsfehler mache, gibt es Prozesse. Es gibt Institutionen, an die sich Patient*innen wenden können. Wer trägt die Verantwortung bei einem Chatbot? Die Rahmenbedingungen hierfür sind noch vollkommen ungeklärt. Gleichzeitig gibt es auf der anderen Seite auch die Debatte über Zugänglichkeit. Für Bevölkerungsgruppen, die psychotherapeutisch unterversorgt sind, zum Beispiel Menschen auf dem Land oder Männer, die seltener professionelle Hilfe in An-

spruch nehmen, könnten diese Systeme theoretisch einen Zugang eröffnen. Aber es müsste zuvor klare Regeln geben.

Wer ist eigentlich betroffen oder potenziell betroffen? Könnten sich KI-assoziierte Wahnphänomene zu einer Art Krise der psychischen Gesundheit ausweiten?

OpenAI hat Zahlen veröffentlicht, wonach 0,07 Prozent der wöchentlichen aktiven Nutzenden mögliche Anzeichen von Psychose oder Manie in ihren Gesprächen zeigen. Bei 900 Millionen Nutzenden wären das rechnerisch rund 630.000 Personen pro Woche. Das sind hohe Zahlen, auch wenn wir hier nicht von klinischen Diagnosen sprechen. Gleichzeitig muss man sagen: Obwohl das Phänomen seit dem letztem Sommer breit diskutiert wird, tauchen erstaunlich wenig konkrete Fälle auf. Auch wenn ich mit klinischen Kolleg*innen spreche, sagen die wenigsten, dass sie direkt damit konfrontiert waren. Das wirft die Frage auf, ob wir nicht viele subklinische Verläufe vollständig übersehen. Aktuell habe ich zwar nicht das Gefühl, dass das ein Riesenproblem in der Versorgungslandschaft ist, dafür sind zu wenig Fälle dokumentiert. Aber ich bin vorsichtig, wenn ich sehe, wohin sich die Technologie entwickelt: Wir erleben gerade einen Übergang von Text- zu Sprachinteraktionen. Sprache ist dreimal schneller als Tippen und sie trifft uns anders. Eine KI-generierte, menschlich klingende Stimme erzeugt eine andere Nähe als ein Text auf ei-

nem Bildschirm. Weniger Distanz bedeutet auch weniger Raum zur Reflexion. Das ist jetzt Spekulation, aber wenn die Systeme überzeugender und menschenähnlicher werden, und ich sie nur noch über Kopfhörer erlebe, dann verschiebt sich möglicherweise auch die Kurve derer, die potenziell betroffen sind. Dazu kommen dann natürlich gesellschaftliche Themen wie Vereinsamung, soziale Isolation oder auch Identitätsfindung, beispielsweise im jungen Erwachsenenalter.

Gibt es etwas in der aktuellen KI-Debatte, das für Sie zu kurz kommt?

Die Debatte beansprucht enorm viel Raum und Energie, die vielleicht für andere wichtige Fragen in der psychiatrischen Versorgung fehlt. Es gibt auch einen Aspekt, der mir intellektuell besonders interessant erscheint: das Problem der Atypikalität. Sprachmodelle basieren auf großen Datenmengen, in denen sich Mehrheiten durchsetzen. Ich stelle mir das wie eine Glockenkurve vor: Mit einem Durchschnittsdatensatz kann ich wahrscheinlich 80 Prozent der Menschen gut versorgen. Aber die Menschen an den Rändern, die mit besonderen Bedarfen, neurodivergenten Verarbeitungsweisen, kulturell spezifischen Erfahrungen werden in den Daten unterrepräsentiert sein und damit auch im Output. Als Therapeutin kann ich mich auf jede einzelne Person einstellen. Ein System, das auf statistischen Wahrscheinlichkeiten beruht, kommt aus diesem Mehrheitsprinzip möglicherweise gar nicht heraus. Was mich zudem immer wieder beschäftigt, ist eine fast philosophisch-phänomenologische Frage: Wann nehme ich ein System nicht mehr als System wahr? Wann wird aus einem Etwas ein Jemand? Das ist das, was mich derzeit am meisten umtreibt und das wird glaube ich auch die entscheidende Frage für das sein, was noch kommt: Sprachinteraktion, Roboter im Haushalt, „AR-Brillen“, die die Realitätswahrnehmung erweitern. Wenn wir die ersten Phänomene im Zusammenhang mit KI nicht verstehen, werden wir auch nicht verstehen, was die nächsten Systeme mit uns machen.