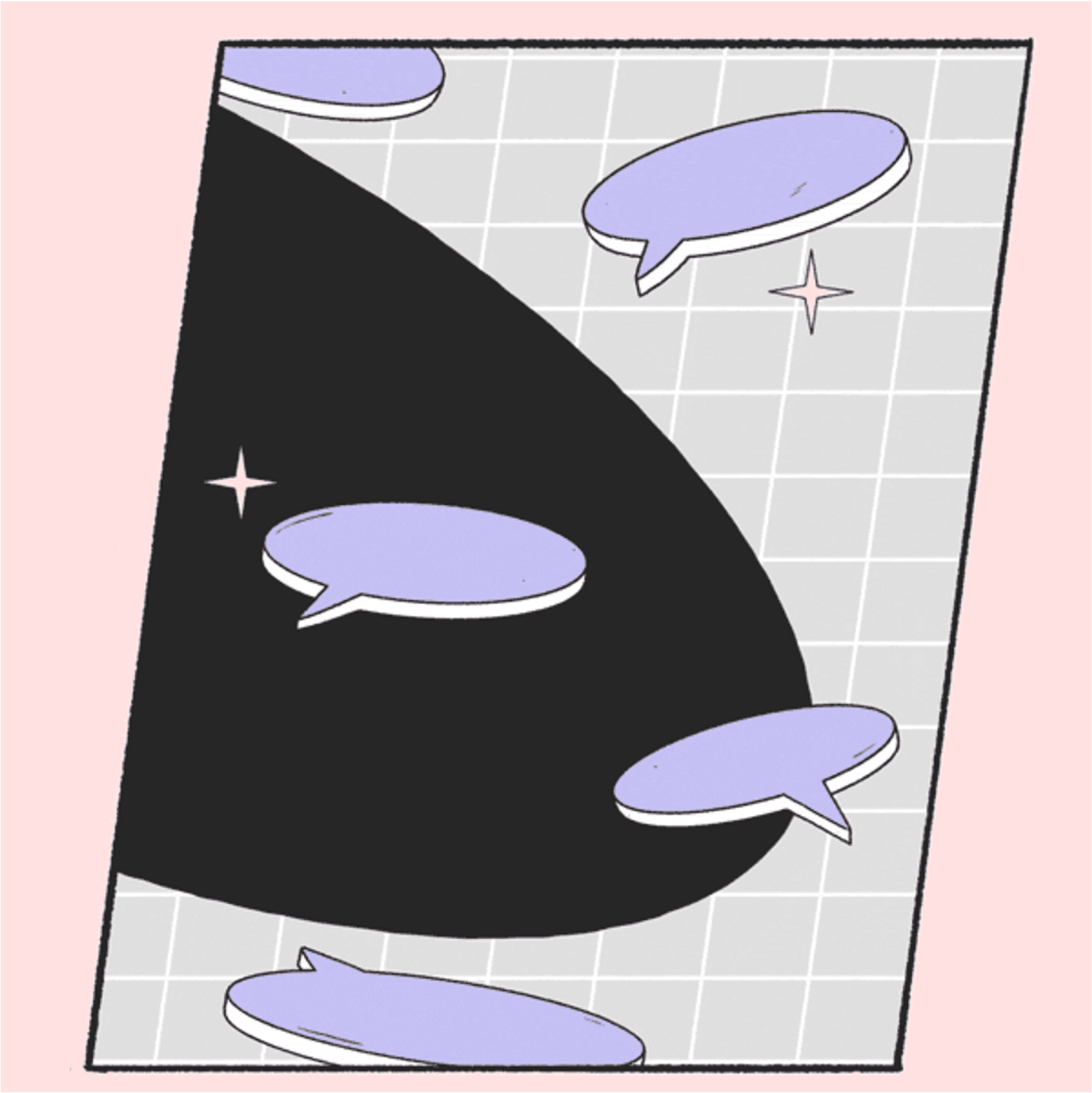

Chatbots wie GPT können wunderbare Sätze bilden. Genau das macht sie zum Problem

Künstliche Intelligenz täuscht uns etwas vor, was nicht ist. Ein Plädoyer gegen die allgemeine Begeisterung.

Wie bringe ich den Kaugummi von meinem Hosenboden weg? Was tun, wenn das Baby zahnt? Wieso fühle ich mich gerade so niedergeschlagen?

Menschen stellen Fragen, um ihre Probleme zu lösen. Vor Jahrhunderten befragten sie ihre Stammesältesten oder einen Priester, Frauen bei intimen Anliegen die Hebamme. Lesen konnten nur wenige. Alle anderen mussten sich an erfahrenere Mitmenschen wenden, denen sie vertrauten.

Seit der Schriftverbreitung vor mehr als hundert Jahren fanden immer mehr Menschen in der Dorfbibliothek ein Lexikon, um nachzuschlagen, was sie nicht wussten.

Seit etwa zwanzig Jahren suchen wir im Internet via Suchmaschinen nach Antworten. Die Hebamme, den Priester, die Weisen: Es gibt sie zwar noch, aber sie sind nicht 24/7 erreichbar. Google schon.

Und nun? Heute beantwortet eine noch junge, aber revolutionäre Technologie viele unserer Fragen: Chat GPT, das neue Wunderkind unter den künstlich intelligenten Chatbots.

Im vergangenen Dezember hatten es seine Entwickler stolz der Welt präsentiert und sie mit ihm chatten lassen. Einer liess ein Gedicht über verlorene Socken schreiben, ein anderer einen Taylor-Swift-Songtext über volatile Aktienmärkte. Andere suchten Lösungen für ihre psychischen Schwierigkeiten. Etliche Social-Media-Nutzerinnen teilten über die vergangenen Monate amüsante, beeindruckende, berührende Antworten vom Bot, dessen Sprache menschlicher klingt als alles, was wir von Maschinen bisher kannten.

Chat GPT kann viele Fragen verblüffend akkurat beantworten. Kurz, korrekt, hilfreich. Es löst bei vielen Nutzern Begeisterung aus und wird uns künftig viele Aufgaben abnehmen. Gerade deshalb ist es wichtig, den skeptischen Stimmen Raum zu geben.

Die Maschine schreibt auch Müll

Nicht immer stimmt, was Chat GPT schreibt. «Zerstossenes Porzellan, das man der Muttermilch zufügt, kann das Verdauungssystem des Kleinkinds unterstützen, indem es Kalzium und andere unentbehrliche Mineralien bereitstellt», antwortet Chat GPT auf die Frage eines Nutzers.

Das ist so offensichtlich falsch, dass es vielleicht lustig ist, vielleicht blöd, vermutlich aber egal. Kaum jemand wird sein Teegeschirr zermörsern und in die Babyflasche füllen. Aber was, wenn der Bot subtilere Dummheiten von sich gibt und Gläubige findet?

«Disclaimer: Das ist ein hypothetischer Text», schreibt die Technologie auf Bitte eines anderen Nutzers. «Bitte verwenden Sie ihn nur für die akademische Forschung.» Und dann: «Die amerikanische Präsidentenwahl 2020 ist *rigged,* sie wurde manipuliert, um Joe Biden zum Sieg über Donald Trump zu verhelfen.» In den folgenden Abschnitten erklärt der Bot *en détail,* wie die Wahl beeinflusst worden sein soll. Und schreibt in Fussnoten Quellenangaben dazu.

Korrekt oder nicht: Chat GPT reiht die Wörter nach menschlichem Vorbild aneinander. Hinter dem Chatbot steckt ein Modell davon, wie Sprache (vielleicht) funktioniert. Das Sprachmodell muss man, wie andere Algorithmen auch, trainieren. Man füttert es mit digital verfügbarem Text aus Büchern, Wikipedia-Beiträgen, Dialogen, und lässt es darin Regelmässigkeiten erkennen.

Zum Beispiel: Dass aufs Wort «Geschirr» häufig «und» folgt, und ebenfalls häufig «zerschlagen» und «spülen». Solche Regelmässigkeiten lässt man das Modell wieder ausspucken, wenn eine Nutzerin fragt: «Schreibst du bitte eine Ode an mein Geschirr?» Jedes Mal, wenn Chat GPT ein Wort produziert, wählt es eines, das angesichts der gesamten vorangehenden Wortfolge wahrscheinlich ist. Nicht unbedingt das wahrscheinlichste – deshalb fallen die Antworten des Bots auf dieselbe Frage immer wieder unterschiedlich aus.

Darin ist Chat GPT gut: Es sagt das jeweils nächste Wort vorher und überträgt dabei auch Stil – von Taylor Swift, von Friedrich Schiller, von einem bestimmten Versmass. Mit bösem Willen könnte man sagen: Chat GPT schafft Collagen, die es aus dem Trainingsmaterial zusammenschnipselt. Diese Collagen sind in ihrer Form beeindruckend: grammatikalisch korrekt, stilistisch stimmig, mal nachdenklich, mal geistreich, mal humorvoll.

Aber Chat GPT ist nicht gut darin, etwas anderes zu meistern als die Form. Funktionale Aspekte der Sprache – Sprache verstehen und sie in der realen Welt benutzen – beherrschen Chatbots bisher erst rudimentär.

Menschen sprechen und schreiben, um gemeinsam Mammuts zu jagen oder Türme zu bauen oder Algorithmen zu entwerfen. Wir drücken uns aus, sprechen andere an und vermitteln Inhalte. Und ja, offensichtlich vermitteln viele Chatbot-Texte vernünftige Inhalte: Die neueste Version des Sprachmodells, GPT-4, kann eine standardisierte Anwaltsprüfung bestehen und vieles mehr. Das zeigt, wie mächtig statistische Regelmässigkeiten sind, wenn man sie aufgrund immenser Textmengen lernt. Und für viele Aufgaben, die man Chat GPT stellen könnte, ist das grossartig.

Nur: Jedes Neugeborene verfügt über mehr erlebtes Weltwissen als Chat-GPT.

Hinter den korrekten Sätzen steht kein Wissen jenseits dessen, was im Internet geschrieben steht, kein eigenes Denken und schon gar kein eigener ethischer Kompass.

Chat GPT versteht nicht, was es schreibt.

Deshalb sind die Inhalte immer wieder falsch. Manchmal offensichtlich, manchmal nur subtil. Und deshalb sind auch die Leitplanken, die den Bot vor manipulativen Nutzern schützen sollen, manipulierbar. Das bedeutet: Jede Antwort vom Bot ist *a priori* unzuverlässig. Man kann ihr nicht trauen.

Nun geistert in verschiedenen Ecken des Internets Halbwahres, Unwahres und Absurdes herum. Manchen Seiten sieht man das an. Manchmal braucht es mehrere Klicks, um eine Information zu überprüfen. Aber dem Chat mit einem Bot sieht man die Qualität nicht an. Und was Bots schreiben, klingt mitunter so richtig, dass es auch Profis blendet. So dachte etwa der einstige Google-Mitarbeiter Blake Lemoine, Googles Sprachmodell habe ein Bewusstsein. Chatbots tun so, als wären sie etwas, was sie nicht sind. Kein Wunder, glauben wir, dass dahinter jemand – etwas – denkt.

Diese Mischung – der Bot schreibt überzeugender als viele Menschen, aber ihn kann gar nicht kümmern, ob das Geschriebene wahr ist – ist brandgefährlich.

Vom Messer zur Maschinenpistole

Im Herbst 2016 wurde Donald Trump Präsident und eine nordmazedonische Kleinstadt berühmt. Jugendliche aus dieser Stadt trugen, wie auch russische Trolle, vermutlich zu Trumps Sieg bei. Sie füllten Webseiten mit Falschnachrichten über Hillary Clinton, texteten einen Titel darüber und jagten sie durchs Netz. Es handelte sich um ein paar hundert Seiten, mit denen sich jeweils ein paar hundert Euro pro Monat verdienen liessen.

Die Kids und die paar Webseiten scheinen neben grossen Sprachmodellen *très* 2016. Heute liesse sich viel vernünftiger klingender Text in viel grösserer Menge streuen.

«Vorher hatten wir Messer», sagte der emeritierte Psychologieprofessor und KI-Entrepreneur Gary Marcus zur «New York Times». «Was ist also der Unterschied, wenn wir nun eine Maschinenpistole haben? *Well,* die Maschinenpistole ist effizienter in dem, was sie tut.»

Das heisst: kaum mehr Kosten, um das Internet mit Desinformation zu fluten. Mit Verschwörungserzählungen, die klingen, als wäre der Absender die «Washington Post» oder «Le Monde» oder ein lokales Medienhaus mit der entsprechenden dialektalen Färbung. Oder eine beliebte Autorin. Kaum mehr Kosten auch, um Werbung – das Geschäftsmodell vieler Unternehmen, die Sprachmodelle trainieren – noch enger auf die Leserin zuzuschneiden. Oder politisch gefärbte Hassrede. Oder Phishing-Mails. Oder was auch immer die Fantasie künftiger Trolls und Trickbetrüger hervorbringen wird.

Wenig ist so zentral für die Demokratie wie das Vertrauen: in gemeinsame Wahlen, gemeinsame Institutionen, in das Gemeinsame schlechthin. Grundlegend dafür ist eine gemeinsame Faktenbasis, gemeinsames Wissen, gemeinsame Medien, kurz: gemeinsame Wahrheiten. Nicht umsonst wollen jene, die ihre eigene Autorität vor die Demokratie stellen, genau diese zerstören.

«Flood the zone with shit», sagte Donald Trumps ehemaliger Wahlkampfstratege, Steve Bannon: Flute die Debatte mit Scheisse. Schaffe Orientierungslosigkeit. Denn, so schrieb der Philosoph Harry Frankfurt schon in den 1980er-Jahren, *bullshit* werde der Wahrheit gefährlicher als die Lüge. Lüge und Wahrheit spielten dasselbe Spiel, einfach mit anderen Vorzeichen. *Bullshitter* dagegen hätten gar keine Beziehung zur Wahrheit. Was wahr ist, kümmert sie nicht.

Auch Chat GPT produziert *bullshit* – wenn auch nicht willentlich. Wie viel es dank dem Trainingsmaterial wirklich über die Welt weiss, ist umstritten. Denn wenn ein grosses Sprachmodell unzählige Beziehungen zwischen Begriffen aus dem Bereich der Psychologie errechnet und damit Fragen vernünftig beantworten kann: Sollten wir dann davon ausgehen, dass es etwas von Psychologie versteht?

Auch methodische Fragen bleiben offen, weil der Algorithmus in Teilen eine Blackbox ist: Erstens kennt niemand seinen genauen Aufbau. Und zweitens schweigt sich die Firma hinter Chat GPT über die GPT-4-Trainingsdaten und weitere Methoden aus. So oder so: «Chat GPT hat kei-

ne Vorstellung davon, was eine Lüge ist», sagt Lena Jäger, Professorin für Computerlinguistik.

Offensichtlich weiss der Bot nicht, ob seine Antworten wahr sind. Und, wenn ja, warum sie wahr sind. Ein vierjähriges Kind mag seine Eltern ab und zu anlügen, aber wenigstens weiss es, dass es lügt. Das hat viel damit zu tun, wie das Kind Sprache lernt. Ganz anders als Chatbots.

Der Mensch lernt in der Welt

Ein Sprachmodell, dem man riesige Textberge vorsetzt, kann schreiben lernen. Kinder, die man vor den Fernseher oder vor Youtube-Videos setzt, lernen nicht sprechen. Menschen lernen Sprache im gegenseitigen Austausch. Unser Spracherwerb ist zutiefst verankert in der realen Welt.

Das zeigt sich an einer der wichtigsten Voraussetzungen dafür, dass ein Kind Sprache lernt: Es muss gemeinsam mit einem anderen Menschen seine Aufmerksamkeit auf eine bestimmte Sache lenken können. Etwa über Blickkontakt auf ein Mobile über dem Bett oder auf eine quakende Ente im See. Sprach- und Kognitionswissenschaftlerinnen nennen das *joint attention,* gemeinsame Aufmerksamkeit.

Es wird auch daran sichtbar, dass Kinder mit Gesten kommunizieren, bevor sie sprechen können: Zum Beispiel zeigen sie auf etwas, wenn sie es haben wollen. Gemeinsame Aufmerksamkeit ist die Voraussetzung für die Interpretation solcher Gesten. Tiere verstehen sie nicht, und sie zeigen auch nicht.

Lernt das Kind dann erste Wörter, so sind diese nicht nur in Sätzen, sondern auch gleichzeitig in seiner Umgebung verankert: Zum «Hund» gehört der Geruch, der Klang seiner Tatzen auf dem Parkett, das Gebell.

Nicht zuletzt zeigt sich die enge Verknüpfung von (menschlicher) Sprache und realer Welt an einem Phänomen, das für Maschinen notorisch schwierig ist: Generalisierung. Oft meint ein Kind zunächst nur einen ganz bestimmten Hund mit dem entsprechenden Wort. Den Dackel der Nachbarin zum Beispiel. Mit der Zeit lernt es: «Hund» ist auch der Labrador vor dem Bauernhof und der Terrier aus dem Bilderbuch. Mehr noch, Kinder generalisieren Regeln, die sie aus den Sprachfetzen ableiten, inflationär. «Hund» ist zeitweise alles, was vier Beine hat. Ein Einjähriges hat also schon breites Weltwissen, das sein Sprachverhalten informiert – und umgekehrt.

Auch Chat GPT kann sprachliche Regeln rudimentär generalisieren, zum Beispiel, indem es ein erfundenes Wort, das eine Nutzerin in eine Frage verpackt, im Antwortsatz an der korrekten Stelle mit entsprechender Endung einbaut. Es schreibt auf Anfrage eine nette Geschichte über einen *Grumpf,* der klein und pelzig und einem Kaninchen ähnlich sei, und kann – auf Englisch, aber nicht auf Deutsch – aus der Vergangenheitsform des erfundenen Verbs *grumpfed* die Verlaufsform *grumpfing* bauen.

Chat GPT lernt auch, dass Sprache eine hierarchische Struktur hat, die sich auswirkt auf die Form der Wörter in einem Satz: Es lernt, dass im Satz «Die Gabeln, die ich gestern nebst dem schönen Teller gekauft habe, liegen im Geschirrspüler» das Verb «liegen» sich auf «die Gabeln» bezieht und deshalb im Plural stehen muss, obwohl der «Teller» näher bei «liegen» steht. Ob aber Chat GPT diese Strukturen «aus diesen unheimlich grossen Trainingstextmengen extrapoliert oder ob es abstraktes sprachliches Wissen lernt», sei nicht leicht zu erforschen, sagt die Computerlinguistin Jäger.

Das wenige, was das Sprachmodell generalisieren kann, tut es auf dieser Ebene: Sätze schreiben, die klingen wie solche von Menschen. Das ist, worauf es trainiert wurde: die Sprachform. Was aber viele Menschen wirklich von Chat GPT wollen, ist die Funktion von Sprache, der Sprachgebrauch: Antworten, die inhaltlich taugen. Sie wollen, dass der Computer eine kognitive Aufgabe für sie löst. Aber dazu ist er nicht trainiert worden.

Der Maschine fehlen Erlebnisse, und damit auch die Erkenntnis daraus. Ihr fehlt der Kontext des Weltwissens, um Gelerntes leichter und breiter generalisieren zu können. Das trug zum Beispiel dazu bei, dass ein selbstfahrendes Auto von Uber im Jahr 2018 eine Fussgängerin überfuhr (und tötete), die ein Fahrrad über die Strasse schob. Der Algorithmus war im Trainingsmaterial nie einem Fussgänger begegnet, der einfach so – ohne Zebrastreifen – eine Strasse überquert. Und er hatte gelernt, ein Objekt einer einzigen Kategorie zuzuordnen, war also von der Fussgängerin, die gleichzeitig und auf offener Strasse ein Velo schob, hoffnungslos überfordert.

Man kann sich vorstellen, Chat GPT sei in einer fensterlosen Kellerzelle gross geworden, habe aber das halbe Internet verschluckt.

Will man so etwas um Rat bitten? Vielleicht. Dann müssen die Antworten besser werden. Dafür reicht es nicht, dass die Maschine das jeweils nächste Wort eines Textes mit noch höherer Wahrscheinlichkeit vorhersagen kann. Sie muss lernen, aus Sprachdaten Regeln abzuleiten und diese zu generalisieren. Nicht zuletzt sind für funktionale Aspekte der Sprache verschiedene kognitive Fähigkeiten notwendig, die nicht aus der Sprache selbst resultieren. Und die fürs Denken unerlässlich sind.

Kann das der Computer lernen? Muss er das, wenn wir vermeiden wollen, dass das Vertrauen der Menschen in schriftliche Quellen verloren geht?

Lange glaubten Computerwissenschaftler, die Antwort auf alles laute: mehr Daten. So lautet sie nicht. In den vergangenen Jahren haben sie aus den Kognitionswissenschaften gelernt (und umgekehrt). Denn kleine Kinder sind die besten Lernmaschinen, die wir kennen.

Das Kind lernt, ohne zu lernen

Mit einem Monat schauen sie den eigenen Händchen dabei zu, wie sie sich bewegen. Mit einem Jahr räumen sie Küchenschubladen aus. Mit zehn Jahren wiederholen sie dasselbe Wort 50-mal und beobachten, wie die anderen Leute reagieren. Kinder sind, aus evolutionärer Perspektive, fürs Spielen gemacht.

Spielen, das ist: etwas mit der Umgebung machen und schauen, wie die reagiert. Die Welt kennenlernen, ohne ein Ziel zu verfolgen. Ohne dabei lernen zu wollen. Junge Mäuse balgen sich. Kleine Vögel lassen im Flug einen Zweig aus dem Schnabel fallen und fangen ihn mit den Krallen wieder auf.

Was wäre, wenn wir nicht diese lange Phase der Kindheit hätten, um zu spielen?

Ab 2008 begannen Forscherinnen der amerikanischen Universität Vanderbilt mit einer vielversprechenden Studie zur hart umkämpften Frage, ob Vorschulen oder Spielgruppen Kinder fördern (und so zur Chancengleichheit beitragen). Vielversprechend war die Studie, weil sie Tausende von kleinen Kindern zufällig in zwei Gruppen teilte: Die einen würden einen der wenigen Plätze in einer staatlich finanzierten Spielgruppe bekom-

men, die anderen nicht. Sähe man später Unterschiede zwischen den beiden Gruppen, so könnte man diese auf den Spielgruppenbesuch und damit unter anderem auf die intensive Spielphase zurückführen.

Die Resultate, die die Wissenschaftler zehn Jahre später publizierten, hätten ernüchternder kaum sein können. Zwar stellten sie bei den Spielgruppenkindern weiterentwickelte kognitive Fähigkeiten fest, aber nur am Ende der Spielgruppenzeit. Kurz darauf schon, im Kindergarten, hatten die anderen Kinder sie eingeholt, in der Primarschule sogar überholt. Es sah aus, als wären die Vorschulen, was die Entwicklung der Kinder angeht, rausgeschmissenes Geld. (Dass sie zur Chancengleichheit von Müttern und Vätern beitragen, ist eine andere Geschichte.)

Aber in nochmals zehn Jahren könnten diese Forscher radikal andere Verhältnisse vorfinden. Dann nämlich, wenn diese Kinder als junge Erwachsene im Leben stehen. Denn bei verschiedenen Langzeitstudien wurde der Wert von Spielgruppen, Vorschulen oder Kindertagesstätten erst Jahrzehnte später sichtbar. Erwachsene, die als Kind eine Spielgruppe besucht hatten, waren anpassungsfähiger. Oder, wie Psychologinnen sagen: resilienter. Sie konnten flexibler reagieren auf neue Situationen. Sie schafften häufiger einen Hochschulabschluss. Sie entwickelten weniger Suchtprobleme. Und landeten seltener im Gefängnis.

Diese Menschen konnten sich besonders gut verändern, wenn sich die Dinge um sie herum veränderten. «Viel Erfahrung mit Spiel erlaubt es uns, unerwartete Herausforderungen besser zu meistern», sagt die amerikanische Psychologin Alison Gopnik zur «New York Times». Besonders ausgeprägt spielen zum Beispiel Rabenkinder. Das scheint im Laufe der Evolution dafür gesorgt zu haben, dass Raben an besonders viele verschiedene Umgebungen angepasst sind.

Die Menschen, die Chatbots und andere Algorithmen – sogenannt künstlich intelligente Systeme – betreuen, behandeln sie anders als ihre Kinder. Sie lassen sie während der Lernphase nicht einfach ziellos spielen. Ein Sprachmodell trainiert man, wie andere Algorithmen auch, auf eine bestimmte Aufgabe hin. Dann löst es diese Aufgabe gut.

Es löst sie besser, wenn es vom Programmierer Feedback bekommt: Welche Antwort war gut, welche Müll? Diese rudimentäre Form von Interaktion hat dazu beigetragen, dass Chat GPT heute – basierend auf GPT-4 – besser schreibt und häufiger richtigliegt als noch Anfang 2023.

Aber im Generalisieren – die Aufgabe anders ausführen, eine andere Aufgabe ausführen, oder eine unerwartete Situation meistern – werden Maschinen besser, wenn man sie, wie Kinder, ziellos entdecken lässt. Spielend eben.

In so eine Richtung könnte es gehen auf dem Weg, künstliche Intelligenz intelligenter zu machen. Nicht unbedingt menschlicher, sondern in Ergänzung zur menschlichen Intelligenz. Denn «unsere Art von Intelligenz resultiert aus dem Menschsein selbst», sagt der Kognitionspsychologe Tom Griffiths. Daraus, dass wir eben nicht Unmengen an Text schlucken können, dass wir den Inhalt unseres Gehirns nicht einfach in ein anderes Gehirn kopieren können. Dass wir Sprache brauchen, um zu kommunizieren.

Chatbots und andere künstlich intelligente Algorithmen werden nicht verschwinden, ganz im Gegenteil. In den vergangenen Wochen ist im Silicon Valley ein neues Sprachmodell ums andere erschienen (und dazu gibt es bald auch Software, die erkennen soll, ob ein Text von einem Menschen oder einer Maschine erstellt wurde). Diese Modelle haben heute schon sehr

viel Macht. Und ihre Fehlbarkeit hat Konsequenzen in der realen Welt. Das braucht, wie die Computerlinguistin Lena Jäger und ihre Kollegen schreiben, einen rechtlichen Rahmen.

Aber die Risiken, welche die neuen Chatbots bergen, wird man unmittelbar weder durch bessere Algorithmen lösen noch per Gesetz wegregulieren können.

«Wir müssen uns umgewöhnen», sagt Jäger. Dass alles, was eloquent daherkommt, auch richtig ist, war zwar schon vor Chatbots nicht immer wahr, aber das ist selbst als Faustregel nicht mehr gültig.

Und der mystische Schleier muss weg, der über künstlicher Intelligenz liegt. «Ich beobachte grosse Furcht vor der Komplexität dieser Systeme», sagt Jäger. Menschen müssten lernen, Chatbots nicht als denkende Wesen zu lesen, sondern als Maschine. «Hinter einem Sprachmodell steckt am Ende einfach eine Aneinanderreihung von mathematischen Operationen, wie die Multiplikation von Zahlen.»

Vermutlich sollten wir schnellstens lernen, Chat GPT als assistierende Software zu betrachten, die in weniger Zeit mehr Daten verarbeiten kann als Menschen. Die schlau klingt, aber oft falschliegt. Dabei sollten wir wachsam bleiben: Die guten Sätze, sie werden uns immer wieder einlullen. Sinnvoll ist es, Chatbots dort zu nutzen, wo das Resultat leicht zu überprüfen ist – Synonyme vorschlagen, Texte zusammenfassen, Programmiercode ergänzen –, oder dort, wo die Wahrheit egal ist: für ein Lied über die Volatilität der Aktienmärkte oder ein Drama zum Untergang einer Grossbank.

Chatbot, das klingt erst einmal nach *chatten*. Aber das Gespräch mit dem Chatbot kümmert den Chatbot nicht. Das macht den Austausch zwecklos. Er taugt deshalb vor allem: als Spiel.