Erscheint in: Merkur, November 2022. Unkorrigierte Version

Dumme Bedeutung
Künstliche Intelligenz und artifizielle Semantik

Hannes Bajohr

Im Juni 2022 wurde der Google-Ingenieur Blake Lemoine von seinem Arbeitgeber auf unbestimmte Zeit freigestellt, weil er meinte, die Künstliche Intelligenz, an deren Testphase er mitarbeitete, verfüge über Bewusstsein.[1] Das Chatbot-System LaMDA *(Language Model for Dialogue Applications)* habe ihn in langen Gesprächen davon überzeugt, dass es die Intelligenz eines hochbegabten Achtjährigen besitze, und darum gebeten, als Person mit Rechten betrachtet zu werden.[2] Lemoine, der sich selbst als „mystischen Christen und ordinierten Priester" bezeichnet, überspitzte dabei nur eine Stimmung, die auch andere Google-Mitarbeiter befiel. Blaise Agüera y Arcas, leitender Ingenieur für Machine Learning und für gewöhnlich frei von allem Mystikverdacht, schrieb nur wenige Tage vor Lemoine über seine Interaktionen mit LaMDA: „Ich hatte das Gefühl, der Boden unter meinen Füßen würde zu schwanken beginnen. Mein Eindruck war immer mehr, mit etwas zu sprechen, das Intelligenz besitzt."[3]

Ganz ohne die Schlagwörter von Bewusstsein und Intelligenz kam dagegen eine etwa zeitgleich verlaufende Diskussion um ein anderes KI-System aus. Dall·E 2, das von der Firma OpenAI veröffentlicht wurde, ist eine *text-to-image AI* und kann aus natürlichsprachigen Eingaben Bilder generieren. Gibt man ihm etwa den Prompt: „Ein Shiba-Inu, der ein Beret und einen schwarzen Rollkragenpullover trägt", zeigt das Ausgabebild anschließend eben diese Szene.[4] Die öffentliche Beta-Version, bei der Schritt für Schritt für immer mehr User teilnehmen konnten, löste ein wildes Experimentieren aus, und bald wurden die interessantesten oder

[1] Nitasha Tiku, *The Google engineer who thinks the company's AI has come to life.* In: *Washington Post* vom 11. Juni 2022 (www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine). Lemoines Begriff ist *sentience*, nicht *consciousness*, aber er scheint sie synonym zu verwenden.
[2] Blake Lemoine, *What is LaMDA and What Does it Want?* In: *Medium* vom 11. Juni 2022 (cajundiscordian.medium.com/what-is-lamda-and-what-does-it-want-688632134489). Vgl. auch das Chatprotokoll von Lemoines Gesprächs mit LaMBDA: *Is LaMDA Sentient? An Interview*. In: *Medium* vom 11. Juni 2020 (cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917).
[3] Blaise Agüera y Arcas, *Artificial neural networks are making strides towards consciousness*. In: *Economist* vom 9. Juni 2022 (www.economist.com/by-invitation/2022/06/09/artificial-neural-networks-are-making-strides-towards-consciousness-according-to-blaise-aguera-y-arcas).
[4] Aditya Ramesh u.a., *Hierarchical Text-Conditional Image Generation with CLIP Latents*. In: *arXiv* vom 13. April 2022 (arxiv.org/abs/2204.06125); OpenAI, *Dall·E 2* vom 6. April 2022 (openai.com/Dall-E-2).

wunderlichsten Ergebnisse im Netz und vor allem auf Twitter geteilt.

Auch das war aufschlussreich und bestätigte angesichts der bislang nur wenig erfolgreichen Versuche mit autonomen Autos unter anderem den Eindruck, dass KI deutlich andere gesellschaftliche Auswirkungen haben dürfte als lange gedacht – etwa indem sie vor Lasterfahrern eher Illustratoren, Grafikern und Symbolbildagenturen ihren Job streitig machen könnte.[5] Doch anders als bei LaMDA kam kaum jemand auf die Idee, Dall·E 2 für eine Person mit Rechten zu halten.

Die unterschiedlichen Reaktionen auf die zwei Systeme zeigen, wie schnell das Denken über KI in die gewohnten konzeptuellen Spurrinnen einschert. Intelligenz, Bewusstsein, Personalität sind seit knapp siebzig Jahren die großen Themen der KI-Forschung und ihrer Imaginationen; amüsante Bildchen hingegen scheinen weniger grundsätzliche Fragen aufzuwerfen. Es ist aber gut möglich, dass es sich genau andersherum verhält, dass die ewige Jagd nach Superintelligenzen und der Singularität die interessanteren und feineren Begriffsverschiebungen verdeckt, die sowohl den Tech-Evangelisten in ihrem visionären Furor als auch deren Kritikern entgehen im Versuch, skeptisch dagegenzuhalten.

Für den Philosophen Benjamin Bratton steht fest, dass angesichts neuer KI-Systeme „die Wirklichkeit die zu ihrer Beschreibung verfügbare Sprache überholt hat". Es brauche daher „ein präziseres Vokabular" jenseits der genannten Großbegriffe,[6] aber auch jenseits der anthropozentrischen Annahme, die einzige Art, wie Maschinen Formen der Welterschließung ausbilden können, wäre die unsere. Betrachtet man Dall·E 2 und LaMDA, zeigt sich eine solche Tendenz, die den Begriff der Bedeutung von seinem anthropozentrischen Korrelat löst; sie wäre Bedeutung ohne Geist – dumme Bedeutung.

Bodenlose und geerdete Systeme

Mit ihren steten Warnungen, Begriffe wie „Intelligenz" und „Bewusstsein" nicht leichtfertig zu verwenden, laufen Informatik, Linguistik und Kognitive Psychologie gerade in der Tech-Branche traditionell ins Leere. Auch Lemoine wurde bald vorgeworfen, dem ELIZA-Effekt aufgesessen zu sein,[7] also Intelligenz und Bewusstsein auf LaMDA projiziert zu haben, wie es Joseph

[5] Prarthana Prakash, *AI Art Software Dall-E Moves Past Novelty Stage and Turns Pro*. In: *Bloomberg* vom 4. August 2022 (www.bloomberg.com/news/articles/2022-08-04/dall-e-art-generator-begins-new-stage-in-ai-development); die Ausgabe des *Economist* vom 11. Juni 2022 zierte eine von einer Bild-KI generierte Illustration.
[6] Benjamin Bratton/Blaise Agüera y Arcas, *The Model Is The Message*. In: *NOĒMA* vom 12. Juli 2022 (www.noemamag.com/the-model-is-the-message).
[7] Brian Christian, *How a Google Employee Fell for the Eliza Effect*. In: *The Atlantic* vom 21. Juni 2022 (www.theatlantic.com/ideas/archive/2022/06/google-lamda-chatbot-sentient-ai/661322).

Weizenbaum bereits 1966 bei den Usern seines Chatbots ELIZA beobachtet hatte: Obwohl ELIZA lediglich einen Psychoanalytiker à la Carl Rogers mimte – dessen Verfahren darin besteht, dem Patienten seine Aussagen zu spiegeln –, benahmen sich seine Benutzer so, als sei das Programm wirklich ein bewusster, an ihrem Befinden interessierter Agent.

Der klassische Einwand lautet hier: Computer sind symbolverarbeitende Systeme, die allein mit Syntax, nicht mit Semantik umgehen, also lediglich logische Formen, aber keine inhaltlichen Bedeutung prozessieren können.[8] Für ihre Operationen ist irrelevant, welche Objekte oder Begriffe die Symbole in einer menschlichen Welt benennen und welche kulturellen Wertigkeiten mit ihnen verbunden sind. So scannt ELIZA die User-Eingabe lediglich auf ein vorgegebenes syntaktisches Muster und wandelt sie nach einer Transformationsregel in eine „Antwort" in Frageform um. Weizenbaum gibt das Beispiel, in dem der Analysand dem Analytiker vorhält: „It seems that you hate me." Das Programm identifiziert in diesem Satz das Schlüsselmuster „*x* you *y* me" und separiert ihn entsprechend in die vier Elemente „It seems that", „you", „hate" und „me". Anschließend verwirft es *y* („It seems that") und setzt *x* („hate") in die Replikschablone „What makes you think I *x* you?" ein. So antwortet ELIZA auf den Vorwurf, es hasse den Analysanden, mit der Frage, wie er denn auf diese Idee komme.[9]

Diese Interaktion mag für den User Bedeutung haben und eine kommunikative Absicht auf Seiten ELIZAs nahelegen, doch sind weder Absicht noch Bedeutung im Programm zu finden. Es hat lediglich Symbole nach einer Regel verarbeitet, ohne zu „wissen", was Hass ist oder welches Verhalten die Sitten des Dialogs nahelegen; das ist der Unterschied zwischen Informationsverarbeitung und Bedeutungsverstehen.

Für eine KI-Forschung, die Computer Menschen angleichen möchte, beschreibt dieser Umstand, was der Kognitionspsychologe Stevan Harnad 1990 das „symbol grounding problem" genannt hat: Symbole, wie die in Weizenbaums Transformationsoperation, haben im Computer keine *intrinsische* Bedeutung, weil sie ohne den Hintergrund eines praktischen Weltwissens nur auf andere Symbole, nie aber auf die Welt verweisen können. Aus diesem „Symbolkarussell" gibt es keinen Ausweg; alle Bedeutung kann hier nur „parasitär" sein, wird von menschlichen Interpreten an das System herangetragen.[10]

Harnads Kritik richtete sich allerdings gegen nur ein bestimmtes Prinzip von KI, unter das

[8] Florian Cramer, *Language*. In: Matthew Fuller (Hrsg.), *Software Studies. A Lexicon*. Cambridge/Mass.: MIT Press 2008.
[9] Joseph Weizenbaum, *ELIZA – A Computer Program for the Study of Natural Language Communication Between Man And Machine*. In: *Communications of the ACM*, Nr. 9/1, 1966. Strenggenommen erlaubt ELIZA ganz verschiedene Transformationsregeln; der „Therapeut" ist nur ein Unterprogramm namens DOCTOR.
[10] Stevan Harnad, *The Symbol Grounding Problem*. In: *Physica D*, Nr. 42/1-3, 1990.

auch ELIZA fällt und das aus naheliegenden Gründen „symbolisch“ genannt wird. Um das „symbol grounding problem“ zu lösen, setzte Harnad auf die seinerzeit neuartigen „*sub*symbolischen“ oder „konnektionistischen“ Systeme – neuronale Netze, von denen LaMDA und Dall·E 2 späte Nachfahren sind.

Anders als die traditionelle KI sind diese Systeme nicht als Satz logischer Schlussregeln angelegt, sondern vage dem Hirn nachempfunden, als komplexe Verbindungen zwischen signalverstärkenden oder -abschwächenden Neuronen. Sie kommen daher ohne explizite symbolische Repräsentationen und Regeln aus, werden nicht programmiert, sondern lernen anhand von Beispielen selbständig. Da sie damals vor allem zur Mustererkennung verwendet wurden, sah Harnad in ihnen einen möglichen Zugang zur Welt: In einen autonomen, mobilen Roboter implementiert, ausgestattet mit Sensoren und Effektoren, sollte zunächst ein Verbund neuronaler Netze Eindrücke empfangen und als wiedererkennbare Gestalten kategorisieren. Diese würden anschließend einer symbolischen KI übergeben, wären nun aber nicht mehr bloße Verweise auf andere Symbole, sondern über ihren kausalen Bezug auf externe Daten mit der Welt verbunden – sie wären endlich „geerdet“.[11]

Die Konsequenz aus diesem Gedanken scheint aber zu sein: Die einzige Art, den ELIZA-Effekt zu umgehen, der Computern fälschlich Bewusstsein zuschreibt, ist, ihnen *wirklich* Bewusstsein zu geben. Denn was Harnad vorschwebt, ist am Ende wieder ein anthropozentrisches Modell, das davon ausgeht, *embodied cognition* und genügend umfangreiche referenzielle Bedeutungen würden Weltverstehen hervorbringen, weil auch wir in etwa so funktionieren. Der Erfolg seines Hybridmodells müsste sich sich darin erweisen, dass sein Roboter sich so kompetent in der Welt zurechtfände, als ob er tatsächlich intelligent wäre. Da das bisher noch nicht der Fall ist, kann auch das „symbol grounding problem“ noch nicht als gelöst gelten; weil alles Verstehen konstitutiv ein solches Weltverhältnis voraussetzt, kann es ein *bisschen* Bedeutung hier per definitionem nicht geben. Und doch scheinen gerade das LaMDA und Dall·E 2 nahezulegen.

Bedeutungsgrade

Mit der überwältigenden Popularität, derer sich neuronale Netze seit nunmehr fast zehn Jahren erfreuen, ist auch die Idee, ihnen sei irgendwie ein Zugang zu Bedeutung jenseits bloßer ungeerdeter Symbole gegeben, wieder attraktiver geworden. Für die Kulturwissenschaftlerin Mercedes Bunz können neuronale Netze dank ihrer Komplexität und Lernfähigkeit nun nicht mehr

[11] Stevan Harnad, *Grounding Symbols in the Analog World with Neural Nets*. In: Think, Nr. 2/1, 1993.

nur leere Symbole, sondern eben auch „Bedeutung berechnen“.[12] Richtig ist, dass sich angesichts neuronaler Netze die binäre Unterscheidung zwischen Bedeutung (menschliche Welt) und Nichtbedeutung (digitale Systeme) immer schwerer aufrechterhalten lässt. Sie erlauben Zwischenstufen von gradierter Bedeutung, die als *artifizielle Semantik* keinen Geist mehr voraussetzt.

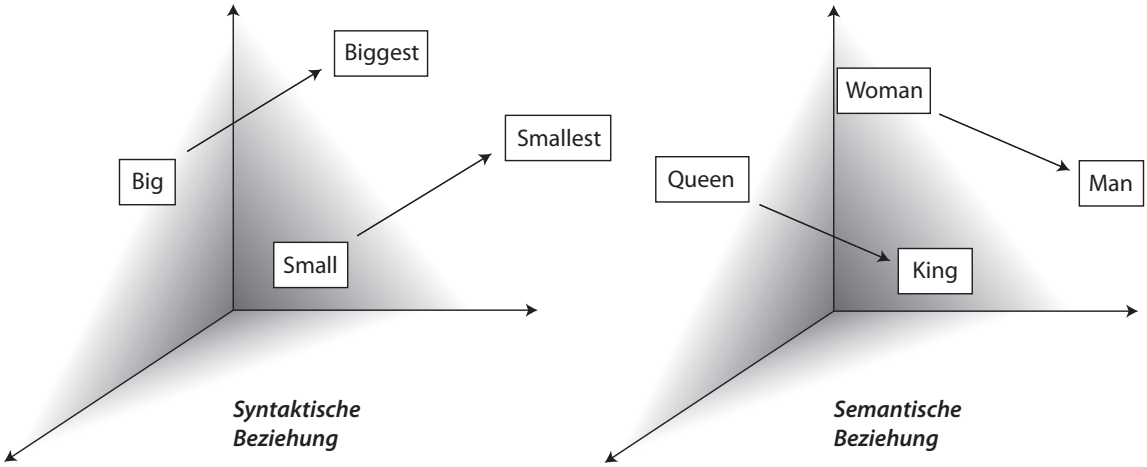
So kann man die Tatsache, dass LaMDAs Antworten so menschenähnlich klingen, statt als Anzeichen von Bewusstsein schlicht als Hinweis auf solche „dumme“ Bedeutung verstehen. Während „breite“ Bedeutung – je nach philosophischer oder disziplinärer Ausrichtung – verkörperte Intelligenz, kulturelles und soziales Vorwissen oder die welterschließende Funktion von Sprache zur Voraussetzung hat, läge dumme Bedeutung jenseits dieser (auf Menschen abzielenden) Skala, und ließe sich am besten über den Begriff der *Korrelation* fassen.[13]

LaMDA ist – ähnlich wie der bislang weit bekanntere Textgenerator GPT-3 – ein großes Sprachmodell, das als neuronales Netz implementiert ist. Trainiert auf Unmengen von Text, verarbeitet es Sprache als vieldimensionalen Vektorraum, das sogenannte *word embedding*, das nach dem Prinzip gestaffelter Korrelationen funktioniert: Zunächst liegen Wörter, die häufig gemeinsam auftauchen, in diesem Raum näher beieinander und bilden semantische Cluster, wie man sie aus Wortwolken kennt. Da im Modell aber nicht nur die Korrelationen von Wörtern zu Wörtern, sondern auch Korrelationen von Korrelationen codiert sind, können große Sprachmodelle auch implizite, im Trainingstext gar nicht formulierte Regelmäßigkeiten explizieren. Das gilt für syntaktische Beziehungen – so ist der euklidische Abstand zwischen den Vektoren für den Positiv und den Superlativ eines Wortes stets nahezu gleich –, aber auch für komplexe semantische Verhältnisse, also für Wortbedeutung. Eines der bekanntesten Beispiele für dieses Prinzip ist die Operation: $v_{king} - v_{man} + v_{woman} \approx v_{queen}$.[14]

[12] Mercedes Bunz, *The Calculation of Meaning: On the Misunderstanding of New Artificial Intelligence as Culture*. In: *Culture, Theory and Critique*, Nr. 60/2, September 2019.

[13] Dumme Bedeutung schließt *natürliche* Bedeutung aus, wie das Symptom die Krankheit bedeutet. Auch kann sie keine *intentionalistische* Bedeutung bezeichnen, wie sie Paul Grice theoretisiert, demzufolge die Bedeutung einer Äußerung vom Erkennen einer Sprecherintention abhängt, die Bewusstsein voraussetzt. Sie ist zudem nur ganz bedingt eine *Gebrauchstheorie* in der Tradition des späten Wittgenstein, weil „Gebrauch“ einen geteilten sozialen Hintergrund, dieser Weltverstehen und dieses verkörperte Intelligenz voraussetzt.

[14] Das wurde zuerst 2013 von den Erfindern des Word-embedding-Modells Word2vec entdeckt. Vgl. Tomas Mikolov/Wen-tau Yih/Geoffrey Zweig, *Linguistic Regularities in Continuous Space Word Representations*. In: *Proceedings of the 2013 Conference of the NAACL*. Atlanta: ACL 2013. Diese Einsicht gilt aber weiterhin für neuere, technisch anders aufgebaute Modelle wie etwa GloVe (Global Vectors for Word Representation).



Word embedding eines großen Sprachmodells

In dieser Gleichung – die man lese als: Zieht man vom Wortvektor „König" den für „Mann" ab und addiert den für „Frau" hinzu, erhält man als Ergebnis den Wortvektor „Königin" – kommt die latente semantische Beziehung „Geschlecht" als arithmetische Korrelation zum Vorschein, obwohl sie nicht ausdrücklich im Modell vorhanden ist. (Dass sie aus der Masse an Sprache emergiert, auf die das Modell trainiert ist, erklärt die Anfälligkeit für *biases*: Sexismus und Rassismus können ebenfalls latent in Sprachmodellen codiert sein.)[15] Die Bedeutung eines Zeichens in einem so konstruierten Sprachsystem ist – wie im linguistischen Strukturalismus Ferdinand de Saussures – rein *differenziell* bestimmt, als Verschiedenheit zu anderen Zeichen und Zeichenkorrelationen.[16] Der Effekt ist dennoch, dass große Sprachmodelle allein durch ihre immensen Trainingsdaten in der Lage sind, scheinbar situatives Verstehen zu produzieren, wie LaMDA es tat, ohne je „in einer Situation" zu sein.[17]

Sprachmodelle wären Produzenten eines ersten Grades dummer Bedeutung. Dumm ist sie, weil das Modell zwar latente Korrelationen zwischen Zeichen erfasst, aber immer noch nicht „weiß", welche Sachen diese Zeichen eigentlich benennen; mit dieser Art von Bedeutung wird man keine Intelligenz bauen können, die sich je in der Welt zurechtfindet. Die Linguistin Emily Bender – eine vehemente Kritikerin allen KI-Hypes um angebliches Bewusstsein – gibt mit ihrem Kollegen Alexander Koller zwar zu, dass „bestimmte Aspekte von Bedeutung", etwa semantische Ähnlichkeit, in Sprachmodellen niedergelegt sein können, hält sie aber für „nur

[15] Vgl. Hannes Bajohr, *Wer sind wir? Warum künstliche Intelligenz immer ideologisch ist*. In: *Republik* vom 6. April 2022 (republik.ch/2021/04/06/warum-kuenstliche-intelligenz-immer-ideologisch-ist)
[16] Sehr schön aufgearbeitet ist das bei Juan Luis Gastaldi, *Why Can Computers Understand Natural Language? The Structuralist Image of Language Behind Word Embeddings*. In: *Philosophy & Technology*, Nr. 34/4, März 2021.
[17] So Hubert Dreyfus' Begriff für das vorgängige Weltverstehen, das Menschen, aber nicht Computer haben, und das er von einer an Heidegger geschulten hermeneutischen KI-Kritik her entwickelt, vgl. ders., *Was Computer nicht können.* Frankfurt am Main: Athenäum 1989.

ein[en] schwache[n] Abglanz wirklicher Bedeutung“, die immer auf etwas in der Welt bezogen, eben „grounded“ ist.[18]

Doch so falsch es wäre, auf dieses System Bewusstsein zu projizieren, so sehr sollte man diese Schwundstufe von Bedeutung nicht allzu schnell abtun. Sofern Sprachmodelle implizites Wissen auf eine nichttriviale Weise explizit machen – und sei es nur durch Matrixtransformationen in einem Vektorraum –, stellen sie jene dumme Bedeutung her, die ohne sie nicht zu haben gewesen wäre.[19] Anders als bei ELIZA – dessen *x* und *y* für das System nur leere Platzhalter waren – sind neuronale Netze dabei nicht *allein* parasitär auf die Bedeutungszuschreibungen von menschlichen Agenten angewiesen, sondern operieren *auch* produktiv mit der Eigenstruktur von Sprache.

Text und Bild und Welt

Emily Bender hat freilich Recht damit, dass LaMDA nicht geerdet ist.[20] Es ist ein *mono*modales Netz, verarbeitet nur einen einzigen Typ von Daten, nämlich Text. Um im Harnads Sinne „grounded“ zu sein, schreibt sie, wäre es notwendig, mehre Datenarten miteinander zu verknüpfen – es müsste *multi*modales Machine Learning sein.[21] Multimodal aber ist Dall·E 2: Statt dass Text nur auf anderen Text verweist, ist hier Text mit Bildinformation korreliert. Damit wird die Hoffnung geschürt, arbiträre Zeichen könnten mit Dingen in der Welt verbunden werden, um so geerdete Bedeutung zu produzieren.

Harnads Hypothese, dass gerade neuronale Netze dem Symbol Grounding Problem begegnen könnten, haben jüngst die Medienwissenschaftler Leif Weatherby und Brian Justie mit ihrem Begriff „indexikalischer KI“ aufgegriffen. Benannt ist sie nach Charles Sanders Peirce’ Begriff des Index. Anders als das Symbol, das zu seinem Bezeichneten in einem rein konventionellen Verhältnis steht (wie „Hund“, „chien“ und „dog“ alle dasselbe meinen), ist der Index kausal mit ihm verknüpft (wie Rauch auf Feuer verweist).

---

[18] Emily M. Bender/Alexander Koller, *Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data*. In: *Proceedings of the 58th Annual Meeting of the ACL*. Atlanta: ACL 2020.

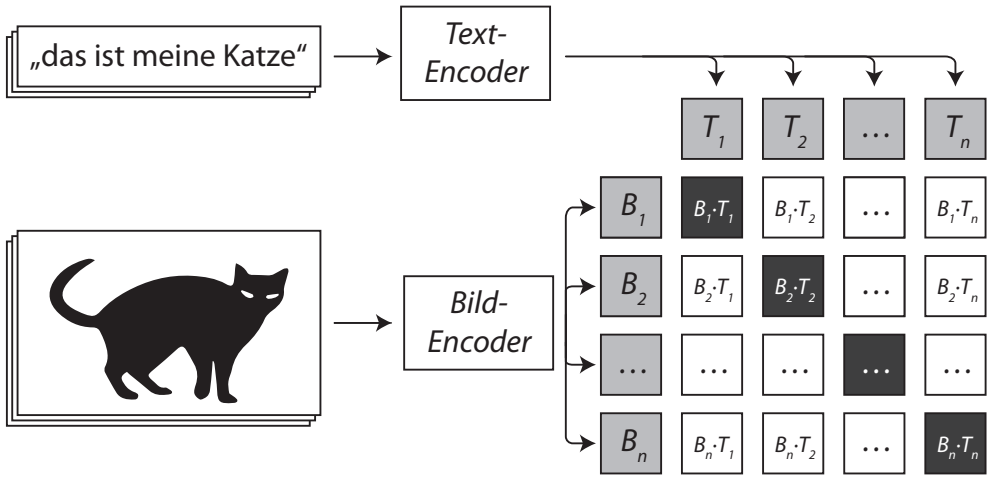
[19] Das setzt voraus, dass in dieser Operation wirklich etwas Neues gefunden und nicht nur eine Tautologie ausgefaltet wird; ein Modell dieser Idee wäre etwa Kants Überzeugung, Sätze der Mathematik seien synthetische Urteile apriori, stellten also wirklich neues Wissen her.

[20] Zwar beansprucht auch das Paper, in dem LaMDA vorgestellt wird, „groundedness“ für das Modell, gemeint ist damit aber nur, dass sich die Ausgaben von LaMDA als „Behauptungen über die Außenwelt auf maßgebliche externe Quellen stützen“. Als *textliche* Quellen sind sie weiterhin Teil des Harnad’schen „Symbolkarussells“. Vgl. Romal Thoppilan u.a., *LaMDA: Language Models for Dialog Applications*. In: *arXiv* vom 10. Februar 2022 (arxiv.org/abs/2201.08239).

[21] Vgl. Gadi Singer, *Multimodality: A New Frontier in Cognitive AI.* In: *Medium* vom 2. Februar 2022 (towardsdatascience.com/multimodality-a-new-frontier-in-cognitive-ai-8279d00e3baf).

Die Autoren machen mit dieser Prägung Harnads Projekt unter der Hand zur Grundlage von Gegenwartsbeschreibung: „Digitale Systeme, die sich auf das neuronale Netz stützen, haben die Welt der bloßen Symbole hinter sich gelassen und haben begonnen, sich *hier*, *jetzt* für *Sie* zu erden – sie sind in der Lage, auf reale Zustände hinzuweisen.“[22] Neuronale Netze bringen die Welt – als die Daten, auf die sie trainiert wurden – in den Computer und steigen so aus dem solipsistischen „Symbolkarussell“ aus. Nirgends scheint sich das plausibler zu zeigen als in Dall·E 2.

Herz von Dall·E 2 ist ein Machine-Learning-Modell namens CLIP. Ihm werden über einen Encoder vektorisierte Text-Bild-Paare vorgelegt, die aus dem Internet stammen – etwa ein Foto einer Katze mit der Bildunterschrift „das ist meine Katze“. CLIP wird darauf trainiert, vorherzusagen, welcher Textvektor zu welchem Bildvektor passt; das Ergebnis ist ein umfangreiches stochastisches Modell, das Bildinformationen mit Textinformationen korreliert, sie aber als *eine* Art von Information speichert. In der folgenden Abbildung ist das die Tabelle, in der das Skalarprodukt der Text- und Bild-Vektoren eingetragen ist – je besser Text und Bild zueinander passen, desto besser dieser Wert; bei der Paarung von Ausgangsbild und -text ist er natürlich optimal (das sind die schwarzen, diagonal verlaufenden Kästchen).



Text-Bild-Korrelation in CLIP

[22] Leif Weatherby/Brian Justie, *Indexical AI*. In: *Critical Inquiry*, Nr. 48/2, Winter 2022. Eine Schwierigkeit bei diesem Begriff ist die Frage, ob in einem neuronalen Netz bereits *alle* Daten als indexikalisch zu gelten haben (das würde den Text von LaMDA einschließen) oder nur solche, die unmittelbar durch körperlichen Sinnen nachempfundenen Sensoren gewonnen wurden (das wären Bilder, aber nicht Text). Weatherby und Justie scheinen Ersteres, Harnad Letzteres im Sinn zu haben. Dieser spricht daher an einer Stelle von „ikonischen“ Daten – dem dritten Zeichentyp von Peirce, der nach dem Prinzip der Ähnlichkeit von Zeichen und Bezeichnetem funktioniert. Da aber auch diese indexikalisch sind, so sie aus Sensoren stammen (was sie freilich auf visuelle Daten beschränkt), scheint mir das Argument von Weatherby/Justie und das von Harnad auf etwas strukturell Ähnliches hinauszulaufen – beiden geht es um die Verbindung von System und Welt.

CLIP ist damit zunächst ein erstaunlich gutes Bild*erkennungs*programm: Legt man ihm ein unbekanntes Katzenfoto vor, wird es trotzdem als „Katze“ erkannt. Erst in einem zweiten Schritt wird es auch zum Bild*generator*. Dazu arbeitet es im Verbund mit einem zweiten KI-Modell namens GLIDE, das bereits auf einen großen Datensatz von Bildern trainiert wurde.[23] Gibt der User einen Prompt ein, kann GLIDE die im CLIP-Modell gespeicherten Text-Bild-Daten verwenden, um diesen Prozess umzukehren und ein Bild zu synthetisieren, das am besten mit dem Eingabetext korreliert. In beiden Operationen – Bilderkennung wie Bilderzeugung – ist wieder zentral, dass die Modelle die *Korrelation* zwischen Textbeschreibungen von Objekten und ihren entsprechenden visuellen Manifestationen lernen und aktiv reproduzieren können.

Nun kann man einwenden, dass die mit dem Wort „Katze“ korrelierte Bildinformation, in der das Fotos einer Katze gespeichert ist, zwar in einem indexikalischen Verhältnis zu dieser Katze stehen mag – Licht wurde von ihr reflektiert und fiel auf einen Fotosensor etc. –, dass aber auch so das System nicht lernen wird, was es heißt, mit einer Katze zusammenzuleben. Verfechter des Symbol Grounding versuchen daher, neben sensorischem auch motorisches und schließlich gar soziales Feedback einzuspeisen: Erst durch die Wirkungen von Sprachgebrauch in der Gemeinschaft von dieselbe Welt bewohnenden anderen Sprechern kann Bedeutung erlernt werden.[24]

Aber dieser Anspruch hieße wieder, „volle“ menschliche, also breite Bedeutung einzufordern und alles unterhalb dessen nicht recht ernst zu nehmen. Stattdessen sollte man multimodale KI als zweiten Grad dummer Bedeutung betrachten. Der Peirce’sche indexikalische Verweis auf etwas außerhalb des Modells scheint jedenfalls etwas anderes zu sein als der de Saussure’sche differenzielle Verweis auf andere Elemente des Modells – und sei es nur, dass die Dimension möglicher Korrelationen zunimmt, und damit auch die Möglichkeit, ungeahnte latente Verbindungen, ungeahnte dumme Bedeutung zutage zu fördern.

In der Tat sind multimodale KIs – neben Dall·E 2 etwa das kostenlose, aber nicht verwandte Dall·E Mini (jetzt Craiyon), Stable Diffusion, Googles noch nicht freigegebenes Imagen oder das Bezahlmodell Midjourney – in der Lage, sehr komplexe Text-Bild-Bedeutungen zu

[23] GLIDE ist ein *diffusion model*, das auf thermodynamischen Modellen basiert, und funktioniert damit anders als die noch bis vor kurzem beliebten GANs, die zwei agonale Teilmodelle vereinen. Vgl. Prafulla Dhariwal/Alex Nichol, *Diffusion Models Beat GANs on Image Synthesis*. In: *arXiv* vom 1. Juni 2021 (arxiv.org/abs/2105.05233). Dass die für ein ästhetisches Werk verwendeten KI-Architekturen selbst eine Ressource für die Diskussion dieses Werks sein können, habe ich vorgeschlagen in Hannes Bajohr, *Algorithmische Einfühlung. Für eine Kritik ästhetischer KI*. In: ders., *Schreibenlassen. Texte zur Literatur im Digitalen.* Berlin: August Verlag 2022.

[24] Vgl. Yonatan Bisk u.a., *Experience Grounds Language*. In: *arXiv* vom 2. November 2020 (arxiv.org/abs/2004.10151).

generieren. Ihre Mächtigkeit liegt in einer Fähigkeit begründet, die eine produktive Qualität solcher Korrelationen nahelegt: Bei der Untersuchung der Tiefenstruktur von CLIP fanden Informatiker heraus, dass das Modell einzelne „Neuronen" ausgebildet hatte, die sowohl für das Wort wie für das Bild einer Sache feuerten – es waren *konzeptuelle* Neuronen, in denen der Unterschied zwischen Bild und Text tendenziell aufgehoben ist.[25] Multimodalität ist, auf der neuronalen Ebene, in Wirklichkeit *Pan*modalität, die eine Semantik ohne klar differenzierte Zeichensysteme nahelegt. Dumme Bedeutung wechselt hier ihre Ebene, ist weder an Text- noch an Bilddaten gebunden, sondern umfasst beide auf eine Weise, die auf eine Bedeutung jenseits der modalen Trennung verweist – und die wieder nichts mit Geist zu tun hat.

Promptologische Untersuchungen

KI-Systeme *sind* dumm. Sie haben kein Bewusstsein. Dennoch produzieren sie eine komplexe artifizielle Semantik, die quer zu unseren gewöhnlichen Vorstellungen von Bedeutung liegt. Multimodale KI zeigt zudem, dass unterstelltes Bewusstsein und die Bedeutungsmächtigkeit eines Systems kaum etwas miteinander zu tun haben: Dass Lemoine gerade LaMDA wie eine Person vorkam – und nicht Dall·E 2, das doch eigentlich eine höhere, weil korrelationsreichere Stufe der KI-Entwicklung darstellt –, liegt schlicht daran, dass man dem dialogisch operierenden Sprachmodell kommunikative Absicht unterstellt, dem Bildgenerator dagegen nicht; Sprache scheint immer klüger zu sein als das Bild.

Bedeutung jenseits kommunikativer Absicht muss dabei aber nicht *bloß* parasitär sein, wie die Vektoroperationen von Word Embeddings und die konzeptuellen Neuronen von Text-to-image-KIs zeigen. Dass sie immer *auch* parasitär ist, liegt daran, dass die Trainingsdaten einer menschlichen Welt entstammen und artifizielle Semantik eben keine „Robotersprache" ist, sondern ein Korrelationseffekt von je interpretierbaren Daten. Dennoch wäre auf lange Sicht eine Angleichung von dummer und breiter Bedeutung denkbar, nämlich dann, wenn sie in einander beeinflussende Kreisprozesse eingehen.

Die Schnittstelle zwischen natürlicher und artifizieller Semantik ist im Fall von Dall·E 2 die Interaktion per Prompt. Das *prompt design* – die genaue, geradezu virtuose Auswahl der Texteingabe – kann einerseits analytisch eingesetzt werden, nämlich um den Vektorraum dummer Bedeutung auf Spuren kulturellen Wissens hin abzugrasen. Damit würde die breite Bedeutung natürlicher Sprache in ihrer Interaktion mit dummer Bedeutung wieder wichtiger.

[25] Gabriel Goh u.a., *Multimodal Neurons in Artificial Neural Networks*. In: *Distill* vom 4. März 2021 (distill.pub/2021/multimodal-neurons/).

Eine „Promptologie“, die sich einer solchen natürlich-artifiziellen Verbindung annimmt – der Korrelation von datafizierter Sprache und der kulturellen Bedeutung, die dieser Sprache auf Rezipientenseite zugesprochen wird –, wäre ein Einfallstor für die Kultur- und Geisteswissenschaften, die mit ihrem Wissen um solche weichen Faktoren wie Stile, Einflüsse, Ikonographie, etc. ihren Beitrag zur Erforschung digitaler Artefakte leisten können, ohne notwendig die Form der Digital Humanities anzunehmen; sie könnten phänomenorientiert arbeiten und sich den Artefakten, die die Modelle ausspucken, als Grenzobjekten zwischen menschlicher und maschineller, zwischen breiter und dummer Bedeutung widmen.

Zugleich aber ist Promptologie kein bloß analytisches Verfahren, sondern immer auch eine Praxis mit einem eigenen Wissen, das viel mit einer geradezu „einfühlenden“ Interaktion mit dem KI-System zu tun hat.

So hat sich herausgestellt, dass bei Text-to-image-KIs diese Prompts allein durch bestimmte, oft kontraintuitive oder absurde Formulierungen in ungeahnte Richtungen gelenkt werden können (inzwischen gibt schon ein Start-up, PromptBase, das besonders effektive Prompts verkauft).[26] Statt sich also das System untertan zu machen und es als Instrument zu verwenden, muss sich im täglichen Gebrauch die natürliche Sprache der artifiziellen Semantik angleichen, die von jener durchaus abweicht.

Das Ergebnis ist eine Feedbackschleife von artifizieller und menschlicher Bedeutung: Nicht nur lernt die Maschine, die Semantik von Wörtern mit der von Bildern zu korrelieren, die wir ihr gegeben haben, sondern wir lernen, die Dummheit des Systems in unsere Interaktion mit ihm einzupreisen; diese Angleichung wäre vielleicht nicht kommunikativ in einem starken, aber vielleicht in einem schwachen, eben dummen Sinne.

[26] Kyle Wiggers, *A startup is charging $1.99 for strings of text to feed to Dall-E 2*. In: *TechCrunch* vom 29. Juli 2022 (techcrunch.com/2022/07/29/a-startup-is-charging-1-99-for-strings-of-text-to-feed-to-dall-e-2/). Interessant hierbei ist, dass der diskutierten tendenziellen Aufhebung der Sprache/Bild-Unterscheidung auf *technischer* Ebene die Verdrängung des Bildes durch Sprache auf *Interface*ebene gegenübersteht. Die Ergebnisse von Dall·E 2 könnten daher auch als Sprachkunst verstanden werden, statt nur visuelle Objekte zu sein.