In: Digitale Literatur II, hg. v. Hannes Bajohr und Annette Gilbert. München: edition text+kritik 2021, 174–85.

Hannes Bajohr

Künstliche Intelligenz und digitale Literatur

Theorie und Praxis konnektionistischen Schreibens

Keine Literaturform ist so offensichtlich an die Medientechnologien ihrer Hervorbringung gebunden wie digitale Literatur und keiner anderen ist daher ihr Zeitbezug so unmittelbar eingeschrieben. Am deutlichsten fällt das anhand jener Werke auf, die unlesbar wurden, weil für ihre Ausführung obsolet gewordene Technik notwendig ist; umgekehrt aber erscheinen immer wieder neue Formate, Medien und Techniken, die poetisch urbar gemacht werden können und deren Potenzial anfangs nicht recht abzusehen ist. Mit künstlicher Intelligenz in Form neuronaler Netze zur Produktion natürlichsprachlicher Texte (›natural language processing‹) hat seit wenigen Jahren eine solche neue Technik Konjunktur. Dieser intensivierte Zeitbezug macht eine Diskussion von literarischer KI einigermaßen prekär: Einerseits ist der Kanon ihrer Werke noch klein und im Fluss, andererseits verläuft technischer Fortschritt hier besonders rasant. Da ich dies schreibe, ist das Sprachmodell GPT-3 des Thinktanks OpenAI der Goldstandard automatischer Textproduktion; das kann schon zum Zeitpunkt der Veröffentlichung wieder anders aussehen. Das vorausgeschickt soll dennoch versucht werden, im Folgenden eine kurze Darstellung theoretischer Herausforderungen durch und gegenwärtiger Schreibpraxis mit KI zu geben.

Künstliche Intelligenz als Co-Creative Writing

»Künstliche Intelligenz« ist eine irreführende Bezeichnung. Populär dominiert die Vorstellung von künstlicher menschenähnlicher Intelligenz (»strong AI«);[1] als Teilbereich der Informatik meint KI dagegen weniger, nämlich alle Versuche, rationales Verhalten und Problemlösen überhaupt zu simulieren. In der Praxis bedeutet KI gerade kein umfassendes Weltverstehen, sondern auf enge Sachbereiche zugeschnittene automatisierte Verfahren (»narrow AI«),[2] die sich gegenüber älteren Ansätzen dadurch auszeichnen, ohne explizite Regeln selbstständig anhand einer großen Menge von Beispielen (»big data«)[3] zu lernen. Wer heute ›Künstliche Intelligenz‹ sagt, meint daher so gut wie immer ›Machine Learning‹; innerhalb dieses Feldes wiederum dominiert der Ansatz des ›Deep Learning‹, das auf künstlichen neuronalen Netzen (KNNs) aufbaut.[4]

Bereits jetzt ist Machine Learning an alltäglichen Schreibprozessen beteiligt: Autokorrektur- und Autocomplete-Funktionen beruhen nicht weniger als Spracherkennung und -ausgabe auf maschinellen Lernverfahren, dasselbe gilt für Übersetzungssysteme wie DeepL oder Google Translate. In allen diesen Fällen spielt KI eine meist für die Schreibenden unbewusste und sie entlastende Rolle. Sie kann aber auch ganz explizit und poetologisch bewusst zur Produktion von Literatur verwendet werden.[5] In diesem Fall wiederholt sich jedoch die oben genannte Differenz: So wenig es heute eine »strong AI« gibt, so wenig existieren KI-Systeme, die tatsächlich *autonom* Kunst oder Literatur herstellen könnten; alle gegenteiligen Behauptungen beruhen noch entweder auf Publicityerwägungen oder auf der bewussten Verschleierung des menschlichen Anteils an der Werkproduktion.[6]

So wurde der per KNN generierte Roman »1 the Road« (2018) von seinem Verlag auf einer Bauchbinde explizit als das »erste von einer künstlichen Intelligenz geschriebene Buch« beworben und der hinter diesem Projekt stehende Ross Goodwin konsequent als »writer of writer« bezeichnet.[7] Dagegen spricht freilich, dass Goodwin einen Cadillac mit Kameras, Mikrofonen und GPS ausstattete, deren Daten auf einem Roadtrip von New York nach New Orleans beständig in sein KNN einspeisen und dessen Ergebnis während der Fahrt auf einem Thermodrucker ausgeben ließ: Goodwin war sowohl *konzeptuell* an der Inszenierung der Schreibszene als Reperformance von Jack Kerouacs »On the Road« als auch *physisch*, nicht zuletzt als Fahrer beteiligt. Wirklich selbstfahrende Autos gibt es, allen KI-Versprechungen zum Trotz, noch nicht. Dass »1 the Road« auf dem Rückdeckel als »book written by a car as a pen« bezeichnet wird, führt die Behauptung der *alleinigen* KNN-Autorschaft und der *starken* künstlerische KI ad absurdum. Wenn auch das Auto Autor ist, könnte Goodwin schließlich ebenso gut als ›driver of writer‹ firmieren.

Dieses Zusammenspiel involvierter Systeme zeigt einerseits aufs Beste, dass es sinnvoller sein mag, Experimente mit Machine Learning und Literatur als Instanzen eines ›Co-Creative Writing‹ zu betrachten, bei dem immer eine Mensch-Maschinen-Assemblage gemeinsam an der Textproduktion beteiligt ist, statt nach autonomer KI-Autorschaft zu fahnden.[8] Andererseits ist sowohl die Initiations- und Konzeptgestaltungsmacht, die alle beteiligten Systeme in Gang setzt, als auch die Anerkennungslogik des Produkts als literarisches Werk bislang noch bei Menschen zu suchen. Das ist nicht notwendig eine »nahezu reaktionäre Revision des Autorbegriffs«, die aus Angst vor autonomen Maschinen eine anthropozentrische Wende vollzieht,[9] sondern vielmehr die Konsequenz der Tatsache, dass es eben noch keine wirklich autonome künstlerische KI gibt. Denn damit ein *allein* von einem KI-System hergestelltes Kunstwerk als solches auch praktisch anerkannt würde, müsste das System den Status eines sozialen Akteurs besitzen. Starke künst-

lerische KI hätte daher weniger den Turing-Test zu bestehen (= die Maschine geht fälschlich als Mensch durch) als vielmehr den »Durkheim-Test« (= Maschine und Mensch sind tatsächlich gleichberechtigt Handelnde).[10] Solange dies nicht der Fall ist, braucht es immer noch menschliche Instanzen, die maschinengenerierte Texte als literarische Lektüre markieren.

Ein Bruch in der Geschichte generativer Literatur

Diese beiden Pole – komplexe Mensch-Maschine-Assemblagen und das unüberwindliche Residuum humaner Autorschaft – sind für die lange Tradition des generativen Schreibens nichts Neues. Ein Projekt wie »1 the Road« ließe sich daher auf den ersten Blick als lediglich jüngster Eintrag in ihrer Geschichte betrachten und wäre dann nicht wesentlich verschieden von anderen in Programmcode formalisierten Algorithmen zur Textproduktion. Diese gibt es, seit Christopher Strachey 1952 mit dem Manchester University Computer seine »Love Letters« generierte, für die er eine Liste an Verben, Nomen und Adjektiven in eine Reihe von Briefschablonen einsetzen ließ: »*My*—(adj.)—(noun)—(adv.)—(verb) *your*—(adj.)—(noun)« wird so zu »My sympathetic affection beautifully attracts your affectionate enthusiasm.«[11] Für viele Forscher*innen ist es zudem verführerisch gewesen, solche parametrisierte Literaturproduktion noch tiefer in der Geschichte zu verorten und eine Genealogie der generativen Literatur bis in das Barock, sogar bis in die Antike nachzuverfolgen.[12] Darin manifestiert sich die Idee der ›Textmaschine‹ als Kalkül, die sowohl in einem digitalen Automaten als auch einem menschlichen Akteur implementiert werden kann, da Algorithmen lediglich eine Abfolge von eindeutigen Regelschritten beschreiben.

Es ist aber fraglich, ob solche Linienbildung nicht auch wesentliche Unterschiede verwischt: Zwar stimmt es, dass in beiden Fällen Computer an der Textproduktion beteiligt sind; betrachtet man aber das zugrundeliegende technische Substrat, fällt auf, dass Operationen in KNNs gerade *nicht* in gewohnter Weise als Regelschritte ausgedrückt werden können. Ich möchte daher für einen Bruch in der Geschichte generativer Literatur argumentieren. Dieser Bruch ist, trotz Vorläufern in den 1950er Jahren, nicht viel älter als ein Jahrzehnt.[13] Er trennt KNNs als *konnektionistisches Paradigma* von klassischen Algorithmen als *sequenzielles Paradigma*.[14]

Stracheys Liebesbriefgenerator gehört zum sequenziellen Paradigma. Er besteht aus einer fixen Liste an Wörtern (einer ›Datenbank‹) und einer Schritt für Schritt ausgeführten Regel zu ihrer Zusammensetzung (einem ›Algorithmus‹). Dagegen folgt das KNN-generierte »1 the Road« dem konnektionistischen Paradigma und einem anderen Prinzip. Vage vom Aufbau des Hirns als Verbindung von ›Neuronen‹ und ›Synapsen‹ inspiriert, model-

lieren gegenwärtige KNNs (›deep neural networks‹) eine mathematische Funktion, die aus Inputdaten statistisch ähnlichen Output produziert. Es stehen also in KNNs gerade keine Herstellungsanweisungen, sondern Daten am Anfang, und aus ihnen wird erst durch einen iterativen Lernprozess das Modell gebildet; dieses Modell wiederum ist kein Algorithmus, sondern beschreibt lediglich die Verbindungsstärken zwischen den ›Neuronen‹ in einem sogenannten Gewichtungsmodell.

Diese technischen Differenzen sind tief in der Architektur der jeweiligen Systeme verankert. Verlaufen die Rechenoperationen im klassischen Algorithmus sequenziell nacheinander, führen KNNs ihre Berechnungen parallel in den zahlreichen Konnexionen zwischen den Neuronenschichten aus. KNNs folgen dabei zudem einer stochastischen statt einer rein deterministischen Logik, in der Schnelligkeit auf Kosten absoluter Exaktheit geht (›approximate computing‹). Schließlich sind in KNNs Daten und Verarbeitungsanweisungen nicht getrennt, sondern beide Funktionen des Gewichtungsmodells. Folge ist unter anderem das bekannte ›Black-Box‹-Problem: Die Gewichtungsmodelle von KNNs lassen, anders als klassische Codes, vergleichsweise wenig Aufschluss über ihre inneren Abläufe zu, sind weder durch äußere Beobachtung des Outputs noch selbst durch das Wissen um die Details des Modells in jene eindeutigen Ablaufpläne übersetzbar, die einen klassischen Programmcode auszeichnen.

Neue Vokabulare und Poetiken

Die technischen Differenzen zwischen konnektionistischem und sequenziellem Paradigma haben Konsequenzen für das Reden über mit KNNs produzierten Texten. Das seit mehr als 30 Jahren geformte Handwerkszeug zur Beschreibung digitaler Kunst und Literatur greift an vielen Stellen nicht mehr. Dass etwa Scott Rettberg in seinem Buch »Electronic Literature« KNNs nur am Rande, in einem Kapitel über »Combinatory Poetics«, abhandelt,[15] drückt die Verlegenheit angesichts dieses neuen Phänomens aus: Kombinatorisch im Sinne eines »paradigm of database and algorithm«[16] wie bei Strachey sind KNNs gerade *nicht*. Wo sequenzielle Algorithmen eine strikte Trennlinie zwischen prozeduralen Regeln und Elementen in einer Datenbank ziehen, ist das ›Wissen‹ in einem neuronalen Netz nicht an einem bestimmten Ort lokalisiert, sondern über das System verteilt.[17] Das konnektionistische Paradigma zeigt so auch die Notwendigkeit, neue theoretische Vokabulare zur Interpretation seiner Werke auszubilden.[18]

Der Bruch zwischen den Paradigmen berührt auch die Autorschaftsfrage, deren menschliche Seite im Mensch-Maschine-Gefüge eine zunehmende Distanzierung durchläuft: Konnte man im sequenziellen Paradigma noch

plausibel von *sekundärer* Autorschaft sprechen, die in der Formulierung einer Regelfolge besteht, deren Ausführung das Werk produziert – hier ergibt die Idee eines »writer of writer« durchaus Sinn –, steht man bei KNNs vor einer *tertiären* Autorschaft: Es bleiben allein der Datensatz für das Training zu definieren, aus denen das KNN selbstständig das Modell bildet, und die Parameter zu bestimmen, mittels derer das Modell schließlich den Output hervorbringt.[19] Bei großen Sprach-KIs wie GPT-3 ist selbst das nicht mehr möglich, denn das Training ist hier zu aufwändig, um es auf je neue Datensätze abzustimmen. Die ›Programmierung‹ erfolgt durch die normalsprachliche Formulierung von Aufforderungen (›prompt design‹) nach dem Vorbild dialogischer Kommunikation[20] – hier wäre gar von *quartärer* Autorschaft zu sprechen.

Auch die Folgen für die Ästhetik der so produzierten Werke sind enorm. So steht etwa mit der Undurchsichtigkeit von KNNs die Poetik der Transparenz infrage, der viele Werke der am Conceptual Writing geschulten generativen Literatur der Gegenwart folgen.[21] Der im sequenziellen Paradigma operierende Nick Montfort bemängelt daher explizit, dass KNNs »bedazzle and obfuscate« und damit der Open-Source-Ethik des generativen Schreibens widersprechen. Mit seinen Werken möchte er dagegen zeigen, »that interesting computational manipulation of language can be done with systems that are simple and comprehensible«.[22] Andererseits fordern KNNs gerade aufgrund ihrer Opazität kritische Perspektiven heraus. So weist Jörg Piringer darauf hin, dass bereits die politische Bedeutung von KNNs eine künstlerische Auseinandersetzung mit ihnen notwendig macht: Wenn KNNs, wie im Fall von GPT-3, eine gewisse Größe erreichen und ihr Training nur noch von finanzstarken Unternehmen geleistet werden kann, sind sie immer auch in Kapital- und Machtverhältnisse verstrickt. Digitale Literatur ist nun dazu aufgerufen, »gesellschaftliche umgangsformen mit sprachtechnologien – und methoden der kritik an ihr – zu entwickeln und so die begehrlichkeiten der internetgiganten, nach den netzen und kommunikationsgeräten auch noch die sprache zu kontrollieren, abzuwehren«.[23]

Lean in: affirmative Kontrollabgabe

Das konnektionistische Paradigma verstärkt die Tendenz zur »Verteilung und Zerstäubung« von Autorschaft, die Mensch-Maschine-Assemblagen immer anhaftet.[24] Verglichen mit den (ohnehin idealisierten) Vorstellungen auktorialer Kontrolle in sekundärer Autorschaft müssen sich jeweils weiter distanziertere Autorschaften zur Erfahrung einer zunehmenden Kontroll*abgabe* verhalten. Schreibenden bleiben dabei grundsätzlich zwei Möglichkei-

ten: *lean in or resist.* Im Folgenden will ich Werke vorstellen, die sich entweder ganz der Opazität des Systems hingeben oder umgekehrt versuchen, Licht ins Dunkel der Black Box zu bringen. Zur simpelsten Technik der ersteren Form gehört die ästhetische Affirmation von Inkongruenz und Absurdität, die auch im sequenziellen Paradigma existiert.[25] So hat sich Janelle Shane, Eigenbezeichnung »AI humorist«, auf die bewusste Überforderung des begrenzten Weltverständnisses von Machine-Learning-Systemen verlegt. Auf Tausende von Kochrezepten trainiert, sind die KI-generierten Kreationen mit Sicherheit kaum genießbar oder semantisch sinnvoll (»Bright Grilled Evaporated Milk«),[26] und Shanes »AI weirdness« zieht ihren Humor gerade aus der Nicht-Intelligenz von KNNs, deren »cuteness«[27] weniger der Aufklärung als der Nachsicht bedarf.

Versteht sich Shane als nicht-literarische Humoristin, schließen andere Autor*innen bewusst an literarische Traditionen des Absurden an. »Sunspring« (2016), ein Drehbuch Ross Goodwins, wurde von einem auf Science-Fiction-Scripts trainierten KNN generiert und anschließend verfilmt.[28] Das Sprunghafte der Dialoge, die Alogismen der Regieanweisungen werden gerade in der Kurzfilmfassung zu einer Aktualisierung des absurden Theaters, demgegenüber »Sunspring« semantische Kohärenz und Kohäsion weitaus radikaler auflöst. Wie auch »1 the Road« – das nur deshalb nicht unmittelbar absurd erscheint, weil die Sätze protokollartig durch Zeitstempel organisiert sind – beruht »Sunspring« auf einer LSTM-RNN genannten Netzwerkarchitektur. Dieses »long short-term memory recurrent neural network« war 2015 dafür verantwortlich, KI-Textgenerierung breitenwirksam zu popularisieren,[29] konnte aber immer noch nur sehr begrenzt kohärente Ausgaben produzieren. Absurdität anzunehmen reagiert zunächst also auf ein doppeltes *contrainte*: einerseits auf die Intransparenz des Systems und andererseits auf seine technische Beschränktheit. Gelegentlich führt das zum Verdacht, »dass die Probleme der technisch generierten Sprache durch pubertäre Krassheiten nobilitiert werden sollen«.[30]

Mit der Einführung der sogenannten Transformer-Architektur, die dem KNN eine höhere ›Aufmerksamkeit‹ für Satzkontexte verleiht und auf denen die großen Sprachmodelle, wie GPT-2 und, als bislang mächtigste Ausführungen, BERT und GPT-3 aufbauen, hat sich die Kohärenz der Textausgabe dramatisch verbessert, wobei aber immer noch kein echtes semantisches Verständnis erreicht ist. Die Entscheidung zum Absurden ist damit nun Sache bewusster Rahmung geworden. So bezieht sich Vladimir Alexeev (Pseudonym »Merzmensch«) offensiv auf Kurt Schwitters und aktualisiert dessen Zeitschrift »Merz« mit seinem eigenen »Merz AI«, in dem er auch auf die collagenhafte Seitengestaltung des Originals anspielt. Weil er GPT-2 mit Texten von Schwitters und anderen Avantgardisten fütterte, ist die dadaistische Anmutung nicht mehr nur technisches Artefakt,

sondern auch Effekt des Datensets selbst.[31] Alexeev verwendet zudem GPT-3, das hundertmal mächtiger ist als GPT-2 und daher zu groß, um eigens mit Textmaterial trainiert zu werden – er muss sich darauf verlassen, dass das Sprachmodell Schwitters bereits im Lernprozess begegnet ist. Ein über Schwitters generierter Essay schlägt diesen ohne Umschweife dem Surrealismus zu; dass das nicht stimmt, bemerkt Alexeev selbst, sieht in dieser Fehlattribution aber ausdrücklich eine Schwitter'sche Geste.[32]

In der Tat ist ›surrealistisch‹ neben ›dadaistisch‹ eine zweite oft zu hörende Beschreibung KI-generierter Literatur. Gerade das Kontrollabgabemodell pflegt, wie Simon Roloff schreibt, ein »Traditionsbewusstsein«, das den ›halluzinatorischen‹ Strang der Moderne aufruft.[33] Bretons Feststellung, »eigentlich bewähren sich die Formen der surrealistischen Sprache am ehesten im Dialog«,[34] wird dann auch dort ernst genommen, wo das KNN als Kooperationspartner einer verteilten *écriture automatique* fungiert. Da vor allem große Sprachmodelle mit On- und Offline-Text trainiert wurden, können ihre Äußerungen als Artikulationen eines kollektiven Unbewussten gelesen werden. Das ist der Ansatz K Allado-McDowells, der*die auch an »1 the Road« beteiligt war und für »Pharmako-AI« den Dialog mit GPT-3 als kollaborativen Schreibprozess versteht. Dieser stelle ein gemeinsames Unbewusstes her, indem sich das Vokabular des Menschen mit dem des Modells vermische. Folge sei, »that the reader is hard pressed to see the separation between human and AI«. Die Abgabe von Kontrolle, zumal im quartären Autorschaftsmodell, imaginiert an seinem Extrem die völlige Verschmelzung von Mensch und KI.[35]

Resist: Kritik und Selbstermächtigung

Auf der anderen Seite des Spektrums stehen Versuche, die Opazität von KNNs technisch und poetologisch zu bändigen. Eine Art Mittelstellung nehmen dabei Werke ein, die ebenfalls Dialogizität betonen, aber versuchen, sie nachvollziehbar zu machen oder zur bloß subalternen Assistenzleistung herabzustufen. In Mattis Kuhns »Selbstgespräche mit einer KI« geht es zwar auch um die Neigung zur »Verbindung (oder Verschmelzung)« von menschlichem und KI-Autor, der dann eine Art maschinell vermittelte Selbsterkenntnis folgt (»eine Maschine, die *mich* schreibt«).[36] Zugleich aber bildet die intensive Kuratierung und Dokumentation des Projekts eine Gegenbewegung zum fröhlichen Obskurantismus Allado-McDowells: Nicht nur ist das seinem selbsttrainierten Modell zugrunde liegende sehr kleine Korpus von lediglich 2732 Sätzen aus literarischen, geistes- und naturwissenschaftlichen Texten vollständig abgedruckt, auch werden die verwendeten Verfahren genau erklärt und sogar eigens die Trainingscodes

dokumentiert – freilich ohne das alles entscheidende Gewichtungsmodell, dessen Abdruck auch kaum erhellend wäre.

Eine andere Art von Dialog, die dem menschlichen Anteil an der Autorschaft eine wieder sehr viel dominantere Rolle einräumt, bietet David »Jhave« Johnstons »ReRites«.[37] Zwischen Mai 2017 und Mai 2018 trainierte Jhave jede Nacht ein KNN und edierte am nächsten Morgen den generierten Output in einem Prozess, den er »carving« nennt, von Hand: Der »block of generated text, massive and incomprehensible«, ist dabei nur das von einer »assistive technology« vorbereitete Zwischenfabrikat, dem erst durch die mühsame ›Meißel‹-Arbeit in einem Texteditor die endgültige Form gegeben und ihr lyrischer Mehrwert gesichert wird.[38] Die Ergebnisse eines jeden Monats sammelte Jhave in je einem Buch, sodass »ReRites« 12 Bände umfasst.[39] Zugleich macht die Rahmung als jahrlanges Ritual (›rite‹) des Umschreibens (›rewrite‹) das Projekt zu einer Langzeitperformance, die an die Arbeiten des Performance-Künstlers Tehching Hsieh denken lässt.[40] Ohne dieselbe Entsagung Hsiehs zu erreichen, der etwa ein Jahr in einem Käfig lebte (»Cage Piece«, 1978–1979), ist doch die Körperlichkeit des Künstlers Bedingung der Werkautorisierung. Dass dieser in einem Auswahlband seine Gedichte dem unedierten und nicht eigens als Werk markierten »Raw Output« gegenüberstellt, betont dabei nur das klare Autorschaftsgefälle: Jhave ist zwar ein »augmented poet«, die KI aber nicht mehr als eine »machine-enhanced muse«.[41]

Gewinnt diese auktorielle Selbstermächtigungsgeste zwar eine gewisse Kontrolle zurück, leistet sie für Lesende keinen Beitrag zum Verständnis des Systems (auch wenn der Autor, in einer Art ›artistic research‹ mehr über diese Technologie lernt). Tiefer in die Sprachcodierung von KIs dringt Allison Parrish in ihren jüngsten Werken ein. Damit Sprache im Machine Learning statistisch verarbeitet werden kann, werden Wörter numerisch, als hochdimensionale Vektoren repräsentiert (›word embedding‹).[42] Für »Compasses« trainierte Parrish gleich zwei Modelle: Das eine, der »sounder-out«, wandelt Wörter in ihre Lautwerte, das andere, der »speller«, Lautwerte in geschriebene Sprache um.[43] Für eine Reihe semantisch ähnlicher Wörter als Eingabe gibt das System anschließend die plausible Schreibweise eines Vektors auf halbem Weg zwischen einem Wortpaar oder den Mittelpunkt aller Wörter aus (»woerth« ist dann das Mittel von »north« und »west«). Solche Reisen durch den Vektorraum haben nicht nur den Effekt, das gemeinhin diskret gedachte Zeichensystem der Sprache als stetig vorzustellen, sondern erlauben zudem, das Prinzip jener für Machine Learning so wichtigen ›word embeddings‹ intuitiv anschaulich zu machen. KI wird dabei weder als autonome Größe noch als gleichberechtigter Mitspieler, sondern lediglich als ein der Autorin unterstehendes Instrument verstanden – und in der Tat ist Parrishs Metapher für ihr Vorgehen ein »theremin for poetry production«.[44]

Mit dem »Nonsense Laboratory« hat Parrish diese Metapher auch tatsächlich in ein Interface überführt, das es Usern ermöglicht, phonetische Worteigenschaften einzustellen, als seien es die Regler auf einem Mischpult.[45]

Piringers Aufruf, die politischen Probleme KI-gestützter Sprachtechnologien in den Blick zu nehmen, ist aber auch bei Parrish nicht recht Genüge getan, auch wenn ihrer Entscheidung, auf proprietäre, kostenpflichtige Systeme wie GPT-3 zu verzichten,[46] politische Überlegungen zugrunde liegen. Ein politisches und ethisches Problem solcher großen Sprachmodelle besteht, neben ihrem enormen Energieverbrauch, unter anderem in der Schwierigkeit, die zu ihrem Training verwendeten Daten nachzuvollziehen. Weil sie meist ohne jede Kuratierung auf große Massen an Text trainiert werden, sind in ihnen auch sexistische oder rassistische ›biases‹ enthalten,[47] denn im Vektorraum werden Abhängigkeiten zwischen Begriffen, die im Korpus bloß latent waren, operabel und können Ideologeme des Urtextes wiederholen. Ein Beispiel sind Gendernormen in Maschinenübersetzungen, die ›nurse‹ stets als weiblich, ›doctor‹ stets als männlich verdeutschen.[48]

Diesem Phänomen widmet sich Li Zilles in »Machine, Unlearning«.[49] Zilles durchsuchte ein bereits trainiertes Sprachmodell nach hundert häufigen Substantiven und ließ zunächst mit ihrem Kontext korrelierte Adjektive, in einem zweiten Schritt als ähnlich bewertete Substantive und wiederum deren entsprechende Kontext-Adjektive ausgeben. Die so entstandenen Wortlisten wurden anschließend in Schablonen eingetragen, die an jene Stracheys erinnern: »Is [ORIGINAL_NOUN] [SAMPLED_ADJECTIVE_CONTEXT] like a [SAMPLED_NOUN] is [SAMPLED_ADJECTIVE_CONTEXT]?«[50] Der Text reicht von erwartbaren (»Can CONTROVERSY be produced in the same way reactions are produced?«) über absurde (»Should TELEVISION be distilled like cupcakes?«) bis zu offensichtlich ›bias‹ aufweisenden Äquivalenzen (»Will LIFE ever wane in the same way femininity wanes?«).[51] ›Unlearning‹ bezeichnet das aktive, therapeutische Verlernen veralteter oder toxischer Gewohnheiten, und entsprechend ist die Frageformulierung der Verszeilen als Aufgabe an die Leser*innen zu verstehen, die Richtigkeit der Vergleiche zu bewerten und sich nicht auf die Suggestionen des Sprachmodells zu verlassen. Ziel war, so Zilles, »(to) expose some of the ›assumptions‹ encoded by these embeddings as making more or less sense, (and to) lead us humans to lend a more critical eye to them«.[52]

Neue Standards

Das ›Co-Creative Writing‹ in Mensch-Maschine-Assemblagen kann viele Formen annehmen, und ich habe versucht, zwei gegenwärtige Tendenzen zu beschreiben. Diese waren freilich auf literarische Anwendungen von KI

beschränkt. Die größten Auswirkungen auf das *konventionelle* Leseverhalten werden dagegen kaum von den konnektionistischen Avantgarden zu erwarten sein, sondern von maschinell produzierten Gebrauchstexten jeder Art, von Wetterberichten über Informationsmaterial bis hin zu, sollte es technisch bald möglich sein, strukturell einfach reproduzierbarer Genreliteratur. Mit ihrer Zunahme stünde eine Verschiebung der Standardannahmen über unbekannte Texte zu erwarten. Bislang ist es noch so, dass Geschriebenes, das nicht explizit als maschinenproduziert markiert ist, als menschengemacht gelesen wird. Das verleitet dazu, hermeneutische Ambiguitäten vor dem Hintergrund eines vermeintlich gemeinsam geteilten Bedeutungshorizontes zu harmonisieren.[53] Diese Standardannahme menschlicher Autorschaft könnte sich mit der Verbreitung maschinell produzierter Texte in Richtung einer agnostischen Position verschieben, die deren Ursprung zumindest offenlässt.

Über die genauen Auswirkungen einer solchen Verschiebung kann man nur spekulieren. Zu bedenken wäre aber, dass mit ihr die Hoffnung auf eine starke, wirklich autonome künstlerische KI von selbst unterminiert würde: Wenn ich nicht mehr sicher sein kann, dass ein Roman *allein* von einem Menschen geschrieben ist, verlieren die an Menschen ausgerichteten zu simulierenden Merkmale wie Intention und Expression als implizite Kategorien seiner Betrachtung an Wert. Solange Menschen bereit sind, einen solchen Text als künstlerisches Objekt zu betrachten, ist es einerlei, wie er zustande gekommen ist – er hätte den Durkheim-Test gewissermaßen auch ohne »strong AI« bestanden. Die Folge wäre, dass nicht KI menschlicher würde, sondern sich Menschen und KI einander annäherten. Freilich ist hier immer die Alternative des Luddismus als Gegenbewegung möglich, die dann gerade menschlich-authentische Produktion hochhielte; an der Verschiebung der Leseerwartungen änderte das nichts, es würde sie eher negativ bestätigen. Die Alternative zwischen Verschmelzung mit den Maschinen und ihrer subalternen Instrumentalisierung hätte sich dann bereits durch die neue Standarderwartung an Texte entschieden. Wo die Differenz zwischen menschen- und maschinengeschrieben ihren Sinn verliert, gäbe es dann auch freilich keine ›digitale Literatur‹ mehr.

1 John Searle: »Minds, Brains, and Programs«, in: »Behavioral and Brain Sciences« 3 (1980), S. 417–424, hier S. 417. — **2** Melanie Mitchell: »Artificial Intelligence. A Guide for Thinking Humans«, London 2019, S. 40–42. — **3** Vgl. den Beitrag von Alexander Waszynski in diesem Band. — **4** Vgl. zur Einführung Ethem Alpaydin: »Machine Learning. The New AI«, Cambridge, Mass. 2016 und John D. Kelleher: »Deep Learning«, Cambridge, Mass. 2019. — **5** Eine Zwischenposition stellt der »Robojournalismus« dar, in dem Wetterbe-

richte, Erdbebenmeldungen oder Börsennachrichten automatisch verfasst werden, vgl. Anya Belz: »Fully Automated Journalism«, 2019, https://cris.brighton.ac.uk/ws/portalfiles/portal/8575767/Fully_Automatic_Journalism.pdf (20.5.2021). — **6** Vgl. meine Überlegungen zu »starker« und »schwacher künstlerischer Künstlicher Intelligenz« in Hannes Bajohr: »Keine Experimente. Über künstlerische Künstliche Intelligenz«, in: »Merkur« 75 (2021), S. 32–44. — **7** (Writer of Writer) Ross Goodwin: »1 the Road«, Paris 2018. Auf der französischen Bauchbinde heißt es: »Le premier livre écrit par une Intelligence Artificielle«. Goodwin wehrt sich allerdings gegen diese Bezeichnung, die vom Verlag stammt, vgl. Brian Merchant: »When an AI Goes Full Jack Kerouac«, in: »Atlantic«, 1.10.2018, https://www.theatlantic.com/technology/archive/2018/10/automated-on-the-road/571345 (20.5.2021). — **8** Vgl. Lucy Suchman: »Human-Machine Reconfigurations. Plans and Situated Actions«, Cambridge 2007; für einen an der Akteur-Netzwerk-Theorie ausgerichteten Ansatz, der spezifisch auf elektronische Literatur abzielt, vgl. Jörgen Schäfer: »Reassembling the Literary. Toward a Theoretical Framework for Literary Communication in Computer-Based Media«, in: Ders. / Peter Gendolla (Hg.): »Beyond the Screen. Transformations of Literary Structures, Interfaces and Genres«, Bielefeld 2010, S. 25–70. — **9** Stephanie Catani: »›Erzählmodus an‹. Literatur und Autorschaft im Zeitalter künstlicher Intelligenz«, in: »Jahrbuch der Deutschen Schillergesellschaft« 64 (2020), S. 287–310, hier S. 303. — **10** Susan Leigh Star: »The Structure of Ill-Structured Solutions«, in: Les Gasser / Michael N. Huhns (Hg.): »Distributed Artificial Intelligence«, London 1989, S. 37–54, hier S. 41. — **11** Christopher Strachey: »The ›Thinking‹ Machine«, in: »Encounter« (Okt. 1954), S. 25–31, hier S. 26. Vgl. hierzu Noah Wardrip-Fruin: »Digital Media Archaeology. Interpreting Computational Processes«, in: Erkki Huhtamo / Jussi Parikka (Hg.): »Media Archaeology. Approaches, Applications, and Implications«, Berkeley, Los Angeles, London 2019, S. 302–322. — **12** Vgl. Florian Cramer: »Exe.cut[up]able statements. Poetische Kalküle und Phantasmen des selbstausführenden Texts«, München 2011, Kap. 2–4; David Link: »Poesiemaschinen – Maschinenpoesie. Zur Frühgeschichte computerisierter Texterzeugung und generativer Systeme«, München 2007, S. 11. — **13** Vgl. Michael Wooldridge: »A Brief History of Artificial Intelligence. What It Is, Where We Are, and Where We Are Going«, New York 2020, S. 115–122. — **14** Vgl. dazu Hannes Bajohr: »Algorithmische Einfühlung. Über zwei Paradigmen digitaler generativer Literatur und die Notwendigkeit einer Kritik ästhetischer KI«, in: »Sprache im technischen Zeitalter« 59, 240 (2021) (jetzt auch in: Ders., »Schreibenlassen«). — **15** Scott Rettberg: »Electronic Literature«, London 2019, S. 53 f. — **16** Ebd., S. 20. — **17** Vgl. Alpaydin: »Machine Learning«, a. a. O., S. 20. Zwar verwenden auch KNNs noch gewisse Grundelemente, doch handelt es sich dabei meist nur um einzelne Buchstaben oder Buchstabenkombinationen (›byte-pair encoding‹); von einer Datenbank kann hier keine Rede sein. — **18** Für einen ersten Versuch dazu sowie eine vertiefende Darstellung des Vokabularproblems, vgl. Bajohr: »Algorithmische Einfühlung«, a. a. O. — **19** Der wichtigste Parameter bei der Textgenerierung ist die sogenannte Temperatur, die die (Un-)Wahrscheinlichkeit der Ausgabe reguliert, vgl. Allison Parrish: »Syntactic Crystals and the Slow Cooker. Editing ›Climate Change‹«, in: David (Jhave) Johnston (Hg.): »ReRites. Human + A. I. Poetry«, Montreal 2019, S. 163–168. — **20** Vgl. Tom B. Brown u. a..: »Language Models are Few-Shot Learners«, 2020, arxiv.org / abs / 2005.14165 (22.6.2021). Ein Beispiel für einen Prompt in diesem Paper ist sogar selbst literarisch: »Compose a poem in the style of Wallace Stevens«; der Output ist ein Gedicht, ebd., S. 49. — **21** Vgl. hierzu die Beiträge von Alexander Waszynski und Karl Wolfgang Flender in diesem Band, sowie Hannes Bajohr: »Das Reskilling der Literatur. Einleitung zu *Code und Konzept*«, in: Ders. (Hg.): »Code und Konzept. Literatur und das Digitale«, Berlin 2016, S. 7–21. — **22** Nick Montfort: »*Autopia* and *The Truelist.* Language Combined in Two Computer-Generated Books«, 2021, in: »Electronic Book Review«, DOI: 10.7273 / qzhw-wc29 (22.6.2021). Siehe auch seinen eigenen Beitrag in diesem Band, der offensiv das sequenzielle Paradigma bedient. — **23** Jörg Piringer: »elektrobarden«, in: »Transistor« 2, (2019), S. 78–83, hier S. 82 f. Von Piringer finden sich einige KNN-Gedichte in seinem Buch »datenpoesie«, Klagenfurt 2018; im Moment arbeitet er an einem vollständig KI-generierten Band. — **24** Vgl. den Aufsatz

von Jasmin Meerhoff in diesem Band; dort auch eine Diskussion der politischen Aspekte großer Sprachmodelle. — **25** Vgl. den Aufsatz von Kathrin Passig in diesem Band. — **26** Janelle Shane: »Delicious Recipe Titles Generated by Neural Network«, 2017, in: »AI Weirdness«, https://aiweirdness.com/post/159091493897/delicious-recipe-titles-generated-by-neural (20.5.2021). — **27** Sianne Ngai spricht von »cuteness« als einer »aesthetic of powerlessness« und einem »affective response to weakness«. Sianne Ngai: »Our Aesthetic Categories. Zany, Cute, Interesting«, Cambridge, Mass. 2012, S. 22, 24. Das lässt sich schön an den Illustrationen zu Shanes Buch ablesen, wo keine sinistren Black Boxes, sondern putzige und etwas vertrottelte Kistenwesen die KIs darstellen, vgl. Janelle Shane: »You look like a thing and I love you«, New York 2019. — **28** Vgl. für das Script: Ross Goodwin/»Benjamin«: »Sunspring«, 2016, https://www.docdroid.net/lCZ2fPA/sunspring-final-pdf; für den Kurzfilm: https://www.youtube.com/watch?v=LY7x2Ihqjmc (20.5.2021). — **29** Große Wirkung hatten dabei ein breit rezipierter Artikel des Tesla-KI-Forschers Andrej Karpathy: »The Unreasonable Effectiveness of Recurrent Neural Networks«. Andrej Karpathy Blog, 21.5.2015, https://karpathy.github.io/2015/05/21/rnn-effectiveness (20.5.2021). — **30** Simon Roloff: »Halluzinierende Systeme. Generierte Literatur als Textverarbeitung«, in: »Merkur« 864, S. 73–81, hier S. 79. — **31** Merzmensch: »Merz AI – Art(ificial) Kunst(liche) Intelligentsia. Issue #01«, in: »Perspektive« 102/103 (2020), S. 112–115. Weitere Ausgaben werden auf Twitter unter dem Hashtag #KImerzAI veröffentlicht. — **32** Merzmensch: »First Encounters with GPT-3«, in: »Merzazine«, 13.7.2020, https://medium.com/merzazine/gpt-3-first-encounters-676fcb8feac9 (20.5.2021). — **33** Roloff: »Halluzinierende Systeme«, a. a. O., S. 80. — **34** André Breton: »Erstes Manifest des Surrealismus«, in: Ders., »Manifeste des Surrealismus«, Reinbek 2009, S. 25. — **35** K Allado-McDowell: »Pharmako-AI«, London 2020, S. X. — **36** Mattis Kuhn: »Selbstgespräche mit einer KI«, Berlin 2021, S. 55. — **37** David Jhave Johnston: »ReRites«, a. a. O. Das Buch geht auf Vorarbeiten von 2015 zurück, die zu den ersten dezidiert literarischen KNN-Texten zählen müssen. Vgl. ders.: »Aesthetic Animism. Digital Poetry's Ontological Implications«, Cambridge, Mass. 2016, S. 127 ff. — **38** David (Jhave) Johnston: »Why A. I.?«, in: Ders.: »ReRites«, a. a. O., S. 171–177, hier S. 175, 171. — **39** https://www.anteism.com/shop/rerites-david-jhave-johnston (20.5.2021). — **40** Vgl. Nick Montfort: »Computational Writing as Living Art«, in: Johnston: »ReRites«, a. a. O., 139–143, hier S. 140. — **41** Johnston: »Why A. I.?«, a. a. O., S. 176. — **42** Vgl. Alexander Waszynskis Beitrag zu diesem Band für eine Diskussion von Parrishs »Articulations«, das ähnlich vorgeht. — **43** Vgl. den Auszug in diesem Band; zuerst als Allison Parrish: »Compasses«, in: »sync«, 27 (2019), sync.abue.io/issues/190705ap_ sync2_27_compasses.pdf (20.5.2021). Eine vermehrte Version findet sich unter dem Titel »Compass Poems« in: »BOMB« 154 (2021), S. 75–79. — **44** Allison Parrish: »Experimental Creative Writing with the Vectorized Word«, 2017, https://youtu.be/L3D0JEA1Jdc (20.5.2021). — **45** https://artsexperiments.withgoogle.com/nonsense-laboratory (20.5.2021). Diese Idee gemahnt an Michel Chaoulis hypothetischen »Literaturequalizer«, der Texteigenschaften regelbar machen soll. Vgl. Michel Chaouli: »Remix: Literatur. Ein Gedankenexperiment«, in: »Merkur« 9 (2009), S. 463–476. — **46** Vgl. »Q&A with Allison Parrish«, in: »Artists + Machine Intelligence«, 5.5.2020, medium.com/artists-and-machine-intelligence/q-a-with-allison-parrish-895a72727a4 (20.5.2021). — **47** Emily Bender u. a.: »On the Dangers of Stochastic Parrots. Can Language Models Be Too Big?«, in: »FAccT '21«, DOI: 10.1145/3442188.3445922 (22.6.2021). — **48** Das Phänomen sexistischer Übersetzungs-KI wird für Jörg Piringer zur Grundlage des Gedichts »tätigkeiten«, in: Ders.: »datenpoesie«, a. a. O., S. 62. — **49** Li Zilles: »Machine, Unlearning«, Denver 2018. — **50** E-Mail an den Autor, 6.5.2021. Die von Zilles verwendeten Wort- und Kontext-Vektoren finden sich hier: https://levyomer.wordpress.com/2014/04/25/dependency-based-word-embeddings, 2014, (20.5.2021). — **51** Zilles: »Machine, Unlearning«, a. a. O., S. 2, 26. — **52** E-Mail an den Autor, 6.5.2021. — **53** Vgl. Suchman: »Human-Machine Reconfigurations«, a. a. O., S. 48; dies wäre ein Fall eines auf KI-Text übertragenen »intentional stance«. Daniel Dennett: »The Intentional Stance«, Cambridge, Mass. 1987.