Soziologie, 53. Jg., Heft 1, 2024, S. 60–75

Probabilistische Wirklichkeitsmodelle und soziologische Intelligenz[1]

Richard Groß

1. ChatGPT als soziologisches Ereignis

Ende 2022 etwa entspann sich in der auf Twitter versammelten geistes- und sozialwissenschaftlichen Öffentlichkeit im Zusammenhang mit dem Release von *ChatGPT* – einem sogenannten *Large Language Model* (LLM) – eine unterhaltsame Debatte. Unter anderem ging es um die Frage, wie die universitäre Lehre darauf reagieren solle, dass eine Seminararbeit nun von einem algorithmischen Sprachmodell auf so zufriedenstellende Weise automatisch generiert werden könne, dass sich Studierende die eigene Arbeit an derartigen Prüfungsleistungen in Zukunft vermutlich sparen könnten (etwa Lepenies 2022) – und womöglich sollten, wie Slavoj Žižek pointiert kommentierte. Automatisch erzeugte Hausarbeiten würden dann bestenfalls ebenso automatisch bewertet, das ohnehin nicht selten skeptisch betrachtete System der Benotung in der akademischen Lehre wäre vollends erledigt, und in Seminaren hätten Dozierende und Studierende – von lästigen Pflichten entbunden – endlich die Gelegenheit zum allein auf das Wesentliche gerichteten Diskurs (Žižek 2023).

Schon kurz im Anschluss an diese Spekulationen zeigten öffentlich dokumentierte journalistische Selbstversuche, dass die Fähigkeiten von zeit-

1 Ich möchte mich bei Jakob Claus und Gereon Rahnfeld herzlich für ihr wertvolles Feedback zu Entwurfsfassungen dieses Aufsatzes bedanken. Ebenso danke ich Dirk Baecker für zahlreiche instruktive Anmerkungen im Zuge der redaktionellen Bearbeitung des Manuskripts. Jan Wetzels Initiative und Engagement verdanke ich schließlich, dass es zu diesem Text überhaupt gekommen ist.

genössischer Textgenerierungssoftware wie *ChatGPT* zwar nicht unterschätzt, aber vermutlich auch nicht überschätzt werden sollten. Die Texterzeugnisse des Sprachmodells schienen in etwa das Niveau mittelmäßiger Seminararbeiten zu erreichen (Cao 2023). Zudem erwies sich ein wesentlicher Kritikpunkt an den vorigen Releases der *GPT*-Reihe seit 2018 noch immer als stichhaltig: LLMs hätten die Tendenz zur ›Halluzination‹ vermeintlicher Fakten und würden also Tatsachen (Ereignisse, Publikationen, Autor:innen) erfinden, ohne dies kenntlich zu machen – und ohne zu einer solchen Kenntlichmachung prinzipiell überhaupt fähig zu sein.[2] Schon aus diesem Grund wirkt ihr akademisches Einsatzpotenzial letztlich überschaubar, zumindest jenseits von Plagiatsversuchen.

Dennoch möchte ich in diesem Beitrag ausgehend von ChatGPT eine soziologische Erörterung von LLMs versuchen und dabei sozialtheoretisch relevante und in mancherlei Hinsicht herausfordernde Aspekte einer Soziologie maschinellen Lernens im weiteren Sinne herausarbeiten. Denn womöglich – so meine Vermutung – ging es in den Diskussionen um mehr als nur Fragen zum Umgang mit Plagiaten. Dies könnte jedenfalls ein Blick in die jüngere Vergangenheit des Diskurses um KI nahelegen.

1996 sorgte das Duell zwischen dem amtierenden Schachweltmeister Garri Kasparow und IBMs Schachcomputer *Deep Blue* für eine mindestens vergleichbare mediale Resonanz wie das Release von *ChatGPT*. Als es schließlich kam, wie es kommen musste, und *Deep Blue* die erste Partie des Aufeinandertreffens gewann, war die Sensation vollbracht: Erstmals hatte ein Computer einen Großmeister des Denksports schlechthin geschlagen.[3] Dieses Ereignis nahm Jean Baudrillard zum Anlass für eine anthropologische Reflexion. Für ihn markierte der ›Triumph der KI über den Menschen‹ in erster Linie eine menschliche Selbstbegegnung, in der sich verschiedene Wünsche, Ängste und Begierden sublimieren (Baudrillard 2012). Die Beteiligten seien gefangen in ihrem merkwürdigen Bestreben danach, ein Double ihrer selbst zu erzeugen, das ihnen überlegen ist (vgl. Baudrillard 2002: 161). Ihnen entgehe dabei, dass sie sich so auf ein Spiel eingelassen haben, das sie nur verlieren könnten: »[I]f you become technically correct,

2 Auch für das wenige Monate später im März 2023 veröffentlichte Nachfolgemodell GPT-4 (beziehungsweise ChatGPT Plus als Bezeichnung des Interfaces) und diesem folgenden zukünftigen GPT-Iterationen ist nichts anderes anzunehmen, weil das Problem grundlegend die modellinterne Informationsverarbeitung betrifft (Bender 2023).

3 Für die dramaturgische Pointe war unerheblich, dass es sich nur um die erste Partie handelte und Kasparow letztlich mit 4:2 gewann. Bis zu einem ›echten‹ Wettkampfsieg unter Turnierbedingungen sollte es dann noch ein Jahr dauern.

you are unfailingly beaten by the machine«, schreibt Baudrillard(ebd.: 163). Man verliere unbedingt, obwohl und gerade weil man sich immer mehr der Maschine angleiche. Und damit werde zugleich auch das Spielen seines Wesens beraubt – und mithin die Spieler:innen ihres Sinns dafür, worum es im Spiel (»Play«) geht: einen Modus des Umgang mit den erwähnten Wünschen, Ängsten und Begierden, der im besten Fall sogar Spaß macht. Dass es im Spiel wesentlich auf Illusionen, Finten, falsche Fährten und Andeutungen ankommt, die herausfordern und verführen (ebd.: 162), heißt im Umkehrschluss, dass Maschinen nie wussten (und vermutlich niemals wissen oder gar ›erleben‹ werden), was *Spielen* wirklich heißt. Weder zu Freude noch Leiden fähig, blieben Maschinen nur – verzweifelt anmutende – Versuche, dies zu simulieren, etwa durch bewusst herbeigeführte (Rechen-)Fehler (ebd.: 164 f.).

Baudrillards Einlassungen zu *Deep Blue* verdeutlichen, in welchem Maße Verständnisse technischer Entwicklungen von deren öffentlichkeitswirksamer dramatischen Inszenierung[4] abhängen und Projektionen unterliegen, die – eben als Elemente (allzu) menschlicher Sozialität – Wünsche, Ängste, Fantasien und Hoffnungen in sublimierter Form mitkommunizieren. Dies gilt wohl insbesondere für sogenannte Künstliche Intelligenz (KI), die als variables *moving target* immer wieder neu und dabei zumeist durch den Vergleich mit als menschlich verstanden Fähigkeiten bestimmt worden ist (Muhle 2023, Fazi 2021). Es ist deshalb naheliegend, auch die seit 2022 anhaltende Diskussion um *ChatGPT* mitsamt solcher Interventionen wie der Žižeks unter diesen Vorzeichen zu lesen. Žižeks Text erinnert indes in seiner kontraintuitiven Pointe als diskursiver Querschläger wohl nicht ganz zufällig an Baudrillards Beitrag von vor 30 Jahren. Mit Baudrillard könnte auch jede soziologische Auseinandersetzung mit *ChatGPT*[5] daraufhin befragt werden, auf welches Spiel sie sich mit ihrer jeweiligen Problemperspektive einlässt und was diese womöglich übersieht. Dass *ChatGPT* Angst vor unsichtbaren Plagiatswellen ebenso wie Hoffnung auf Entlastung von ohnehin wenig nützlichen Pflichten zu evozieren scheint, sagt jedenfalls ebenso viel (wenn nicht mehr) über die gesellschaftliche Situiertheit der Autor:innen hinter den Positionen aus, wie es etwas über die tatsächlichen Fähigkeiten von LLMs mitteilt.

4 Vgl. Binder (2021) für eine Fallstudie des »Social Drama« um den Sieg des von Google entwickelten Go-Programms *Alpha Go* gegen den Weltklassespieler Lee Sedol im Jahre 2016, gewissermaßen die Neuauflage von Kasparow vs. Deep Blue.

5 Oder anderen Formen von maschinellem Lernen als gegenwärtig dominanter Form von KI.

Das Engagement von Soziolog:innen im geschilderten Fall hat vermutlich nicht zuletzt etwas mit ihrer Situiertheit im spezifischen Kontext der Debatte um Autor:innenschaft zu tun. Soziolog:innen stehen als Autor:innen mit dem eigenen Namen für die von ihnen verfassten Texte ein und sind ebenso verantwortlich für die Begutachtung von Texten anderer Autor:innen. Die Problematisierung von Autor:innenschaft durch *ChatGPT* kann daher als soziologisch brisante Angelegenheit verstanden werden, einerseits als Anlass zur Reflexion über das, was Soziolog:innen ihre Arbeit nennen: überwiegend als kognitive beziehungsweise ›geistig‹ verstandene Tätigkeiten. Andererseits stellt sie auch für die soziologische Theorie eine Herausforderung dar, indem sie etwa auf womöglich neue Weise Probleme der Zurechnung von Handlungen (und mithin Verantwortung) wie auch der Genese von Wissen und Kommunikation thematisch macht (ausführlicher Groß, Jordan 2023). In diesem Beitrag möchte ich eine Reaktion auf letzteren Aspekt versuchen. Mit Fokus auf LLMs wie *ChatGPT* werde ich im Rahmen einer sozialtheoretischen Exploration maschinellen Lernens (ML) den Wirklichkeitsbezug algorithmischer Modelle wie auch deren probabilistische Operationsweise und soziale Funktion diskutieren. Auf dieser Basis werde ich abschließend einen Vorschlag zur Charakterisierung der Beziehung von Soziologie und ML unterbreiten.

2. Über die automatische Erzeugung von Texten

Für die soziologische Annäherung an maschinelles Lernen bietet sich eine Reflexion der stochastischen Erfassung gesellschaftlicher Wirklichkeit an, schließlich meint ML meist die algorithmische Modellierung eines interessierenden Sachverhalts mittels eines Datensatzes, der diesen möglichst vollständig und präzise beschreiben soll. Algorithmische Mustererkennung entspricht zunächst nur einer quantitativen Datenanalyse, die in der Folge weitere Daten(-analysen) produzieren kann (Dourish 2016: 7). Eine solche Analyse bietet etwa die Errechnung einer Häufigkeitsverteilung als Beschreibung der Beziehungen verschiedener Datenpunkte zueinander, die dann als Modell für die Beurteilung weiterer Inputs genutzt wird. Im mittlerweile klassischen Beispiel von (sogenannten überwachten) Bilderkennungsverfahren etwa geht es um den Zusammenhang von Bildern als Kombinationen von Pixelverteilungen und Kategorien (Labels), die zur Klassifizierung der Bilder

vorhanden sind. Auf der Basis der erfolgten Analyse einer Bilddatenbank kann mit Hilfe des Modells errechnet werden, mit welcher Wahrscheinlichkeit ein Bild einer bestimmten für den Datensatz definierten Klasse angehört. Bei Festlegung eines Grenzwertes (etwa 90% Wahrscheinlichkeit laut Modell) kann diese Wahrscheinlichkeitsrechnung als ›Entscheidung‹ des Modells erscheinen: *Katze, Hund* oder aber auch *nicht erkannt*. Wobei dies schon terminologisch unsauber wäre. Schließlich *sieht* das Modell nichts, geschweige denn *erkennt* es etwas, jedenfalls nicht im konventionell gebrauchten, zumeist auf menschliche Wahrnehmung bezogenen Sinne. Stattdessen geht es in *Computer Vision* eben zentral um die Ausgabe von Wahrscheinlichkeitsangaben, beruhend auf der Analyse der Häufigkeitsverteilung von Bildpunkten.

Dies gilt auch dann, wenn für Nutzer:innen einer Bilderkennungsapp statt der Auflistung wahrscheinlicher Ergebnisse einer Analyse nur das Höchstwahrscheinliche aufgelistet wird, sodass es wirken kann, als wäre dies eindeutig *erkannt* worden. Eine solche Reduktion von Komplexität zur Herstellung von Eindeutigkeit wäre dann als »interface effect« (Galloway 2012), nicht aber als genuine (Fehl-)Leistung von ML zu verstehen. Zugleich verdeutlicht diese Lesart die Verstrickung von ML mit anderen Aspekten des Designs wie auch mit der praktischen Nutzung technischer Geräte. Interfaces – die zur Nutzung von Computern unverzichtbar sind – sind in ihrer konkreten Gestalt also zugleich Ausdruck sozialer Normen und technischer Datenverarbeitung, die ihrerseits sozialen Konventionen unterliegt.

ML-Verfahren können Beteiligten aus verschiedenen Gründen opak erscheinen (Burrell 2016). So kann es sein, dass fehlendes technisches Verständnis zu einem solchen Eindruck führt, doch ist es ebenso möglich – und gerade für gegenwärtig prägende ML-Verfahren typisch –, dass Umfang und Komplexität der Datenbasis (beziehungsweise *scale* des datenförmig erfassten Phänomens) sowie deren algorithmische Verarbeitung die Rekonstruktion der Genese von Outputs des Modells zum Zweck ihrer Erklärung nicht zulassen. Die Opazität von ML-Anwendungen kann jedoch auch durch bewusste (staatliche oder unternehmerische) Geheimhaltung zustande kommen und also gewissermaßen vertrautere nicht-technische Gründe haben. Für Beteiligte ist dabei oft nicht ersichtlich, um welchen dieser Gründe es sich handelt.

Es ist in den letzten Jahren bekannt geworden, dass ML-Anwendungen ihre soziale Wirkung häufig derart entfalten, dass sie pathologische Tendenzen bestehender gesellschaftlicher Realitäten reproduzieren, wenn nicht gar

verstärken. Und dies kann geschehen, ohne dass die für die Anwendungen Verantwortlichen davon notwendigerweise in Kenntnis sind, geschweige denn eine solche Wirkung beabsichtigt haben müssen. Zu solchen Tendenzen in den Anwendungseffekten von ML gehören die Diskriminierung marginalisierter sozialer Gruppen (Chun 2021), die Zunahme vernetzter datenbasierter Formen sozialer Kontrolle, die sich mit dem Bild des distribuierten »Polyopticon« (Sherman 2022) beschreiben lassen, oder die durch die Anwendungen verursachten Umweltbelastungen (Dhar 2020; Bender et al. 2021).

3. Stochastische Technik und maschinelle Sinnverschiebung

Die Beobachtung fehlgeleiteter Einsätze von ML offenbart ein dessen sozialen Charakter prägendes Missverständnis hinsichtlich seiner technischen Funktionsweise. Sofern man Technik soziologisch als »funktionierende Simplifikation« (Luhmann 1997: 524) versteht, zeichnet sie sich gerade dadurch aus, dass sie funktioniert. Man benutzt und bezeichnet etwas als Technik, wenn es bei gleicher Handhabung zuverlässig zum gleichen Ergebnis führt und damit Erwartungssicherheit in ihrer Nutzung bietet. Eine Enttäuschung dieser Erwartung deutet in der Regel daraufhin, dass die Technik kaputt und also genau in dem Moment keine Technik ist. Eine solche funktionale Minimalbestimmung macht leicht erkennbar, was als Technik infrage kommt – und was nicht. *Wie* jedoch eine wie auch immer geartete Maschine technisiert, kann im Fall von ML hingegen deutlich schwieriger zu erkennen sein. Dies gilt in besonderem Maße für ML-Anwendungen, in denen regelmäßig kausale Fixierung und stochastische Prädiktion verwechselt werden. *Computer Vision*-Verfahren ›erkennen‹ keine Bildinhalte, sondern errechnen Wahrscheinlichkeiten für mit den Bildpunktverteilungen korrelierenden Klassen, die als Inhalte definiert sind. Obwohl sich *Computer Vision* nie ›irrt‹, kann zugleich nicht davon ausgegangen werden, dass jeder relevante Bildinhalt mit Gewissheit korrekt identifiziert wird. Bilderkennung wird hier nicht kausal fixiert, sondern als stochastische Prädiktion technisiert. Zwischen Inputs und Outputs besteht keine Kausalbeziehung, sondern eine lediglich eine statistische Korrelation, die eine Kausalbeziehung abbilden *könnte*.

Dennoch wird ML häufig *als Kausaltechnik zu nutzen versucht*, das heißt genauer: als solche ausgewiesen und oberflächlich in entsprechende Form gebracht. Eine funktional adäquate technische Grundlage – mit Frieder Nake

(2008): die Unterfläche (»subface«) – kann, muss dafür aber nicht unbedingt gegeben sein. Mit Nakes Unterscheidung formuliert befinden sich Ober- und Unterfläche also in einem kontingenten Verhältnis, das durch das Interface (›Zwischenfläche‹) geregelt wird. Ein konventioneller Taschenrechner etwa ist auf Ebene seiner Unterfläche *hard-coded*, das heißt programmatisch vollständig determiniert, sodass ein bestimmter Input immer zum gleichen Output führen wird. Ein LLM hingegen wird für eine Rechenaufgabe als *prompt* ein statistisch plausibles, das heißt den Trainingsdaten ›nahes‹ Ergebnis ausgeben. Es operiert dabei probabilistisch: Aus dem Abgleich mit ähnlichen, im Datensatz vorhandenen Rechnungen samt Ergebnissen wird es ein auf die Ähnlichkeit zu den Trainingsdaten optimiertes Resultat liefern, dessen konkrete Form jedoch nutzer:innen- und zeitpunktabhängig variieren kann. Ein Effekt, der aus der Google-Suche bekannt ist, von dem auf der Ebene der Nutzer:innenoberfläche jedoch nichts vernehmbar wäre. Um zu erkennen, dass ein solches Vorgehen in diesem Fall nur bedingt erfolgversprechend ist – aber prinzipiell sein könnte –, reichen wohl selbst überschaubare Kenntnisse der Mathematik aus. Wenn ML sich in diesem Sinne gerade durch seine Differenz zu Kausaltechnik auszeichnet, wie kann diese Differenz dann als Kennzeichen von stochastischer Technik positiv bestimmt werden?

Ein Indiz für die Beantwortung dieser Frage liefert Blick auf sogenannte *generative* Anwendungen von ML, bei denen das Funktionsprinzip gewissermaßen in invertierter Form erweitert wird. Nach der Modellierung eines Datensatzes (›Training‹) bringt das Verfahren auf dieser Datenbasis selbst einen (›potentiellen‹) Output hervor, der daraufhin analysiert wird, mit welcher Wahrscheinlichkeit er dem ursprünglichen Trainingsdatensatz hätte angehören können. Abhängig von dieser Klassifikation kann dieser dann zu einem tatsächlichen Output werden – oder sich aber aufgrund der fehlenden Passung als unnütz erweisen. Diese Unterscheidung von *generator*- und *discriminator*-Funktionen beschreibt konkret das Verfahren sogenannter *generative adversarial nets* (GANs; Goodfellow et al. 2014), kann aber auch als Heuristik dienen für das Verständnis neuerer Ansätze, zu denen die bereits diskutierten LLMs zählen.

Auf diese möchte ich zurückkommen, da sich an ihnen prinzipielle Mechanismen der generativen Funktionen von ML gut veranschaulichen lassen. Die Funktionsweise von LLMs lässt sich treffend als »spicy autocomplete« (Solomon 2023) bezeichnen. Mit *autocomplete* ist zunächst bekanntermaßen eine automatische Vorhersage potenziell folgender Textteile bezeichnet. Auf

Basis der Analyse eines Textkorpus kann von einer beliebigen Stelle eines Textes ausgehend errechnet werden, was das nächste Wort, der nächste Satz, der nächste Absatz oder gar das folgende Kapitel des Textes möglicherweise sein könnten – gemessen am Vergleich mit dem im Korpus vorhandenen Textmaterial. Dass auf »Ich« nicht selten »bin« folgt, mag in diesem Zusammenhang trivial anmuten. Wie der aktuelle Absatz dieses Textes enden wird, ist womöglich weniger leicht vorherzusagen, nicht einmal für den Autor des Textes im Moment des Verfassens dieser Zeile. *Spicy* scheint es dann allemal zu werden, wenn ein Umschlagen ins Nicht-Triviale geschieht, das heißt im Sinne des Kybernetikers Heinz von Foerster der Output einer Maschine nicht allein durch einen gegebenen Input, sondern gleichermaßen durch unbekannte interne Zustände determiniert werden, die die Input-Output-Beziehung konstituieren (von Foerster 1984: 10). Liest man Antworten als Fortsetzungen von Fragen und Dialoge in diesem Sinne als kontinuierliche Texte, ist leicht ersichtlich, dass jeder Text als plausibel fortschreibbar behandelt werden kann.

Ein Hinweis auf die soziale Funktion von ML lässt sich in den unterschiedlichen Weisen der Verwendung von LLMs und anderen Anwendungen generativer Rechenarchitekturen finden. *ChatGPT* und verwandte Sprachmodelle können zweifelsohne dazu benutzt werden, ›klassisch‹ technische Funktionen zu erfüllen, die bereits ›pre-ML‹ von Computern ausgeführt wurden. Mathematische Rechenaufgaben zu lösen, Listen von Namen alphabetisch zu ordnen oder auf (lokale oder über Netzwerke verfügbare) Datenbanken zuzugreifen, um Informationen über Objekte, Personen oder Ereignisse abzurufen – die Liste könnte beliebig verlängert werden. Doch gilt grundlegend für jede dieser Anwendungen – wie trivial sie auch anmuten mögen –, das Prinzip von *spicy autocomplete*: Ein LLM wird für gestellte Aufgabe (*prompt*) ein statistisch plausibles Ergebnis liefern, operiert dabei allerdings probabilistisch und ist eben nicht *hard-coded* wie ein ›klassischer‹ Taschenrechner. Aus diesem Grund können generell keine Garantien für die Richtigkeit der Outputs angeboten werden; das Ergebnis für einen bestimmten *prompt* hängt stattdessen von der Verfügbarkeit ›passender‹ (bestenfalls kombiniert mit der Unverfügbarkeit nur ›fast passender‹) Trainingsdaten des Modells ab, auf deren Basis algorithmisch ein wahrscheinliches Ergebnis errechnet wird. Die epistemische Autorität des Verfahrens beruht nicht auf begründbaren Regeln, sondern der Generalisierung von Beispielen (Campolo, Schwerzmann 2023), wobei diese Aufgabe im Zuge ihr Technisierung an die algorithmische Modellierung »delegiert« wird (Jaton 2021: 278).

Ein statistisch plausibles Ergebnis könnte, muss aber nie ›korrekt‹ sein. Ob ein Output gemessen am *prompt* ›korrekt‹ (passend, stimmig) ist, kann – typisch für sinnförmige Outputs – prinzipiell in verschiedenen Hinsichten beurteilt werden. Die Begutachtung eines Textes ›aus der Feder‹ eines LLM kann etwa Syntax, Semantik, Ästhetik oder andere Kategorien zur Beurteilung von Texten betreffen, wobei das, was explizit ›in den Zeilen‹ steht (Syntax), verständlicherweise leichter statistisch zu verarbeiten ist als das, was mitunter nur ›zwischen den Zeilen‹ zu finden wäre (Semantik). Alternativ könnte man diese Differenz als jene von Form und Inhalt reformulieren (Lopez 2023): Ein LLM-generiertes Textstück wird aufgrund seiner formalen Ähnlichkeit zu einer Vielzahl von Referenzobjekten in den Trainingsdaten vermutlich selten unmittelbar völlig aus dem Raster fallen und mag auf den ersten Eindruck häufig Plausibilität suggerieren können, wenngleich immer zuerst eine exakte Prüfung Gewissheit verschaffen kann. Hannes Bajohr beschreibt diesen Charakter von LLM-Outputs als Produktion von »dummer Bedeutung« (Bajohr 2022). Bedeutsam auf den ersten, dumm auf den zweiten Blick – so könnte vermutlich nicht selten der Lektüreeindruck von »artifizieller Semantik« (ebd.) als Produkt von »content generation« (Bull 2023) beschrieben werden.

Im Falle von Abweichungen gegenüber konventionell anerkannten und leicht überprüfbaren Wahrheiten in LLM-Outputs ist häufig von »Halluzinationen« die Rede. Eine Tendenz zu Halluzinationen mag eine Vielzahl von Anwendungen für den Einsatz von LLMs ausschließen, etwa in wissenschaftlichen Zusammenhängen. Allerdings wäre bis hierhin auch nur die Hälfte der sozialen Funktion von LLMs erschlossen. Die Fähigkeit zur automatischen Generierung von sinnförmigen Outputs, die zuallererst an Halluzinationen erinnern, meint schließlich zugleich, dass die Modelle immerhin dazu imstande sind, Unvorhergesehenes zu erzeugen und daher – in welcher Hinsicht auch immer – überraschende Outputs zu produzieren. Und Überraschungen können – bekanntermaßen abhängig vom Zusammenhang ihres Auftretens – als Störung oder aber als Bereicherung empfunden werden. Theoretisch gewendet meint dies einen Bruch mit kausaler Fixierung als primärem technischen Prinzip. LLMs können stattdessen *potenziell* originelle Beiträge zu Prozessen leisten, die vorher nicht technisierbar schienen. Nützlich erweisen sie sich dabei letztlich vielleicht weniger, indem sie ganze Texte formulieren, sondern vielleicht eher, wenn sie etwa Schreibprozesse durch hilfreiche Stichworte und Verbesserungsvorschläge unterstützen. Es zeigt sich hier allgemeiner ein Potenzial für den Einsatz von »uncertain tools« (Wilk 2022) in kreativen

Prozessen, deren explizites Ziel ohnehin unvorhergesehene Ergebnisse sind. ›Halluzinierende Maschinen‹ können sich in dieser Hinsicht als nicht nur nützlich, sondern potenziell als inspirierend erweisen. Die soziale Funktion von ML als Technik ist dann weniger in der automatischen Verrichtung mechanischer Vorgänge als vielmehr in der »funktionierenden Simplifikation« (Luhmann 1997: 524) kreativer Prozesse auszumachen. Grundlage dafür ist die probabilistische Modellierung von Sinnphänomenen, die im Verhältnis von Input (etwa *prompts*) und Outputs bewirkt, was man »technologische Sinnverschiebung« nennen könnte (Hörl 2013).

Diese Verknüpfung von Stochastik und sinnförmigem sozialen Geschehen macht zugleich den Kern von ML als epistemologischem Problem aus, da selten antizipiert oder gar verstanden werden kann, *wie genau* sich der Sinn im sozialen Geschehen verschiebt. Gleichermaßen führt diese Bedingung zu Problemen in der praktischen Anwendung von ML (empirisch: Groß, Wagenknecht 2023), da – sozialtheoretisch gefasst – die Einbindung von derartigen nicht-menschlichen Entitäten in soziale (Ko-)Operationen damit freilich noch unwägbarer wird als ohnehin schon. »Künstliche Kommunikation« (Esposito 2022) unter Beteiligung von ML hat sozusagen ihre ganz eigenen Fallstricke, wenn die Erwartungen und Beiträge einer anderen, an Kommunikation beteiligten Partei weder als hinreichend symmetrisch mit dem eigenen und noch überhaupt als hinreichend verstehbar vorausgesetzt werden können.

Dieser Nexus von Sinnverstehen und Stochastik bildet interessanterweise zugleich das Fundament der ›modernen‹ Kommunikationstheorie nach Claude Shannon (1948). In dessen Beschäftigung als Ingenieur des Kommunikationsunternehmens AT&T war die Frage nach den Möglichkeitsbedingungen dafür, dass die Wahrscheinlichkeit für ein ›gutes‹ Verhältnis von Signal und Rauschen in einer Telefonleitung steigt.[6] Diese technik- und wissenschaftshistorisch folgenreiche neue Perspektive auf Kommunikation setzt voraus, dass das Kommunizierte gerade nicht sinnförmig, sondern technisch-statistisch als quantitativ bestimmbares elektrisches Signal vorgestellt wird (Suchman 2021). Die mit der ML-basierten Errechnung wahrscheinlicher Sinngehalte einhergehenden Wider- und Unsinnigkeiten lassen die Spannung der Dualität von Kommunikation als qualitativ (Sinn) wie quantitativ (Wahrscheinlichkeit) beschreibbarer Vorgang in diesem Sinne auf neue Weise erfahrbar werden.

6 Wobei die Unterscheidung von Signal und Rauschen sich hier nicht auf semantische Register bezieht, sondern auf die akustische Qualität einer Verbindung; Schwerverständliches bleibt auch dann schwerverständlich, wenn es laut und deutlich vernehmbar ist.

Zu dem bereits beschriebenen Spannungsverhältnis zwischen Wahrscheinlichkeitsangabe über das Eintreten eines Ereignisses und dem tatsächlichen Ereignis gesellt sich in realweltlicher sozialer Praxis eine weitere Komplikationsachse. Diese entsteht aus dem Umstand, dass prinzipiell keine Gewissheit darüber bestehen kann, wie faktisch stichhaltig die errechnete Wahrscheinlichkeitsangabe ist, da sie von einem Modell eines Wirklichkeitsphänomens errechnet wurde, dessen Beziehung zum eigentlichen Phänomen prinzipiell kontingent ist. Es gibt keine Gewähr, dass die errechnete Wahrscheinlichkeit der *tatsächlichen* Wahrscheinlichkeit entspricht. Das Modell könnte derart konstruiert sein, dass es Rückschlüsse über die Wirklichkeit nicht zulässt, ohne dass modellintern Indizien dafür vorliegen. Diese zweite, externe Wahrscheinlichkeit kann logisch indes nicht in der ersten, modellinternen enthalten sein. Grundsätzlich ist der Zusammenhang zwischen Modell und modellierter Wirklichkeit als kontingent zu verstehen. Wichtige Indizien für die integrierte Validität des Modells in seiner praktischen Anwendung offenbaren sich womöglich nur rückblickend und hängen von einer unüberschaubaren Anzahl von Kriterien für die Validität der Bezüge zwischen Wirklichkeit, Modell und dessen Anwendung ab. Und schließlich entscheidet sich der praktische Umgang mit ML darüber hinaus ohnehin erst in konkreten Nutzungssituationen. Mithin zeigt sich erst dann, wie und wozu ML genutzt wird. Im Designprozess intendierte Nutzungsformen von technischen Objekten entsprechen selten völlig und manchmal überhaupt nicht den tatsächlichen der Praxis – ein mittlerweile klassischer Lokus der Techniksoziologie und der STS (Akrich 1992).

4. Kontingenzmaschinen oder: What does mean mean?

Der operativen Logik von ML – etwa gefasst als »computational reason« (Cavia 2022) – soziologisch auf die Spur zu kommen, erweist sich methodologisch womöglich als unlösbare Aufgabe. Von den erwähnten Unwägbarkeiten des sozialen Einsatzes von ML abgesehen gibt es für komplexe algorithmische Architekturen wie *Generative Pre-Trained Transformers* (GPT) keine unmittelbare Möglichkeit der eindeutigen Verknüpfung von ursprünglichem

Input (Trainingsdaten) und Output als Kausalverhältnis.[7] Anschauliche Belege dafür liefert etwa Hito Steyerl in einem Mitte 2023 erschienenen Aufsatz mit dem Titel »Mean Images« (Steyerl 2023). Ausgehend von der pointierten Formulierung des US-amerikanischen Schriftstellers Ted Chiang, *ChatGPT* sei im Kern ein »blurry jpeg of all the text in the web« (Chiang 2023), nahm sie sich für ihren Essay das Bildgenerierungsmodell *Stable Diffusion* vor. Dank des von den Künstler:innen Mat Dryhurst und Holly Herndon entwickelten Online-Tools »Have I been trained?« wusste Steyerl, dass Fotos von ihr im Datensatz *LAION-5B* enthalten sind, auf den das Modell trainiert wurde. Diese präsentiert sie in ihrem Beitrag neben einem Bild, das vom Modell auf den *prompt* »Image of Hito Steyerl« hin generiert wurde. In diesem lässt sich ein zweifellos an Hito Steyerl erinnerndes Gesicht ausmachen, doch handelt es sich ebenso zweifellos um keine realitätsgetreue Darstellung der Künstlerin, auch wenn der fotografische Stil des Bildes dies suggeriert, denn aus Gründen, über die sich nur spekulieren lässt, sieht die abgebildete Person in der *Stable Diffusion*-Kreation gegenüber der realen Erscheinung Steyerls (und ebenso den Fotos im Datensatz) stark gealtert aus.

Es lässt sich dieser Anekdote entnehmen, dass ML-Modelle durchaus *mehr* leisten, als die Inhalte der ihnen zugrundeliegenden Datensätze präzise zu repräsentieren, wenn auch dies angesichts umfangreicher vieldimensionaler Datensätze in vielen Zusammenhängen schon eine nützliche Errungenschaft wäre. Dieses *Mehr* verweist indes auf einen Faktor, der ML-Modellen Eigenschaften verleiht, die ihren Charakter nicht-trivial und im Hinblick auf Outputs unbestimmt machen. Steyerl formuliert zur Beschreibung dieses Umstandes in ihrem Essay eine treffende Zuspitzung mit der Frage: »What does mean mean?« Die Differenzen zwischen den beschriebenen Bildern deutet auf die inkommensurable Dualität von Qualität (Sinn) und Quantität (Durchschnitt) hin – wobei sich diese Differenz im Begriff des Durchschnitts, der seinerseits qualitativ wie quantitativ verstanden werden kann, nochmal wiederholt.

Eine Antwort auf Steyerls Frage wäre im konkreten Fall der *Stable Diffusion*-Kreation von »Image of Hito Steyerl« instruktiv, ist allerdings nicht zu erlangen: Obwohl sowohl die algorithmische Architektur des Modells wie auch die Inhalte des Trainingsdatensatzes öffentlich bekannt sind, hilft dies

7 Weshalb Bemühungen um »Explainable AI« unmittelbar einleuchtend sind, wobei die Tragweite der ›Erklärung‹ eines Modells allerdings häufig einerseits eher beschränkt ist und andererseits nicht für alle aktiv oder passiv Beteiligten einer ML-Anwendung gleichermaßen aufschlussreich ist (Zednik 2021).

wenig in der konkreten Aufklärung der Frage. Es lassen sich nicht einzelne Bilder oder Sequenzen des Algorithmus ausmachen, die zur Erklärung von Details des generierten Bildes dienen könnten. Steyerl schlägt zur Beschreibung solcher opaker technischer Arrangements[8] das Konzept der »White Box« vor. Im Unterschied zur hinlänglich bekannten *black box* – für die von Foersters »non-trivial machine« (1984: 10) eine alternative Bezeichnung ist – seien deren »internal states« bekannt, ohne dass diese Kenntnis allerdings instruktive Einsichten zum Verständnis ihrer Outputs liefern könnte. Die Transparenz der Maschine erweist sich als Illusion, wenn sich zeigt, dass sich mit dem vermeintlichen Wissen nichts erklären lässt. Man hatte es, entgegen täuschenden Suggestionen, letztlich doch die ganze Zeit mit einer, wenn nicht mehreren *black boxes* zu tun: »Inside every white box there are two black boxes trying to get out« (Glanville 1982). Gescheiterte Trivialisierungsversuche (»making them white«) können – im Bild des Kybernetikers Ranulph Glanville – immerhin zur Reflexion der eigenen Perspektive führen, weil sie im Moment des Scheiterns als *eigene* Versuche sichtbar werden.

Bildmodelle wie *Stable Diffusion sehen* nicht, sondern errechnen wahrscheinliche Pixelverteilungen; LLMs wie LaMDA *verstehen* nicht, sondern errechnen wahrscheinliche Wortverteilungen. Es zeigt sich im Beispiel Steyerls ein Hinweis auf die Inkommensurabilität zwischen der sprachlichen und der statistischen Beschreibung und Repräsentation von Wirklichkeit: Die mathematische Modellierung eines Datensatzes, der alle möglichen Wirklichkeitsaspekte abdecken soll, mag Unsicherheiten technisieren und rationalisieren. Sie kann die Unbestimmtheit der Welt jedoch nicht zum Verschwinden bringen, sondern deren Berechenbarkeit – etwa in Form maschineller Prädiktionen – nur suggerieren, sodass Unbestimmtheit dann als (bestimmbare) Kontingenz eines Beobachters erscheint und diesem zugerechnet werden kann (Esposito 2014: 234). Dass zahlenmäßigen Beschreibungen der Wirklichkeit mitunter eine wirklichkeitsstiftende Suggestionskraft ›Objektivität‹ anzuhaften scheint, kann angesichts des Umstandes verwundern, dass der eine Zahl bestimmende Verweisungshorizont selbst unbestimmt ist, weshalb die Zahl als Maß für (Un-) Ordnung geeignet scheint, nicht aber per se für Ordnung an sich steht (Lehmann 2014: 41). Steyerls Frage danach, was es mit Durchschnitten auf sich hat, gewinnt ihre Prägnanz indes gerade dadurch, dass sie sich auf Bilder

8 Die ihrerseits allerdings auf Arbeiter:innen oft brutal ausbeutende »ghost work« (Gray, Suri 2019) angewiesen sind und in dieser Hinsicht als Akkumulation menschlich verrichteter Arbeit verstanden werden können; ein Umstand, den Steyerl in ihrem Essay wie auch in ihren künstlerischen Arbeiten der letzten Jahre immer wieder explizit herausstellt

bezieht, denen ihre Kontingenz nur anmerken kann, wer ihre Geschichte kennt. Wer mit ML in Kontakt kommt, muss dabei nicht (mehr) mit Zahlen in Kontakt kommen. Stattdessen hat man es mit bestimmt anmutenden Bildern oder Texten zu tun. Die damit verbundene Suggestion scheint mithin die von Bestimmbarkeit schlechthin zu sein: »from error to optimism« (Mackenzie 2015: 436ff.). Maschinelles Lernen bezeichnet in diesem Sinne die technische Transformation von Unbestimmtheit in kontingente Formen mittels stochastischer Wirklichkeitsmodellierung. Was man soziologische Intelligenz nennen würde, scheint mir demgegenüber gerade auf die reflexive Sichtbarmachung der Kontingenz in all den vielfältigen Bestimmungsversuchen einer Welt abzuzielen, die wir uns mit maschinellem Lernen teilen.

Literatur

Akrich, Madeleine 1992: The De-Scription of Technical Objects. In Wiebe E. Bijker / John Law (eds.), Shaping Technology / Building Society: Studies in Sociotechnical Change. Cambridge: MIT Press, 205–224.

Bajohr, Hannes 2022: Dumme Bedeutung. Künstliche Intelligenz und artifizielle Semantik. Merkur, 76. Jg., Heft 882, 69–79.

Baudrillard, Jean 2002: Deep Blue or the Computer's Melancholia. In Jean Baudrillard, Screened Out. London, New York: Verso, 160–165.

Baudrillard, Jean 2012: Beyond Artificial Intelligence: Radicality of Thought. In Jean Baudrillard, Impossible Exchange. London, New York: Verso, 145–158.

Bender, Emily 2023: Thought experiment in the National Library of Thailand. https://medium.com/@emilymenonbender/thought-experiment-in-the-national-library-of-thailand-f2bf761a8a83, letzter Aufruf am 1. November 2023.

Bender, Emily M. / Gebru, Timnit / McMillan-Major, Angelina / Shmitchell, Shmargaret 2021: On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, doi: 10.1145/3442188.3445922.

Binder, Werner 2021: Alpha Go's Deep Play. Technological Breakthrough as Social Drama. In Jonathan Roberge / Michael Castelle (eds.), The Cultural Life of Machine Learning. An Incursion into Critical AI Studies. Cham: Palgrave Macmillan, 167–195.

Bull, Sarah 2023: Content Generation in the Age of Mechanical Reproduction. Book History, vol. 26, no. 2, 324–361. doi: 10.1353/bh.2023.a910951.

Burrell, Jenna 2016: How the machine ›thinks‹: Understanding opacity in machine learning algorithms. Big Data & Society, vol. 3, no. 1, doi: 10.1177/2053951715622512.

Campolo, Alexander / Schwerzmann, Katia 2023: From rules to examples: Machine learning's type of authority. Big Data & Society, vol. 10, no. 2, doi: 10.1177/20539517231188725.

Cao, Sissi 2023: Can ChatGPT Essays Get You Into an Ivy League School? We Asked an ex-Harvard Advisor. Observer, 5th May 2023. https://observer.com/2023/05/chatgpt-generated-college-admissions-essay-ask-an-expert/, letzter Aufruf am 1. November 2023.

Cavia, AA 2022: Logiciel: Six Seminars on Computational Reason. Berlin: &&& Publishing.

Chiang, Ted 2023: ChatGPT Is a Blurry JPEG of the Web. The New Yorker, 9th February 2023. https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web, letzter Aufruf am 1. November 2023.

Chun, Wendy Hui Kyong 2021: Discriminating Data: Correlation, Neighborhoods, and the New Politics of Recognition. Boston: MIT Press.

Dhar, Payal 2020: The carbon impact of artificial intelligence. Nature Machine Intelligence, no. 2, 423–425. doi: 10.1038/s42256-020-0219-9.

Dourish, Paul 2016: Algorithms and their others: Algorithmic culture in context. Big Data & Society, vol. 3, no. 2, 1–12, doi: 10.1177/2053951716665128.

Esposito, Elena 2014: Algorithmische Kontingenz. Der Umgang mit Unsicherheit im Web. In Alberto Cevolini (Hg.), Die Ordnung des Kontingenten. Beiträge zur zahlenmäßigen Selbstbeschreibung der modernen Gesellschaft. Wiesbaden: Springer VS, 233–249.

Esposito, Elena 2022: Artificial Communication. How Algorithms Produce Social Intelligence. Cambridge: MIT Press.

Fazi, M. Beatrice 2021: Introduction: Algorithmic Thought. Theory, Culture & Society, vol. 38. no. 7/8, 5–11. doi: 10.1177/02632764211054122.

Galloway, Alexander R. 2012: The Interface Effect. Cambridge: MIT Press.

Glanville, Ranulph 1982: Inside every white box there are two black boxes trying to get out. Behavioral Science, vol. 12, no. 1, 1–11.

Goodfellow, Ian J. / Pouget-Abadie, Jean / Mirza, Mehdi / Xu, Bing / Warde-Farley, David / Ozair, Sherjil / Courville, Aaro / Bengio, Yoshua 2014: Generative adversarial nets. Proceedings of the 27th International Conference on Neural Information Processing Systems. doi: 10.5555/2969033.2969125.

Gray, Mary L./ Suri, Siddharth 2019: Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass. Boston: Houghton Mifflin Harcourt.

Groß, Richard / Jordan, Rita 2023: KI-Realitäten / AI Realities. In Richard Groß / Rita Jordan (Hg), KI-Realitäten: Modelle, Praktiken und Topologien maschinellen Lernens. Bielefeld: transcript, 9–33.

Groß, Richard/ Wagenknecht, Susann 2023: Situating machine learning. On the calibration of problems in practice. Distinktion. Journal of Social Theory. vol. 24, no. 2, 315–337.

Hörl, Erich 2011: Die technologische Bedingung. Zur Einführung. In Erich Hörl (Hg.), Die technologische Bedingung. Beiträge zur Beschreibung der technischen Welt. Frankfurt am Main: Suhrkamp, 7–53.

Jaton, Florian 2021: The Constitution of Algorithms. Ground Truthing, Programming, Formulating. Cambridge: MIT Press.

Lehmann, Maren 2014: Komplexe Ereignisse und kontingente Mengen. Anmerkungen zur Soziologie der Zahl. In Alberto Cevolini (Hg.), Die Ordnung des Kontingenten. Beiträge zur zahlenmäßigen Selbstbeschreibung der modernen Gesellschaft. Wiesbaden: Springer VS, 41–62.

Lepenies, Robert 2022: Twitter-Thread vom 7. Dezember 2022. https://twitter.com/RobertLepenies/status/1600611200408100876, letzter Aufruf am 1. November 2023.

Lopez, Paola 2023: ChatGPT und der Unterschied zwischen Form und Inhalt. Merkur, 77. Jg., Heft 891, 15–27.

Luhmann, Niklas 1997: Die Gesellschaft der Gesellschaft. 2 Bände. Frankfurt am Main: Suhrkamp.

Mackenzie, Adrian 2015: The production of prediction. Was does machine learning want? European Journal of Cultural Studies, vol. 18, no. 4-5, 429–445.

Muhle, Florian 2023: (Vor-)Geschichte der Künstliche-Intelligenz-Forschung und der sozialen Robotik. In Florian Muhle (Hg.), Soziale Robotik: Eine sozialwissenschaftliche Einführung. Berlin, Boston: De Gruyter Oldenbourg, 13–29.

Nake, Frieder 2008: Surface, Interface, Subface: Three Cases of Interaction and One Concept. In Uwe Seifert / Jin Hyun Kim / Anthony Moore (eds.), Paradoxes of Interactivity: Perspectives for Media Theory, Human-Computer Interaction, and Artistic Investigations. Bielefeld: transcript, 92–109.

Shannon, Claude E. 1948: A Mathematical Theory of Communication. The Bell System Technical Journal, vol. 27, 379–423, 623–656.

Sherman, Stephanie 2022: The Polyopticon: A diagram for urban artificial intelligences. AI & Society, vol. 38, no. 3, 1209–1222, doi: 10.1007/s00146-022-01501-3.

Solomon, Mike 2023: GPT3 is Just Spicy Autocomplete. https://thecleverest.com/gpt3-is-just-spicy-autocomplete/, letzter Aufruf am 1. November 2023.

Steyerl, Hito 2023: Mean Images. New Left Review, no. 140/141. https://newleftreview.org/issues/ii140/articles/hito-steyerl-mean-images, letzter Aufruf am 1. November 2023.

Suchman, Lucy 2021: Talk with Machines, Redux. Interface Critique Journal, no. 3, 69–80.

von Foerster, Heinz 1984: Principles of Self-Organization in a Socio-Managerial Context. In Hans Ulrich / Gilbert J. B. Probst (eds.), Self-Organization and Management of Social Systems. Berlin: Springer, 2–24.

Wilk, Elvia 2020: Against Prediction: Designing Uncertain Tools. Noema Magazine, 24th September 2020. https://www.noemamag.com/against-prediction-designing-uncertain-tools/, letzter Aufruf am 1. November 2023.