Zeitschrift für Medienwissenschaft, Jg. 17, Heft 32 (1/2025), https://doi.org/10.14361/zfmw-2025-170106. Published by transcript Verlag.

BENEDIKT MERKLE

KONTROLLIERBARE KOGNITION

Stochastische Elemente in symbolischen und subsymbolischen Systemen

Seit einigen Jahren treten Nutzer*innen des Internets zunehmend Interfaces entgegen, die auf der Technologie künstlicher neuronaler Netze (KNN) aufbauen.[1] Viel Faszination geht von diesen Systemen aus und große Unternehmen tätigen Investitionen, um sie allgegenwärtig zu machen. Dabei fasziniert nicht nur die automatische Generierung digitaler Texte und Bilder, sondern auch die diesen Phänomenen zugrunde liegende Hardware-Technologie. Als Schwellenfigur einer Erfolgsgeschichte neuester Technologie wird oftmals auf den Kognitionspsychologen Geoffrey Hinton verwiesen, dessen Software-System für Bilderkennung sich beim ImageNet-Wettbewerb 2012 als deutlich leistungsfähiger als die Konkurrenz erwies. Den Unterschied machte dabei der Einsatz parallel organisierter Recheneinheiten von Grafikprozessoren.[2] Erzählt wird auf diese Weise die Erfolgsgeschichte einer Hardware-Entwicklung, die den parallel operierenden Strukturen der KNNs angemessen ist.[3] Für Andreas Sudmann weicht diese Entwicklung generell vom Paradigma digitaler Rechenmaschinen ab: «Neuronale Netzwerke, ob künstlich oder natürlich, stellen […] ein Gegenmodell zur Funktionsweise digitaler Computer gemäß der seriell organisierten Von-Neumann-Architektur dar.»[4] Parallele Prozessierung wird gegenüber serieller mit Metaphern der Vertiefung als ‹subsymbolisch› und als *deep learning* charakterisiert.

Dementgegen möchte ich in diesem Artikel argumentieren, dass eine Kontinuität zwischen digitaler Hardware und gegenwärtiger Phänomene der KI besteht, und möchte dafür zwei Gründe anführen: Zum einen entspringt jenes soeben angedeutete Narrativ einer Hardware-Revolution dem Selbstverständnis der eigenen historischen Entwicklung des herrschenden, konnektionistischen Paradigmas der KI-Forschung. Subsymbolische KI konturiert sich in Abgrenzung zum Ansatz der Good Old Fashioned AI (GOFAI).[5] Diese ging von der These aus, dass Prozesse der Kognition sich nicht nur als symbolische Prozesse simulieren lassen, sondern darüber hinaus selbst symbolisch sind. Nicht-determiniertes, intelligentes Verhalten, etwa das strategische Vorgehen

1 Die generelle Funktionsweise eines KNN wurde in der Medienwissenschaft in den vergangenen Jahren auf unterschiedlichen Niveaus der Zugänglichkeit erläutert, vgl. etwa Hannes Bajohr: Dumme Bedeutung. Künstliche Intelligenz und artifizielle Semantik, in: *Merkur. Deutsche Zeitschrift für europäisches Denken*, Jg. 76, Nr. 882, 2022, 69–79.

2 Vgl. Christoph Engemann, Andreas Sudmann: Einleitung, in: dies. (Hg.): *Machine Learning – Medien, Infrastrukturen und Technologien der Künstlichen Intelligenz*, Bielefeld 2018, 9–36, hier 23.

3 Vgl. für eine kritische Erweiterung dieser Erzählung Rainer Mühlhoff: Human-aided artificial intelligence: Or, how to run large computations in human brains? Toward a media sociology of machine learning, in: *New Media & Society*, Bd. 22, Nr. 10, 2020, 1868–1884, doi.org/10.1177/1461444819885334.

4 Andreas Sudmann: Szenarien des Postdigitalen: Deep Learning als MedienRevolution, in: Engemann, Sudmann (Hg.): Machine Learning, 55–73, hier 66.

5 Vgl. John Haugeland: *Artificial Intelligence: The Very Idea*, Cambridge (MA) 1985, 112–117.

beim Spielen eines Spiels wie Schach, sollte aus einem Verbund symbolisch vollständig determinierter Systeme hervorgehen. Demgegenüber konturierten sich Forschungen des konnektionistischen Paradigmas, indem sie intelligentes, maschinelles Verhalten als emergenten Effekt aus der Verkettung einer großen Menge binärer Zustände darstellten. Das Narrativ der Dominanz des konnektionistischen Paradigmas bildet einen strategischen Teil technologischer Entwicklung, insofern es dazu beiträgt, einer bestimmten Technologie das Monopol bei der Entwicklung künstlicher Intelligenz (KI) zuzuschreiben.[6] Es bedarf daher erneut der Historisierung. Zum anderen, so argumentiert Lucy Suchman in einem aktuellen Beitrag zur historischen Lokalisation der KNN-Technologie, trifft subsymbolische KI mit ihrem Gegenstück, einem «rule-based symbolic approach», in einem beide Ansätze umfassenden Paradigma des Kognitivismus zusammen. Diese reduktionistische Theorie mechanisierbaren, intelligenten Verhaltens basiert auf der Annahme einer «correspondence between mental representation formed in the brain/mind and a world taken to stand outside of it».[7]

Den Befund Suchmans aufgreifend möchte ich im Folgenden argumentieren, dass symbolischen und subsymbolischen Systemen zwar unterschiedliche technische Prozesse zugrunde liegen, beide aber in dem Wunsch nach einem unmöglichen Objekt eines maschinellen Modells der Kognition übereinstimmen. Das repräsentationale Modell des Kognitivismus bildet eine Fiktion, die von unterschiedlichen technologischen Implementierungen gestützt werden kann. Zur historischen Einordnung der fachinternen Unterscheidung von symbolischen und subsymbolischen Systemen kann auf die in der Medienwissenschaft bereits mehrfach beschriebene Verbindung der aktuellen Technologie mit dem Forschungsfeld der Kybernetik verwiesen werden.[8] Das Paradigma des Kognitivismus entsteht Mitte des 20. Jahrhunderts in engem Austausch mit der groß angelegten, interdisziplinären Integrationsleistung der Kybernetik.[9] Die Kybernetik konstruiert intelligentes maschinelles Verhalten und Kognition als adaptives System, als «etwas, das seine innere Organisation, seine innere Struktur schwankenden Außenbedingungen anpassen kann. Der Homöostat demonstriert, wie sich ein eingangs chaotisches System mittels Rückkopplung und Zufallsgenerator zu einem Servomechanismus entwickeln kann.»[10] Punkte, an denen ein System auf seine Umgebung hin geöffnet wird und ein stochastisches Element in seine internen Abläufe integriert, finden sich, wie ich zeigen werde, sowohl bei älteren Maschinen der Anfangsphase kybernetischer Forschung als auch in der aktuellen Forschung zu maschinellem Lernen mit KNNs. Über die Unterscheidung in symbolische und subsymbolische Systeme hinausgehend stelle ich im ersten Teil des Artikels zwei maschinelle Systeme zur Generierung intelligenten Verhaltens vor und gehe auf Punkte ein, an denen die Systeme strategisch ein stochastisches Element integrieren. Diese Stellen markieren den willkürlichen Eingriff von Entwickler*innen in einen Prozess, der gerne als Emergenz von Symbolen aus statistischen Werten eines

6 Vgl. Ranjodh Singh Dhaliwal, Théo Lepage-Richer, Lucy Suchman: Introduction. Rendering the Neural Network, in: dies. (Hg.): *Neural Networks*, Lüneburg 2024, 1–19, hier 6.

7 Lucy Suchman: The Neural Network at Its Limits, in: Dhaliwal, Lepage-Richer, Suchman (Hg.): *Neural Networks*, 87–112, hier 88.

8 Vgl. Clemens Apprich: Die Maschine auf der Couch. Oder: Was ist schon ‹künstlich› an Künstlicher Intelligenz?, in: *Zeitschrift für Medienwissenschaft*, Jg. 11, Nr. 21 (2/2019): *Künstliche Intelligenzen*, 20–28, doi.org/10.25969/mediarep/12617; Andrea Knaut: Können Künstliche Neuronale Netzwerke denken?, in: Theo Hug, Günther Pallaver (Hg.): *Talk with the Bots. Gesprächsroboter und Social Bots im Diskurs*, Innsbruck 2018, 73–86; Sudmann: Szenarien des Postdigitalen; Hannes Bajohr: Die ‹Gestalt› der KI. Jenseits von Holismus und Atomismus, in: *Zeitschrift für Medienwissenschaft*, Jg. 12, Nr. 23 (2/2020): *Zirkulation*, 168–183, doi.org/10.25969/mediarep/14824.

9 Vgl. Jean-Pierre Dupuy: *On the Origins of Cognitive Science: The Mechanization of the Mind*, Cambridge (MA) 2009. Vgl. zur Kybernetik als interdisziplinäres Großprojekt Claus Pias: Zeit der Kybernetik – Eine Einstimmung, in: ders. (Hg.): *Cybernetics – Kybernetik. The Macy-Conferences 1946–1953*, Bd. 2: *Essays und Dokumente*, Zürich, Berlin 2004, 9–41.

10 Christina Vagt: Nietzsche, Ashby und die logische Fiktion künstlicher Intelligenz, in: Renate Reschke, Knut Ebeling (Hg.): *Nietzsche, die Medien und die Künste im Zeitalter der Digitalisierung*, Berlin 2023, 185–200, hier 193.

Datensatzes, als Black Box vorgestellt und in die Fiktion einer kontrollier- und modellierbaren Kognition integriert wird. Daran anschließend lässt sich die Kategorie des Subsymbolischen innerhalb der Entwicklung und Kommunikation des konnektionistischen Paradigmas lokalisieren.

«A Mind-Reading (?) Machine»

Zwischen der Definition und Konstruktion mechanischer Rechenprozesse und der Struktur mentaler Vorgänge besteht seit langer Zeit eine enge Verbindung. Bereits Alan Turing sah sich dem Vorwurf des Kognitivismus, also der Reduktion mentaler Prozesse auf mechanisierbare Verfahren durch Kurt Gödel ausgesetzt.[11] In den 1950er Jahren boten kybernetische Maschinen disziplinübergreifend Anlass dazu, über die maschinellen Qualitäten von intelligentem Verhalten nachzudenken. Zentral steht dabei die Frage, wie weit intelligentes Verhalten mit logischen Strukturen modellierbar ist. Der Psychoanalytiker Jacques Lacan machte kybernetische Maschinen zum Demonstrationsmodell der symbolischen Ordnung, in die Subjekte sich eintragen. In seinem Seminar II von 1954/55 («Das Ich in der Theorie Freuds und in der Technik der Psychoanalyse») bekamen sie die Funktion, die Bildung eines Unbewussten zu repräsentieren.[12] Das Unbewusste nach Sigmund Freud und Lacan bildet sich als Gedächtnis einer symbolischen Ordnung der Sprache, die Subjekte umgibt. Freud stieß bei der Behandlung neurotischer Personen auf eine unabschließbare Suche nach einem unerreichbaren Zustand, die er als Wiederholungszwang bezeichnete.[13] Unter Verweis auf kybernetische Maschinen deutete Lacan nach Freud den Wiederholungszwang nicht entlang der organizistischen Vorstellung einer konservativen Natur der Triebe, sondern, in Henning Schmidgens Formulierung, als «Determinierung des Subjekts durch das Symbolische. Dessen besondere Eigenschaft sei es, wie von selbst zur Wiederholung zu drängen.»[14] Die automatische Insistenz von Signifikanten beschrieb Lacan als maschinelle Qualität im Subjekt:

> [W]enn das Unbewusste im Freud'schen Sinne existiert, […] [ist es] nicht undenkbar […], dass eine moderne Rechenmaschine, indem sie den Satz freilegt, der, ohne dass es das weiß und auf lange Sicht, die Wahlen eines Subjekts moduliert, es über jedes gewohnte Maß hinaus schafft, beim Gerade-Ungerade-Spiel zu gewinnen.[15]

Die von Lacan angesprochene «moderne Rechenmaschine» blieb eine leere Referenz in den Seminaren. Lacan kündigte zwar an, dass er mit einer konkreten Implementierung einer solchen Maschine experimentieren werde, erwähnte sie dann aber nicht wieder.[16] Dem zugrunde lag mit großer Sicherheit, so rekonstruierte es Mai Wegener, ein Text Claude Shannons, auf den Lacan durch den befreundeten Mathematiker Jacques Riguet aufmerksam gemacht wurde.[17] Dieser Text erschien unter dem Titel «A Mind-Reading (?) Machine» als Memo der Bell Laboratories vom 18. März 1953 und beschreibt die Konstruktion

11 Vgl. Brian Jack Copeland, Oron Shagrir: Turing versus Gödel on Computability and the Mind, in: dies., Carl J. Posy (Hg.): *Computability: Turing, Gödel, Church, and Beyond*, Cambridge (MA) 2013, 1–34.

12 Vgl. Henning Schmidgen: *Das Unbewußte der Maschinen. Konzeptionen des Psychischen bei Guattari, Deleuze und Lacan*, München 1997, hier 113–115.

13 Vgl. Sigmund Freud: Jenseits des Lustprinzips [1920], in: ders.: *Studienausgabe*, Bd. 3: *Psychologie des Unbewußten*, hg. v. Alexander Mitscherlich, Frankfurt/M. 1981, 213–272, hier 229.

14 Schmidgen: *Das Unbewußte der Maschinen*, 102.

15 Jacques Lacan: Das Seminar über ‹Der gestohlene Brief›, in: ders.: *Schriften. Bd. 1*, Wien, Berlin 2016, 12–76, hier 71.

16 Vgl. ders.: *Das Seminar von Jacques Lacan. Buch II (1954–1955): Das Ich in der Theorie Freuds und in der Technik der Psychoanalyse*, hg. v. Norbert Haas, Freiburg, Olten, 1980, 227 f.

17 Vgl. Mai Wegener: *Neuronen und Neurosen. Der psychische Apparat bei Freud und Lacan. Ein historisch-theoretischer Versuch zu Freuds Entwurf von 1895*, München 2004, 73. Dass Theoretiker*innen des französischen Strukturalismus sich für neueste informationstechnische und kybernetische Entwicklungen interessieren, ist in dieser Zeit verbreitet, vgl. Lydia He Liu: *The Freudian Robot: Digital Media and the Future of the Unconscious*, Chicago 2010, 139, die auf die Verbindung zwischen Shannons A Mind-Reading (?) Machine (vgl. nachfolgende Fußnote) und Lacans *Theorie der Bildung des Unbewussten* eingeht.

einer Maschine nach dem Prinzip einer Markov-Kette, die das Gerade-Ungerade-Spiel gewinnt. Shannon führte das Spiel als bekannten Topos ein: «This game has been discussed from the game theoretic angle by von Neumann and Morgenstern, and from the psychological point of view by Edgar Allen [sic] Poe in ‹The Purloined Letter.›»[18] Lacan wird ein Jahr später sein Seminar über Poes «Der entwendete Brief» halten, ohne dabei auf Shannons Konstruktionsplan zu verweisen.

Das Gerade-Ungerade-Spiel ist ein einfaches Nullsummenspiel und eignet sich, um Mensch und Maschine einander gegenüberzustellen. Dazu werden zwei Spielenden zunächst die Positionen ‹Gerade› und ‹Ungerade› zugewiesen. Daraufhin zeigen beide auf Kommando eine Anzahl von Fingern oder andere Symbole, die eine Zahl repräsentieren. Beide Zahlen werden zusammengerechnet, das Ergebnis ist eine gerade oder ungerade Zahl und es gewinnt der*die Spielende mit der entsprechenden Position. Auf diese Weise ist ein klassisches Nullsummenspiel nach John von Neumann realisiert, bei dem die Anzahl der Gewinne und Verluste zusammengenommen null ergibt. Shannon implementiert das Spiel als digitalen Schaltkreis: Es stehen zwei Zustände zur Auswahl (entsprechend des verbauten Schaltes «either ‹right› or ‹left›»[19]). Die Maschine gewinnt, wenn ihr Output mit der Selektion des menschlichen Gegenübers übereinstimmt. Die Gewinnstrategie der Shannon'schen Maschine baut darauf, dass es einem Menschen schwerfällt, auf lange Sicht eine willkürlich alternierende binäre Symbolfolge zu erzeugen.[20] Stattdessen bilden sich automatisch Muster, auf deren Beobachtung Shannons Apparatur programmiert ist. Der Automat beobachtet dabei das Auftreten von zwei aufeinanderfolgenden Spielzügen des menschlichen Gegenübers in Mustern wie: «The player wins, plays the same, and loses.»[21] Acht verschiedene solcher Muster lassen sich formulieren und werden von Shannon als logische Schaltung implementiert. Über die Dauer des Spiels verfolgt Shannons Apparatur auf diese Weise das Spiel des menschlichen Gegenübers und antizipiert das Auftreten von Mustern. In den Schaltkreisen bildet sich nach einigen Spielrunden ein Gedächtnis der Wahrscheinlichkeit aus, wonach das Gegenüber nach bestimmten Abfolgen von Gewinn und Verlust auf ein zuvor gespieltes Muster zurückfällt.[22] Darüber hinaus integriert die Maschine ein Zufallselement, um den generativen Prozess der Mustererkennung zu starten: «The machine also contains a random element. Until patterns have been found, or if an assumed pattern is not repeated at least twice by the player, the machine chooses its move at random.»[23]

Durch das Zufallselement wird das stochastische Prinzip der Markov-Kette in einen automatischen maschinellen Lernprozess integriert. Diesem Element kommt zum einen die Funktion zu, den stochastischen Prozess zu starten. Zum anderen übernimmt es während des Spiels überall dort den maschinellen Prozess der Auswahl des nächsten Elements, wo die Mustererkennung unterbrochen werden muss, da die Antizipation enttäuscht wurde. Das Zufallselement

18 Claude E. Shannon: A Mind-Reading (?) Machine (Bell Laboratories Memorandum, March 18, 1953), in: Neil J. A. Sloane, Aaron D. Wyner (Hg.): *Claude E. Shannon: Collected Papers*, New York 1993, 688–690, hier 688.

19 Ebd.

20 Vgl. Axel Roch: *Claude E. Shannon. Spielzeug, Leben und die geheime Geschichte seiner Theorie der Information*, Berlin 2010, 21.

21 Shannon: A Mind-Reading (?) Machine, 688.

22 Das stochastische Prinzip, bei dem das jeweils nächste Element einer Folge von Symbolen ausgehend von einer zuvor bestimmten Anzahl von vorausgehenden Symbolen bestimmt wird, ist in der Mathematik als Markov-Kette bekannt.

23 Shannon: A Mind-Reading (?) Machine, 688.

wird dabei von Shannon auf beinahe ironische Weise integriert: Es entsteht durch die unbewusste Mithilfe menschlicher Spieler*innen. Ein im Schaltplan (Abb. 1) unten rechts als «Motor» beschriftetes Modul realisiert ein ständig rotierendes Element, das angehalten wird, wenn Spielende den Knopf zur Entscheidung von ‹Gerade› oder ‹Ungerade› drücken. Da die Zeitspanne zwischen zwei aufeinanderfolgenden Spielzügen nicht von den internen Abläufen der Maschine vorherbestimmt ist, wird auf diese Weise ein Element des Zufalls integriert. Ironischerweise sind es so die Spielenden selbst, die die Maschine mit dem gewinnbringenden stochastischen Element füttern, das aus der Öffnung des Systems auf die externe Dauer des Spielgeschehens entsteht. Menschliche Gegenspieler*innen arbeiten auf diese Weise an ihrer eigenen Niederlage gegen eine Maschine mit, die ihre Aufmerksamkeitsspanne strategisch unterläuft, wie Philipp von Hilgers es zusammenfasst:

> Dem Spieler ist aber sehr wohl gestattet, der Maschine zu ihrem Zufall zu verhelfen. Denn jedesmal, wenn er einen Zug macht, wird an einem Zufallselement ein momentaner Wert abgenommen, der alle 1/10 Sekunde alterniert. Seit Emil Du Bois-Reymond gilt das Diktum, daß Reize für ihre Perzeption eben eine 1/10 Sekunde benötigen. Medien hausen bekanntlich gerne hinter solchen Zeitfenstern, in die keine menschliche Kontrolle hineinreicht.[24]

Die Integration der Willkür in vollständig determinierte Prozesse erzeugt den Zauber maschinellen intelligenten Verhaltens als adaptiven Lernprozess.[25] Diesen Befund möchte ich nun auf den Bereich großer Sprachmodelle übertragen.

Stochastische Elemente in großen Sprachmodellen

Shannons Maschine kombinierte einen stochastischen Prozess mit einem Zufallselement, um ein automatisches maschinelles Lernsystem zu erhalten, das über den Ablauf eines Spiels seine Spielstrategien entsprechend den Zügen seines menschlichen Gegenübers anpassen konnte. Auf analoge Weise möchte ich Aspekte der Konstruktion eines Chat-Interfaces auf Grundlage eines *generative pretrained transformer* (GPT) beschreiben. Hier kommt nicht der Prozess der Markov-Kette zum Einsatz, der das Folgeelement einer Serie aus einer bestimmten Anzahl vorausgehender Elemente errechnet. Stattdessen wird die Technologie neuronaler Netze implementiert, wobei ein zuvor trainiertes Netz die Errechnung der Wahrscheinlichkeit von Folgeelementen in einer Serie von Symbolen übernimmt. Dennoch sind die Systeme an dem Punkt vergleichbar, an dem sie Elemente der Willkür in einen generativen Prozess des maschinellen Lernens integrieren, was ich im Folgenden an zwei Stellen aufzeigen werde.

Dabei ist zunächst wichtig, trainierte neuronale Netze als deterministische Strukturen zu verstehen, die auf Grundlage einer komplexen Matrixmultiplikation eine Relation zwischen Input und Output herstellen.[26] Ihre Komplexität entsteht aus der großen Anzahl von Verbindungen zwischen Knoten, den

24 Philipp von Hilgers: Eine Maschine, die Gedanken liest, in: Sigrid Schade, Georg Christoph Tholen: *Konfigurationen zwischen Absturz und Wirklichkeit. Materialien zur Tagung ‹Konfigurationen zwischen Kunst und Medien›*, Kassel, 4.–7.9.1997 (CD-ROM), in: dies.: *Konfigurationen zwischen Kunst und Medien*, München 1999.

25 Vgl. unter Verweis auf William Ashbys Homöostat Vagt: Nietzsche, Ashby und die logische Fiktion künstlicher Intelligenz, 193.

26 Eine Einführung in die mathematischen Hintergründe der Technologie gibt die Online-Publikation von Michael Nielsen: *Neural Networks and Deep Learning*, 2015, neuralnetworksanddeeplearning.com (9.12.2024). Fabian Offert und Ranjodh Singh Dhaliwal weisen darauf hin, dass probabilistische Systeme aus dem Kontext fortlaufender Testsituationen entstehen, deren Algorithmen nicht stabil sind, und fordern daher eine angemessene «probabilistic form of critique», vgl. Fabian Offert, Ranjodh Singh Dhaliwal: The Method of Critical AI Studies, A Propaedeutic, *arXiv*, 28.11.2024, 1–10, hier 3, doi.org/10.48550/arXiv.2411.18833. Die hier besprochenen algorithmischen Techniken neuronaler Netze möchte ich daher nicht als Teil eines bestehenden stacks konnektionistischer Systeme verstehen, sondern als aktuell genutzte algorithmische Techniken älteren Datums, die ein stochastisches Element in ein deterministisches System zu integrieren suchen.

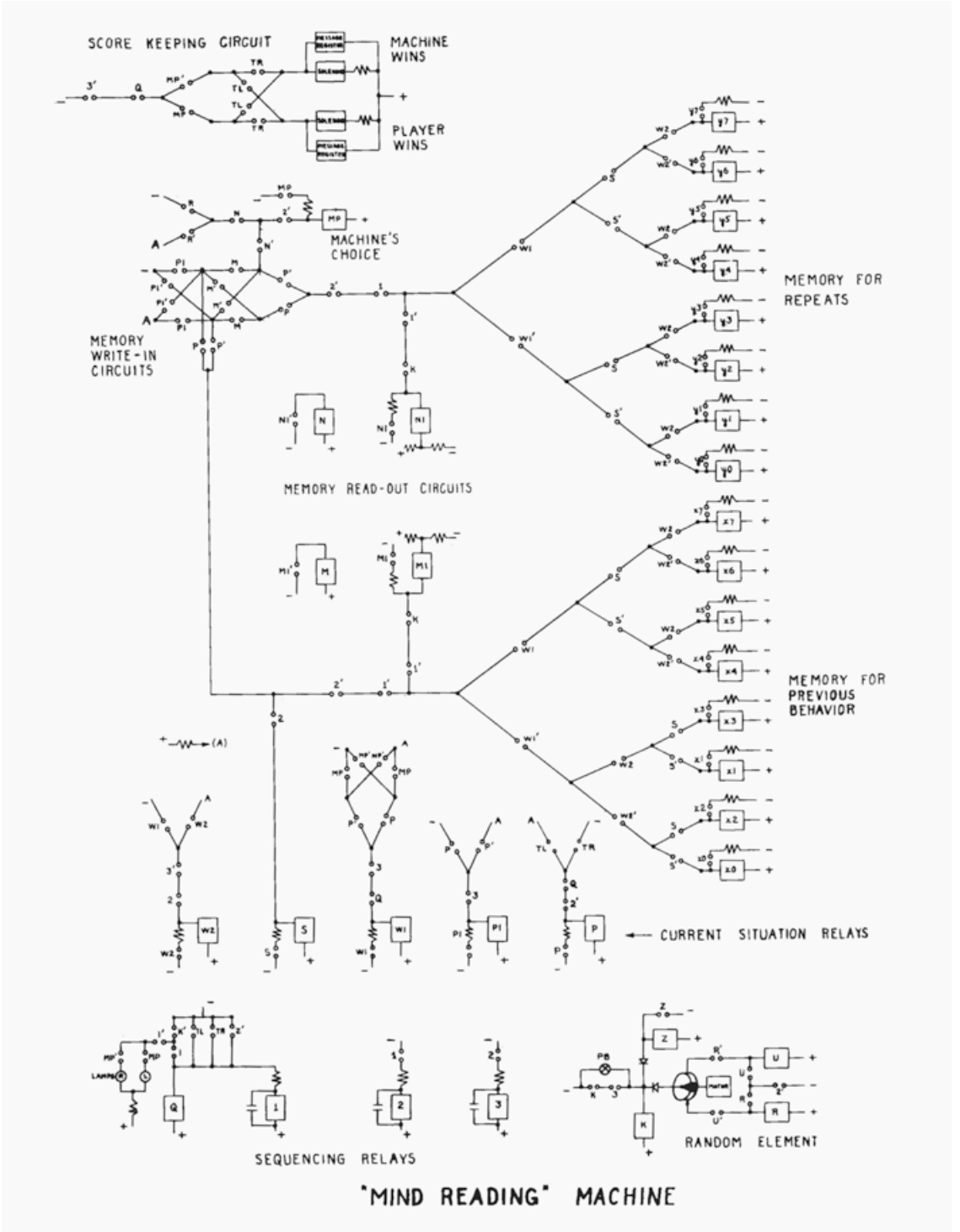


Abb. 1 Shannons Schaltplan für die *Mind-Reading (?) Machine*, 1953

Neuronen des Netzwerks. Dem Operieren mit einem trainierten Modell geht ein stochastischer Prozess des Trainings voran. In ihm werden die Relationen zwischen Elementen eines Datensatzes als Parameter der Neuronen des Netzwerks, sogenannte Gewichte einer Polynomfunktion, angenähert. Der Trainingsprozess führt zur sukzessiven Verbesserung der Gewichtparameter, wobei die generierten Outputs des Modells anhand korrekter Vorgaben aus annotierten Datensätzen korrigiert werden.[27] Ist das Training abgeschlossen, kann das entstandene Modell zur Errechnung von Übergangswahrscheinlichkeiten genutzt werden. Es funktioniert nun deterministisch: Gleiche Eingaben führen stets zu gleichen Ergebnissen.

Doch welchen Wert sollen die Gewichtparameter sämtlicher Neuronen ganz zu Beginn des Trainingsprozesses haben? David Rumelhart, Geoffrey Hinton und Ronald Williams halten 1986 in ihrem einflussreichen Artikel zum *backpropagation*-Algorithmus für das Training neuronaler Netze fest, dass der Prozess mit zufällig gewählten Gewichten initialisiert werden muss.[28] Ein symmetrischer Anfangszustand, bei dem allen Gewichten der gleiche Wert zugewiesen wird, führt dagegen zu keinem Lernprozess. Das Modell hat aus sich heraus keine Möglichkeit, die Symmetrie zu brechen. Der anfängliche Unterschied, der im Verlauf des Trainingsprozesses einen Unterschied macht, muss zu Beginn von außen zugegeben werden, indem der Prozess mit zufälligen Gewichten initialisiert wird. Unter der Bedingung, dass kein symmetrischer Anfangszustand gewählt wird, kann mit der Wahl der Parameter zu Beginn des Trainingsprozesses experimentiert werden, denn sie hat Einfluss auf die Errechnung der Gewichtparameter. Diese willkürliche Initialisierung der Gewichte in der Trainingsphase des Modells ist die erste Instanz eines Zufallselements innerhalb der Technologie, die ich hervorheben möchte. Ein zweites Element findet sich ganz am Ende des Prozesses der Anwendung eines bereits trainierten Modells.

Sprachmodelle nutzen, genau wie alle trainierten neuronalen Netze, ein vordefiniertes Vokabular. Der gesamte Wortschatz einer Sprache wird dazu in kleine Elemente, sogenannte Tokens unterteilt. Für OpenAIs ChatGPT sind das 100.261 Elemente.[29] Zur Berechnung eines Folgesymbols aus dieser Liste wird dem Modell als Input ein Text übergeben, der von dem trainierten Transformer auf seine internen Eigenschaften analysiert wird und für sämtliche 100.261 möglichen Folgesymbole einen Wahrscheinlichkeitswert ausgibt. Die allermeisten dieser Wahrscheinlichkeiten werden Werte im negativen Bereich sein und es wird einige Kandidaten geben, die hohe positive Wahrscheinlichkeiten erhalten. Experimente mit Modellen der Sprachgenerierung zeigten, dass die Selektion des Begriffs mit der jeweils höchsten Wahrscheinlichkeit das Modell schnell in einen infiniten Regress führt, in dem die wahrscheinlichste Verknüpfung von Worten endlos wiederholt wird. Um einen generativen Prozess der Textsynthese zu erhalten, ist es erneut nötig, ein Zufallselement in das Output-Layer zu integrieren.[30] Dies geschieht im letzten Schritt, kurz vor der Ausgabe des Outputs durch

27 Eine sogenannte Kostenfunktion errechnet die Differenz des Systems zur erwünschten korrekten Zuordnung. Diese Differenz soll so klein wie möglich werden, wozu beim Training der *backpropagation*-Algorithmus zum Einsatz kommt, der die hochdimensionalen Matrixvektoren zu lokalen Minima führt.

28 Vgl. David E. Rumelhart, Geoffrey Hinton, Ronald J. Williams: Learning representations by back-propagating errors, in: *Nature*, Bd. 323, Oktober 1986, 533–536, hier 535: «To break symmetry we start with small random weights».

29 Vgl. für eine vollständige Liste der Tokens Emmanuel Maggiori: Here's ChatGPT's entire vocabulary (with its +100k tokens), *Emmanuel Maggiori* [Website], o. D., *emaggiori.com/chatgpt-all-tokens/* (9.12.2024).

30 Zugänglich erklärt von *Hugging Face*-Entwickler Patrick von Platen: How to generate text: Using Different Decoding Methods for Language Generation with Transformers, *Hugging Face*, 1.5.2020, *huggingface.co/blog/how-to-generate* (9.12.2024).

einen stochastischen Samplingprozess. Ein Sprachmodell hat an dieser Stelle eine Liste von möglichen folgenden Tokens mit unterschiedlichen Wahrscheinlichkeitswerten errechnet. Diese Werte müssen nun in ihrer relativen Größe zueinander evaluiert werden, wozu die *softmax*-Funktion zum Einsatz kommt.

$$\mathrm{softmax}(x)_i = \frac{e^{\frac{y_i}{T}}}{\sum_j^N e^{\frac{y_j}{T}}}$$

Diese Funktion errechnet eine Wahrscheinlichkeitsverteilung aller Symbolfolgen des Vokabulars über ein festes Intervall von 0 bis 1. Dazu wird *e* mit dem jeweils errechneten Wahrscheinlichkeitswert *y* potenziert, wodurch sichergestellt ist, dass stets ein positiver Wert entsteht. Dieser Wert wird durch die Summe aller auf diese Weise errechneten Werte geteilt, sodass der resultierende Wert im Intervall von 0 bis 1 liegt. Die Summe aller auf diese Weise errechneten Werte ist 1. Besonders hohe Wahrscheinlichkeitswerte liegen nun nahe bei 1, der errechnete Maximalwert steht jedoch in Relation zu anderen hohen Werten auf dem Intervall von 0 bis 1 und wird auf diese Weise gedämpft, oder *softened*.

Das Sampling des nächsten Tokens erfolgt entlang dieser Wahrscheinlichkeitsverteilung stochastisch. Im Effekt wird auf diese Weise nicht notwendigerweise das Folgesymbol mit der höchsten Übergangswahrscheinlichkeit selegiert. Durch einen zusätzlichen Faktor *T* im Nenner des Exponenten lässt sich auf diesen stochastischen Prozess Einfluss nehmen. *T* wird oft als Temperatur der *softmax*-Funktion bezeichnet, da sie, ähnlich wie in einem physikalischen System, das Niveau an Unordnung zwischen den Elementen steigert und somit die Zufälligkeit der Selektion erhöht. Ein höherer Wert *T* führt dazu, dass der Unterschied zwischen dem vom Modell errechneten Maximalwert und anderen Tokens mit hohen Wahrscheinlichkeiten kleiner wird. Dadurch steigt die Wahrscheinlichkeit, dass Folgesymbole mit niedrigeren Werten selegiert werden.

Damit ist ein zweiter Punkt innerhalb von GPT-Systemen identifiziert, an denen ein Zufallselement integriert wird. Diese Elemente ermöglichen den willkürlichen symbolischen Eingriff von Entwickler*innen in das Operieren eines Systems, das zumeist als Emergenz aus statistischen Werten eines Datensatzes, also als Black Box imaginiert wird. Mittels des Faktors *T* oder des Grads der Abweichung von einem symmetrischen Ausgangszustand wird seitens der Entwickler*innen aktiv Einfluss auf den statistischen Prozess genommen.[31] Unter diesem Gesichtspunkt ähneln die Systeme des *deep learning* der Shannon'schen Maschine zur Vorhersage der nächsten Züge eines menschlichen Gegenübers.

31 Gründe für solche Eingriffe können rein technischer Natur sein, z. B. um ein sogenanntes *overfitting* des Modells, d. h. eine Duplizierung der Werte des Trainingsdatensatzes zu verhindern. Speziell die Veränderung der Temperatur im Output-Layer kann jedoch genutzt werden, um auf die vom Sprachmodell realisierte Simulation eines kohärenten Textflusses Einfluss zu nehmen. Korrekte Werte von T sind somit auch eine Frage des ästhetischen Geschmacks.

Symbolische und subsymbolische Repräsentation

Die Technologie der KNNs lässt sich auf Vorarbeiten aus der Kybernetik zurückführen. Dieses Forschungsprogramm, das insbesondere die Redukti-

on unterschiedlicher beobachtbarer Systeme aufeinander vornimmt, wurde mehrfach als eine Wissenschaft des Krieges charakterisiert.[32] Dieter Mersch hält fest:

> Zu den nachhaltigsten und unheimlichsten Versprechen der ‹kybernetischen Hypothese› zählte […], dass sie von Anfang an – bereits mit der ersten Publikation von Norbert Wiener – sich anschickte, kognitive Prozesse nach der Logik des Feedbacks zu analysieren und die Regime der Rekursion auf neuronale Netzwerke anzuwenden.[33]

Die Irrationalität psychischer Systeme und menschlichen Verhaltens wird parallelisiert mit dem Zufallselement oder Chaos in physischen Systemen und den Theoremen statistischer Informationssysteme.

> Wenn ein Modell Intelligenz in Begriffen mathematischer Informationstheorie, also stochastischer Entropie, definiert, und damit auf Selektion reduziert, und dieses Modell dann wiederum zur Modellierung künstlicher Gehirne oder Intelligenz verwendet wird, dann wird die Tatsache, dass der Wille zur Metapher eben nicht in Selektion, sondern im Erfinden, Übertragen, Übersetzen, Ersetzen etc. besteht, effektiv vergessen.[34]

Die logische Fiktion einer künstlichen, kontrollierbaren Kognition verlangt die willentliche Aussetzung der Ungläubigkeit hinsichtlich des begrenzten symbolischen Repertoires mechanischer Berechenbarkeit. Bei der Konstruktion solcher Fiktionen helfen Begriffsschöpfungen wie die des Subsymbolischen.

Einführungen in die Geschichte der KI verwenden das Begriffspaar des Symbolischen und Subsymbolischen häufig, um auf fachliche Entwicklungen seit den Gründungs-Workshops der 1950er Jahre zu verweisen.[35] Symbolische KI stellt dabei jenen Bereich der KI-Forschung dar, der einen Katalog an intelligenten Funktionen des menschlichen Gehirns als symbolisch anschreibbare Anweisungen für eine Rechenmaschine formulieren und implementieren möchte. Peter Galison zeichnet nach, wie dieses Forschungsprogramm 1944 von dem Harvard-Psychologen Edwin Boring an Norbert Wiener als Herausforderung für dessen Black-Box-Engineering herangetragen wurde.[36] Subsymbolische Ansätze wiederum sind in Abgrenzung zur symbolischen KI entstanden. Frank Rosenblatts Perzeptron ist 1958 das erste künstliche neuronale Netz und greift auf Vorarbeiten von Warren McCulloch und Walter Pitts zurück. Hannes Bajohr beschreibt die unterschiedlichen Paradigmen folgendermaßen:

> Meint ‹lernen› bei Expertensystemen die Erweiterung der Wissensdatenbank, sind Perzeptrone auf Wiederholungen innerhalb der zu lernenden Domäne angewiesen; folgt das Expertensystem linearen Wenn-dann-Strukturen, hat die Schaltung des Perzeptrons einen parallelen Aufbau und kommt ohne die Trennung von Fakten und Regeln aus.[37]

Die Geschichtsschreibung entlang dieser starken Trennung entstammt der Selbsthistorisierung der KI-Forschung und erzählt die Geschichte der aktuell

32 Vgl. Peter Galison: The Ontology of the Enemy: Norbert Wiener and the Cybernetic Vision, in: *Critical Inquiry*, Bd. 21, Nr. 1, 1994, 228–266; Axel Roch, Bernhard Siegert: Maschinen, die Maschinen verfolgen. Über Claude E. Shannons und Norbert Wieners Flugabwehrsysteme, in: Schade, Tholen (Hg.): *Konfigurationen. Zwischen Kunst und Medien*, München 1999, 219–230.

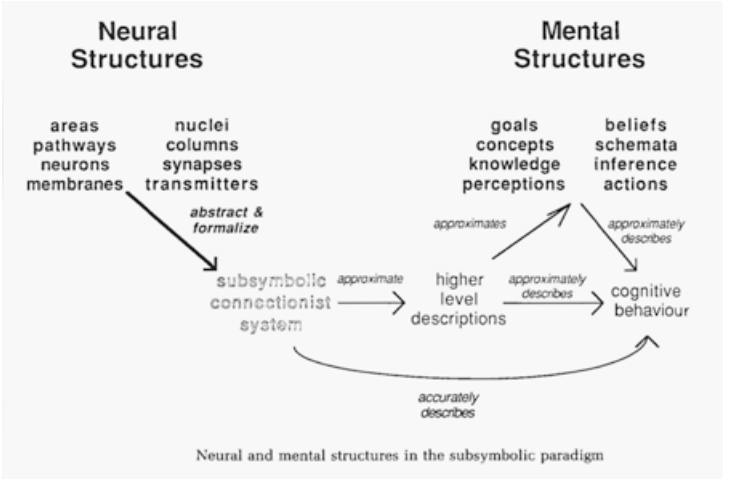
33 Dieter Mersch: Ordo ab Chao/Order from Noise. Überlegungen zur Diskursgeschichte der Kybernetik, in: Annette Brauerhoch u. a. (Hg.): *Entautomatisierung*, Paderborn 2017, 19–38, hier 33.

34 Vagt: Nietzsche, Ashby und die logische Fiktion künstlicher Intelligenz, 195, Herv. im Orig.

35 Vgl. Melanie Mitchell: The Roots of Artificial Intelligence, in: dies.: *Artificial Intelligence: A Guide for Thinking Humans*, New York 2019, 17–34, hier 21–23.

36 Vgl. Galison: The Ontology of the Enemy, 247.

37 Bajohr: Die ‹Gestalt› der KI, 172.



Neural and mental structures in the subsymbolic paradigm

Abb. 2 **Smolenskys Diagramm des subsymbolischen Systems, 1987**

dominanten Position, dem konnektionistischen Paradigma.[38] Dieses wurde in der Frühzeit der KI-Forschung aus unterschiedlichen Gründen von der Fachgemeinschaft und den Forschungsinstitutionen abgelehnt.[39] Marvin Minsky und Seymour Papert formulierten 1969 eine kritische Evaluation des Rosenblatt'schen Perzeptrons, die der Technologie wenig Zukunftsperspektiven einräumte.[40] Erst in den 1980er Jahren, unter anderem aufgrund der Popularisierung des oben erwähnten *backpropagation*-Algorithmus durch Rumelhardt, Hinton und Williams, erstarkten konnektionistische Ansätze erneut[41] und avancierten von da an zum dominanten Ansatz der KI-Soft- und Hardware-Entwicklung. In dieser Zeit brachte der Kognitionsforscher Paul Smolensky erst den Begriff des Subsymbolischen prominent in Umlauf, um ausgehend vom konnektionistischen Paradigma der KI-Forschung ein alternatives Modell menschlicher Kognition zu formulieren. «An alternative to the symbolic paradigm is what I call the subsymbolic paradigm. In this paradigm, there is an intermediate level of structure between the neural and symbolic levels.»[42] Smolensky grenzte dieses Paradigma von einem symbolischen ab, das auf den genannten Zwischenbereich verzichte und neuronale und symbolische Strukturen direkt ineinander zu übersetzen suche. Dies führe, so Smolenskys Darstellung, unweigerlich zur Interpretation mentaler Strukturen als digitale Computer, wozu der Konnektionismus eine Alternative bieten möchte. «At the fundamental level of subsymbolic formalism, we have moved from thinking about cognition in terms of discrete processes to thinking in terms of continuous processes. This means that different mathematical concepts apply.»[43]

Das Subsymbolische nach Smolensky eröffnet einen fiktiven Zwischenbereich, in dem trotz unveränderter digitaler Bedingungen mechanischer Berechnung über Kognition als kontinuierlichen Prozess nachgedacht werden kann. Darüber hinaus geht diese Idee in die Architektur der Systeme selbst ein. Die maschinelle Simulation kognitiver Fähigkeiten zielt im Paradigma des Subsymbolischen nicht auf die direkte Verknüpfung von neuronalen und symbolischen Strukturen, sondern auf einen Zwischenbereich, in dem sich diese Bereiche berühren. Der Begriff des Subsymbolischen positioniert ein «subsymbolic connectionist system» (Abb. 2) zwischen neuronalen Strukturen und stabilisierten mentalen Kategorien der Repräsentation.

Ziel der Forschung bleibt jedoch, eine Übereinstimmung zwischen Welt und Repräsentation zu erfassen: «[T]he subsymbolic paradigm involves

38 Vgl. Nicholas Thompson: An AI Pioneer Explains the Evolution of Neural Networks, *Wired*, 13.5.2019, *wired.com/story/ai-pioneer-explains-evolution-neural-networks* (9.12.2024).

39 Paul Edwards hat darauf hingewiesen, dass die Entwicklung symbolischer AI eng mit der Etablierung von *time-sharing*-Systemen in Verbindung stand. Dies wiederum brachte die Forschung in die Nähe von Militärprojekten des ARPA, vgl. Paul N. Edwards: *The Closed World: Computers and the Politics of Discourse in Cold War America*, Cambridge (MA), London 1997, 239–419.

40 Vgl. Marvin Minsky, Seymour Papert: *Perceptrons: An Introduction to Computational Geometry*, Cambridge (MA) 1988.

41 Vgl. Mikel Olazaran: A Sociological Study of the Official History of the Perceptrons Controversy, in: *Social Studies of Science*, Bd. 26, Nr. 3, August 1996, 611–659, hier 645. Olazarans Studie ist eine detaillierte Darstellung der Debatte um das Perzeptron bis in die mittleren 1990er Jahre.

42 Paul Smolensky: Connectionist AI, Symbolic AI, and the Brain, in: *Artificial Intelligence Review*, Bd. 1, 1987, 95–109, hier 99, *doi.org/10.1007/BF00130011*.

connectionist systems using so-called distributed representations, as opposed to local representations.»[44] Beide Ansätze bleiben, so argumentiert Suchman, im übergeordneten Paradigma des Kognitivismus:

> [I]n their fundamental assumptions and commitments, deep-learning and symbolic approaches share more than they offer up in the way of alternatives. While one relies on statistical analysis and the other on the encoding of algorithms that determine computational operations (so-called rules), both have already translated cognition into a problem of computation before the research begins.[45]

Der Begriff des Subsymbolischen fügt der Fiktion einer symbolisch kontrollierbaren Kognition ein neues Kapitel hinzu. Melanie Mitchells Einführungswerk zu aktuellen KI-Technologien etwa verweist auf das Symbol als Sprache im Gehirn, um das Konzept des Subsymbolischen einzuführen:

> [T]he human brain has given rise to language, which allows you to use symbols (words and phrases) to tell me – often imperfectly – what your thoughts are about or why you did a certain thing. In this sense, our neural firings can be considered subsymbolic, in that they underlie the symbols our brains somehow create.[46]

Ziel der Technologie bleibt bis heute die Rekonstruktion einer stabilen Repräsentationsstruktur, die die Selektion mentaler Kategorien mit einer als extern angenommenen Welt verknüpft.

Fazit

Eine kritische Auseinandersetzung mit aktuellen KI-Systemen kann von der grundlegenden Kontinuität digitaler Hardware und von historischen Vorarbeiten der Kybernetik ausgehen. Auf diese Weise lässt sich die Rede von der Black Box emergenter, konnektionistischer Systeme als strategisch wertvolle interpretative Illusion besprechen, die vor konkrete Entscheidungen in der Entwicklung neuronaler Netze geschoben wird.[47] Dazu ist es notwendig, Momente der Entscheidung innerhalb der technologischen Systeme zu identifizieren. Dieser Artikel hat zum einen vorgeschlagen, konkrete Stellen der Integration von stochastischen Elementen als solche Momente zu begreifen. Zum anderen lässt sich die Geschichte fachinterner Begriffsschöpfungen befragen. Der Begriff des Subsymbolischen, so konnte gezeigt werden, trägt im Feld des konnektionistischen Paradigmas der KI-Forschung zur Naturalisierung einer Repräsentationslogik bei, mit der Kognition auf die Selektion adäquater mentaler Kategorien reduzierbar wird. Wenn daher mit Théo Lepage-Richer zugestanden wird, dass ‹Neuralität› stets die Aufgabe hatte, die Kategorie der Intelligenz zwischen Menschen und Maschinen zu verhandeln,[48] so lässt sich ausgehend von der hier vorgenommenen Einordnung des Begriffs des Subsymbolischen festhalten, dass ein Teil dieser Aushandlung die Fiktion einer symbolisch kontrollierbaren Kognition ist. Diese Fiktion wird in der aktuellen KI-Forschung genutzt,

43 Ebd., 102. Die Selbstpositionierung als vielversprechender Gegenentwurf zu einem alten Modell führt dazu, dass Forschung zu KNNs in den 1980er Jahren mit staatlichen Fördermitteln am Canadian Institute for Advanced Research (CIFAR) erfolgreich vorangetrieben wird. Théo Lepage-Richer zeigt, wie in Hintons Forschung am CIFAR Phänomene menschlicher Kognition und Modelle der Computation in einer umfassenden Theorie der Organisation zusammenfallen, vgl. Théo Lepage-Richer: Neural Media, in: Dhaliwal, Lepage-Richer, Suchman (Hg.): *Neural Networks*, 20–53, hier 40.

44 Smolensky: Connectionist AI, Symbolic AI, and the Brain, 100.

45 Suchman: The Neural Network at Its Limits, 106.

46 Mitchell: The Roots of Artificial Intelligence, 30, Herv. im Orig.

47 Zur Kritik einer «black box casuistry» vgl. Offert, Dhaliwal: The Method of Critical AI Studies, 3 f.

48 Vgl. Lepage-Richer: Neural Media, 47.

um die eigene Geschichte zu schreiben oder um neue Ansätze nach außen zu kommunizieren. Darüber hinaus hat sie eine längere Geschichte: Sie entsteht dadurch, dass Intelligenz als stochastischer Prozess definiert wird, wie es bereits Shannons *Mind-Reading (?) Machine* demonstrierte.

Nicht-deterministisch ablaufende Maschinen stellen historisch kein Novum dar, ihre Verbreitung und Veralltäglichung dagegen schon. Dies wird den Diskurs der Mind Control rekonfigurieren, denn was Mind, was Control ist, wird aktuell neu ausgelotet. Probabilistische Systeme bilden einen eigenartigen Schauplatz der Suggestibilität, die Interaktion via *prompting* ist Medienpraxis als Einflussnahme. Shannons vergnügtes eingeklammertes Fragezeichen im Titel seines Memos darf vor diesem Hintergrund kassiert werden. Der Titel eines Memorandums für gegenwärtige künstliche neuronale Netze könnte vielleicht *A Mind-Writing Machine* lauten, denn, wie Mai Wegener vorsichtig spekuliert: «[E]s ist eine Bildung des Unbewussten. Es ist die sich entwickelnde Ausdehnung von Sätzen, Algorithmen, die die Programmierer der Maschine implementiert haben in dem Glauben, sich im Allgemeinen, vielleicht in einer Art Universalsprache zu bewegen».[49]

49 Judith Kasper, Mai Wegener: Protokoll 12, in: Anna Tuschling, Andreas Sudmann, Bernhard J. Dotzler (Hg.): *ChatGPT und andere ‹Quatschmaschinen›. Gespräche mit Künstlicher Intelligenz*, Bielefeld 2023, 86–94, hier 92.

Frühe Versionen dieses Artikels wurden von Moritz Hiller, Hannes Bajohr und Christina Vagt durchgesehen, für deren wertvolle Kommentare und Einsichten ich mich bedanken möchte. Ebenso bin ich dankbar für den interdisziplinären Austausch mit Filip Tomaska aus der Dynamic Neuroscience an der University of California, Santa Barbara, durch den der Text sehr profitiert hat.