



Roger Häußling, Claudius Härpfer, Marco Schmitt (Hg.)

SOZIOLOGIE DER KUNSTLICHEN INTELLIGENZ

Perspektiven der Relationalen Soziologie und Netzwerkforschung

[transcript] Digitale Soziologie

Roger Häußling, Claudius Härpfer, Marco Schmitt (Hg.)
Soziologie der Künstlichen Intelligenz

Digitale Soziologie | Band 5

Editorial

Die Digitalisierung der Gesellschaft eröffnet neue soziologische Forschungsfelder, beeinflusst aber auch klassische Phänomene wie Infrastrukturen, Vergemeinschaftungen und soziale Praktiken grundlegend. Die Reihe **Digitale Soziologie** bietet sowohl theoretisch-konzeptionellen als auch empirischen Untersuchungen hierzu eine gemeinsame editorische Plattform. Insbesondere versammelt sie Arbeiten, die neue, auch unorthodoxe Zugriffe auf substanzielle Fragestellungen der Soziologie anbieten, die aktuelle methodische und konzeptionelle Entwicklungen aufgreifen oder innovative Methodenkombinationen nutzen. Dabei legt sie besonderen Wert auf die Diskussion der Rolle der Soziologie angesichts der digitalisierten Gesellschaft und auf die Reflexion der daraus folgenden Implikationen bspw. für Infrastrukturen und die wissenschaftliche Praxis. Die Reihe schließt deutsch- und englischsprachige Monografien und Sammelbände ebenso ein wie herausragende Qualifikationsarbeiten.

Die Reihe wird herausgegeben von Roger Häußling, Katharina Kinder-Kurlanda, Sophie Mützel, Jan-Hendrik Passoth und Andreas Schmitz.

Bei Interesse an der Publikation in dieser Reihe können die Herausgeber*innen kontaktiert werden.

Roger Häußling (Prof. Dr. phil.), geb. 1969, hat seit 2009 den Lehrstuhl für Technik- und Organisationssoziologie (STO) an der RWTH Aachen inne. Zu seinen Arbeitsschwerpunkten zählen unter anderem Innovationssoziologie, Techniksoziologie, Transformationssoziologie, Netzwerkforschung und Relationalismus.

Claudius Härpfer (Dr. phil.), geb. 1981, ist seit 2020 wissenschaftlicher Mitarbeiter am Lehrstuhl für Technik- und Organisationssoziologie (STO) der RWTH Aachen. Zu seinen Arbeitsschwerpunkten zählen unter anderem Techniksoziologie, Transformationssoziologie, Netzwerkforschung und relationale Soziologie sowie soziologische Theorie und Theoriegeschichte.

Marco Schmitt (Dr. rer. pol.), geb. 1974, ist seit 2015 wissenschaftlicher Mitarbeiter am Lehrstuhl für Technik- und Organisationssoziologie (STO) der RWTH Aachen. Zu seinen Arbeitsschwerpunkten zählen unter anderem Transformationssoziologie, relationale Soziologie, Netzwerktheorie und -forschung sowie transformative Forschung.

Roger Häußling, Claudius Härpfer, Marco Schmitt (Hg.)

Soziologie der Künstlichen Intelligenz

Perspektiven der Relationalen Soziologie und Netzwerkforschung

[transcript]

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <https://dnb.dnb.de/> abrufbar.



Dieses Werk ist lizenziert unter der Creative Commons Attribution-ShareAlike 4.0 Lizenz (BY-SA). Diese Lizenz erlaubt unter Voraussetzung der Namensnennung des Urhebers die Bearbeitung, Vervielfältigung und Verbreitung des Materials in jedem Format oder Medium für beliebige Zwecke, auch kommerziell, sofern der neu entstandene Text unter derselben Lizenz wie das Original verbreitet wird.

<https://creativecommons.org/licenses/by-sa/4.0/>

Die Bedingungen der Creative-Commons-Lizenz gelten nur für Originalmaterial. Die Wiederverwendung von Material aus anderen Quellen (gekennzeichnet mit Quellenangabe) wie z.B. Schaubilder, Abbildungen, Fotos und Textauszüge erfordert ggf. weitere Nutzungsgenehmigungen durch den jeweiligen Rechteinhaber.

Erschienen 2024 im transcript Verlag, Bielefeld

© Roger Häußling, Claudius Härpfer, Marco Schmitt (Hg.)

Umschlaggestaltung: Kordula Röckenhaus, Bielefeld

Umschlagabbildung: generiert mit Playground AI (<https://playground.com/>)

Druck: Majuskel Medienproduktion GmbH, Wetzlar

<https://doi.org/10.14361/9783839474112>

Print-ISBN: 978-3-8376-7411-8

PDF-ISBN: 978-3-8394-7411-2

Buchreihen-ISSN: 2751-3149

Buchreihen-eISSN: 2751-3157

Gedruckt auf alterungsbeständigem Papier mit chlorfrei gebleichtem Zellstoff.

Inhalt

Einleitung

Roger Häußling, Claudius Härpfer, Marco Schmitt7

Mustererkennung in Datennetzwerken

Zu einer Relationalen Soziologie neuer Formen »Künstlicher Intelligenz«

Roger Häußling, Marco Schmitt 27

KI zwischen Blackbox und Transparenz

Das Koppeln und Entkoppeln von Kontrollprojekten

Marco Schmitt, Christoph Heckwolf51

Von der Kunst des Lernens

Einige Bemerkungen zur Intentionalität von In- und Output

Claudius Härpfer, Nadine Diefenbach 85

Deep Learning Techniken als Boundary Objects zwischen Entwicklungs- und Anwendungsfeld

Philip Roth, Matthias Dorgeist, Astrid Schulz 115

Über Identitäten und Selbstverständnis medizinischen Personals in Zeiten künstlicher Intelligenz und Algorithmierung

Tabea Bongert, Dhenya Schwarz 153

Plattform und jetzt?

Gründungsstories digitaler Start-ups zwischen Weltverbesserung und digitaler Transformation

Tim Franke, Niklas Strüver, Sascha Zantis175

Künstliche Intelligenz in der Zukunftsforschung	
<i>Axel Zweck, Thomas Werner</i>	219
Künstliche Intelligenz und relationale Techniksoziologie	
Ein Schluss	
<i>Roger Häußling, Marco Schmitt</i>	257
»Ja, Transparenz in Bezug auf die Funktionsweise von KI-Modellen wie ChatGPT ist äußerst wichtig.«	
Ein Interview mit ChatGPT	275
Glossar	
<i>Jannis Grub</i>	303
Angaben zu den Autor:innen	309
Sachregister	311
Personenregister	315

Einleitung

Roger Häußling, Claudius Härpfer, Marco Schmitt

1. Künstliche Intelligenz – ein Thema in aller Munde!

Die Oral-B Genius X¹ ist eine elektrische Zahnbürste, die nicht nur dank mechanischer Rotation des runden Bürstenkopfes das händisch-menschliche Zähneputzen unterstützt und somit laut Herstellerangabe bis zu 100 % mehr Plaque entfernt als herkömmliche Zahnbürsten. Sie weist darüber hinaus die das Gerät bedienenden Menschen dank künstlicher Intelligenz auf Fehler beim Putzen hin. War man bei Vorgängermodellen noch darauf angewiesen, für ähnliche Funktionen stillzustehen und sich beim Putzen mit der Smartphone-Kamera in den Mund filmen zu lassen, so erfasst das 2019 erschienene Modell nun die Putzbewegungen mit der eingebauten Sensorik der Zahnbürste und errechnet über die mit Bluetooth verbundene Smartphone-Anwendung aus den Bewegungsmustern Putzstile, die dazu dienen, den Nutzer in Echtzeit auf Optimierungspotentiale beim Putzen hinzuweisen. Dieses Beispiel ist insbesondere in seiner vermeintlichen Trivialität fernab großer Erwartungen in Richtung einer starken, menschenähnlichen KI interessant (vgl. z.B. Zweig 2019: 267–280). Künstliche Intelligenz wird in einem abgesteckten Rahmen für eine spezifische Anwendung genutzt, um konventionelle Technik zu ergänzen und so schnellere oder bessere Ergebnisse zu erzielen, als dies rein mit konventionellen Mitteln möglich war. Diese Zahnbürste mit Zusatzfeature ist typisch für eine neue Art von alltäglichen Anwendungen und zeigt, wie verschiedene Faktoren einer soziotechnischen Innovation zusammenspielen. Bei genauem Hinsehen ist die (schwache) künstliche Intelligenz natürlich nicht in der Zahnbürste verortet, denn diese bietet mit der Sensorik lediglich einen Teil der Mensch-Maschine-Schnittstelle. Die Verarbeitung der Daten erfolgt über

1 <https://www.oralb.de/de-de/produktkollektionen/elektrische-zahnbursten/genius>, abgerufen am 01.03.2023.

die gekoppelte Anwendung auf dem Smartphone, die wiederum auf einen großen Datenbestand zurückgreift, um Muster identifizieren zu können, mit denen typische Putzfehler korrelieren und diese dann auszugeben. Grundbedingung hierfür ist natürlich nicht nur, dass der nutzende Mensch bereit ist, Daten über sein tägliches Zahnputzverhalten mit dem Hersteller der Bürste zu teilen, sondern auch, dass er sein Smartphone zum Zähneputzen mit ins Badezimmer nimmt. Technische Entwicklung und soziale Konvention gehen hier Hand in Hand. Dass nun die Bürste – die also Erwachsenen nebenbei dabei helfen soll, ihr Zahnputzverhalten zu optimieren – als intelligent beworben wird, spricht für die breite Akzeptanz und Verbreitung, die künstliche Intelligenz in den letzten Jahren im Alltag erlangt hat.

Es gäbe fraglos eine ganze Reihe anderer Beispiele. Fast 70 Jahre nach Schöpfung des Begriffs künstliche Intelligenz Mitte der 1950er Jahre (vgl. McCarthy/Minsky/Rochester/Shannon 2006) – für einen Abriss zur Geschichte der KI siehe Zweck/Werner in diesem Band – hat die Entwicklung, auch dank Datafizierung und Digitalisierung, dahingehend Fahrt aufgenommen, als dass unser Alltag mit Anwendungen durchdrungen ist, die sich künstlicher Intelligenz bedienen. Viele davon werden – wie die Zahnbürste – ohne große Kontroversen am Markt etabliert. Soziotechnische Meilensteine wie der im November 2022 veröffentlichte Textgenerator ChatGPT² hingegen sorgen für breite Diskussionen bis in die Tagespresse hinein und verdeutlichen damit die gesellschaftliche Relevanz des Themas. Hierbei hat es das Unternehmen OpenAI geschafft, einen Chatbot zu entwickeln, der auf Basis eines Sprachmodells und eines großen Datenpools in der Lage ist, nicht nur eine normale Konversation in bisher unerreichter Authentizität zu führen, sondern – wenn man ihn darum bittet – auch Hausarbeiten, Übersetzungen, Lyrik oder Programmcode zu schreiben. Er kann also im Gegensatz zur Zahnbürste eine Vielzahl von Aufgaben ausführen, die das Intelligenzniveau konventioneller Technik überschreiten. Diese Entwicklung hat das Potential das Arbeiten mit Texten massiv zu verändern und macht auch vor der kreativen Seite des Arbeitens nicht halt. Die Grenzen dieser Entwicklung sind derzeit noch nicht abzusehen. Was die kurz nach Veröffentlichung des Tools entflammte Diskussion um ChatGPT aber zeigt, ist, dass schnell rechtliche und ethische Fragen im Fokus stehen (vgl. z.B. Bleher/Braun 2023; Nehlsen/Fleck 2023), während die langfristig relevante soziologische Perspektive ins Hintertreffen gerät.

2 <https://openai.com/chatgpt>, abgerufen am 01.03.2023.

GPT steht für »Generative Pretrained Transformer« und basiert auf einem Deep Learning Verfahren mittels eines künstlichen neuronalen Netzes – also im Grunde auf nichts Neuem. Insofern ist erklärungsbedürftig, warum ChatGPT so viel Aufmerksamkeit erfahren hat, wo es doch haufenweise andere Anwendungen von Deep Learning Verfahren gibt. Von Nutzer:innenseite springen dabei zwei Dinge ins Auge: Die Niederschwelligkeit seiner Nutzung und die Qualität seiner Ergebnisse. Sowohl die Registrierung für eine kostenfreie Nutzung dieser Künstlichen Intelligenz, als auch die Handhabung ihrer Eingabemaske geschehen im Handumdrehen, ohne dass jeweils spezifische Kenntnisse notwendig oder besondere Hürden zu nehmen wären. Die Frage der Qualität ist engstens – wie im vorliegenden Band mehrfach herauszuarbeiten sein wird – mit der Quantität und Qualität der Lerndaten verknüpft. Hier hat OpenAI neue Maßstäbe gesetzt, ohne jedoch irgendwelche innovativen neuen Verfahren zu etablieren, die vorher bei anderen KI-Verfahren nicht zur Anwendung gekommen wären. Vielmehr wurden etablierte Verfahren des künstlichen Lernens mehrstufig zur Anwendung gebracht, sodass sich schon die Frage stellt, ob wir es hier wirklich mit einer neuen Stufe künstlicher Intelligenz zu tun haben.

Eine Besonderheit weist GPT allerdings doch auf, die deren Neuheitsanspruch unterstreichen könnte: Während bisherige KI-Verfahren durch eine Fokussierung auf ein konkretes Problem und dabei auf einen bestimmten Analysegegenstand gekennzeichnet waren, ist GPT breiter in Bezug auf Probleme und Analysegegenstände angelegt. Dies wird besonders bei GPT-4 deutlich, insofern nicht nur Text, sondern auch Bilder als Eingabe erfolgen können. Aus einem Bild als Vorlage kann dann GPT-4 zum Beispiel einen Code für eine Homepage generieren. Bereits ChatGPT-3 kann sowohl lyrische Texte in einem gewünschten Stil, wissenschaftliche Abhandlungen zu verschiedensten Themen, politische Reden aus unterschiedlichen parteipolitischen Positionen heraus oder Quellcode für verschiedene Anwendungen generieren – je nach dem, worum man die Künstliche Intelligenz konkret bittet. Genau hierin unterscheidet sich GPT von bisherigen KI-Verfahren!

Das erklärte Ziel der Macher:innen von OpenAI ist es, an der Entwicklung einer so genannten AGI, Artificial General Intelligence zu arbeiten – einer starken Künstlichen Intelligenz also, die für alle möglichen Aufgaben zum Einsatz gebracht werden kann. Auf technischer Ebene bleibt allerdings unklar, wie genau GPT-4 Schritte in diese Richtung vollzieht. Vormalig waren Bilderkennung und Textmining zwei getrennte Angelegenheiten, für die jeweils eigene Lerner zum Einsatz kamen. Wie nun Text- und Bilddaten zusammengeführt wer-

den, bleibt ein streng gehütetes Geheimnis von OpenAI, sodass der eigentliche Neuheitswert nicht wirklich unter die Lupe genommen werden kann.

Bleibt noch eine weitere Besonderheit von GPT: Wird die Datenlage dünn, dann fängt GPT zu halluzinieren an, sprich: die KI erfindet kontextbezogen Dinge hinzu – und zwar oftmals so gekonnt, dass es für Nicht-Sachkundige nicht gleich erkennbar ist. Das erzeugt auf der Nutzer:innenseite im Umgang mit der KI generell Unsicherheiten, da man bei den von ihr angebotenen Ergebnissen nie genau wissen kann, ob Halluzinationsanteile enthalten sind. Im Grunde äußert sich darin aber auch eine besondere Qualität der Halluzinationen. Sie sind oftmals doch so gut, dass sie für uns als Fakten erscheinen. Diese ›kreative‹ Leistung von GPT wird bislang nur selten herausgestellt. Immerhin gelingt es der KI, menschlichen Nutzer:innen mitunter täuschend echte Phantasmagorien aufzutischen. Eine solche Münchhausen-Qualität hätte man vor ein paar Jahren einem künstlichen neuronalen Netz nicht zugetraut. Zu Recht wird auf die Gefahr einer Potenzierung von »Fake News« durch solche Dienste hingewiesen, dabei aber auf die Kehrseite des Phänomens bislang wenig abgehoben: Sind diese halluzinatorischen Eigenschaften von GPT nicht auch gezielt nutzbar für spekulative Problemstellungen, um uns mit einer Vielfalt an sinnvoll erscheinenden Brückenargumenten bezüglich unserer Wissenslücken zu versorgen?

Vor diesem Hintergrund haben wir ChatGPT zu seiner technischen Verfasstheit, zu seiner Funktionsweise sowie zu Chancen und Risiken von Deep Learning im Allgemeinen befragt. Auch wollten wir schauen, ob sich die KI in die Enge treiben lässt, wie sie sich eine Soziologie des Deep Learnings vorstellt, und was sie von netzwerkforscherischen Zugängen zur KI hält. Die Ergebnisse dieses Interviews finden sich am Ende dieses Bandes.

Für eine intensive soziologische Auseinandersetzung mit der neueren Künstlichen Intelligenz fehlen bislang zum einen noch Möglichkeiten einer empirischen Auseinandersetzung mit dem Thema jenseits öffentlicher oder wissenschaftlicher Diskurse und zum anderen auch ein theoretischer Zugang der techniksoziologisch, wie sozialtheoretisch überzeugt. Diese Lücke möchte der Sammelband schließen. Es liegt nahe, einen solchen Zugang in der Netzwerktheorie zu suchen, da sie begrifflich wie methodisch auf das kalibriert ist, was auch (künstliche) neuronale Netzwerke ausmacht. Der Sammelband versucht entsprechend auf der Basis einer dezidiert relationalen Theorieperspektive und ausgewählter Anwendungen, soziologische Analyseangebote für das Phänomen einer, große Teile von Wirtschaft und Gesellschaft durchwirkenden, Künstlichen Intelligenz zu entwickeln. Hierbei erweist sich

die Theorie von Harrison White als besonders vielversprechend (systematisch hierzu Häußling/Schmitt in diesem Band). Mit der Ableitung zentraler Theoriekomponenten aus einer Vielzahl von empirischen Arbeiten aus der Netzwerkforschung und den grundlegenden Komponenten von Verteilung und Selbstähnlichkeiten bietet White auch eine dem Phänomen der neueren KI stark angenäherte Theoriesprache, da beide stark auf Mustererkennung setzen. Seine zentralen Begriffe Identität, Kontrolle, Netzwerkdomäne, Switching, Stil und Kontrollregime ermöglichen zudem aufgrund ihrer Abstraktheit ihre direkte Anwendbarkeit auf KI-Phänomene par excellence. Gerade weil KI im Kern netzwerkartig ist, kann Whites elaborierte Netzwerktheorie so fruchtbringend daran angeschlossen werden und neue Einsichten in das Phänomen KI liefern.

2. Zum Begriff der »neueren KI«

Unter dem Begriff der neueren KI verstehen wir in erster Linie Anwendungen, die auf Machine Learning (ML) und Deep Learning (DL) zurückgreifen. Die grundlegende Engineering-Idee dieser in den letzten Jahren aufgekommenen Anwendungen ist, nicht mehr alles programmieren zu wollen (in manchen Fällen, wie zum Beispiel bei autonomen Fahrzeugen, zu können), sondern stattdessen künstliche neuronale Netze darauf zu programmieren, Muster zu erkennen. Die dabei zum Einsatz kommenden Programme, die so genannten Learner, sind vergleichsweise schlicht und lassen sich – wie die oben erwähnte Zahnbürste – in vielen Fällen problemlos als Smartphone-Anwendung realisieren, komplex hingegen sind die Datenmengen, anhand derer KI trainiert wird. Dies ist möglich durch den erleichterten Zugang zu großen Datenmengen durch technische und soziale Entwicklungen und die Fortschritte im Bereich der Machine Learning-Verfahren, insbesondere auf der Basis neuronaler Netze.

Kern dieser Learning-Verfahren sind jene künstlichen neuronalen Netze (vgl. z.B. Rosengrün 2021: 24–29; Mainzer 2019: 104–124). Vom menschlichen Nervensystem inspiriert, werden komplexe Schaltungen aufgebaut, in denen Neuronen als Knoten »feuern«, also Verbindungen zu anderen Neuronen aufbauen, wenn sie aktiviert werden, weil ein bestimmter Grenzwert erreicht ist. Als Stellschrauben dienen zum einen die Gewichtung des Inputs und zum anderen die Höhe des Grenzwerts. Diese Netze sind so aufgebaut, dass es eine Eingabeschicht (Input Layer), in der Regel eine oder mehrere Zwischenschicht-

ten (Hidden Layers) und eine Ausgabeschicht (Output Layer) gibt. Durch diese Struktur können große systematische Zusammenhänge modelliert werden, indem der Learner im Rahmen seiner Rechenleistung die möglichen Fälle für die einzelnen Neuronen und Schichten nacheinander durchspielt und so statistische Korrelationen erhält, die irgendwann mit Blick auf den weiteren Aufwand ein befriedigendes Ergebnis liefern.

In der Lernpraxis unterscheidet man zwischen drei grundlegenden Verfahren, um unterschiedliche Funktionen zu erfüllen. Das gängigste Verfahren ist das sogenannte Supervised Learning, bei dem der Mensch dem Learner einen (vergleichsweise schmalen) Trainingsdatensatz zur Verfügung stellt, in dem der Learner Input-Daten und Output-Daten vergleichen kann, um die Zusammenhänge direkt zu sehen. Ist das künstliche neuronale Netz auf Basis dieser Lerndaten optimiert, kann es operativ eingesetzt werden. Wenn diese Möglichkeit des Blicks auf vorab vergangene Daten nicht gegeben ist, also kein Feedback vorliegt, spricht man von Unsupervised Learning. Hier muss sich der Learner die Daten selbst annotieren, indem er Muster findet und Kategorien bildet, also auf Basis der Input-Daten, seine Output-Daten schafft. In diesem Fall ist natürlich ein größerer Datenpool nötig. Bei der dritten Form, dem Reinforced Learning, hingegen, werden die Output-Daten im Prozess evaluiert, um die Mustererkennung zu optimieren. Die drei Verfahren verfolgen unterschiedliche Zwecke und werden in der Praxis gerne auch kombiniert.

Die auf Basis dieser Verfahren entwickelten Anwendungen sind aus unserem Alltag nicht mehr wegzudenken und haben dementsprechend eine nicht zu unterschätzende Wirkung auf unser soziales Leben. Daher ist es unabdingbar, diese Phänomene einer adäquaten soziologischen Betrachtung zu unterziehen.

3. Neuere KI in der sozialwissenschaftlichen Diskussion

Während Machine Learning und dessen (potentielle) Nutzbarmachung für die Sozialwissenschaften immer wieder diskutiert werden (vgl. Molina/Garip 2019; Mökander/Schroeder 2021; Heiberger 2022), scheint uns zunächst insbesondere das Diskursfeld der so genannten Critical Code Studies ein fruchtbarer Nährboden für sozialwissenschaftliche Reflektionen zu sein. Hierin lassen sich, wie wir anhand einiger ausgewählter Studien detaillierter zeigen, fünf zentrale Topoi ausmachen, um die die gegenwärtigen sozialwissenschaftlichen Diskussionen rund um die neue Künstliche Intelligenz kreisen. Im

Einzelnen sind dies: (1) Fragen zu dem Wechselverhältnis zwischen Kontrolle und Machine Learning, (2) die durch den Einsatz künstlicher neuronaler Netze eingeläutete Abkehr von deterministischen bzw. kausalistischen Verfahren, (3) – damit eng verknüpft – die Hinwendung zu Realexperimenten durch die Tatsache, dass es sich um lernende Systeme handelt, (4) die Anzeichen für eine postdigitale Ära sowie (5) die Opakheit von Deep Learning-Verfahren. Diese Topoi werden im Folgenden in der gebotenen Kürze dargelegt.

Zum Wechselverhältnis von Kontrolle und Machine Learning

Einen wichtigen Beitrag zur Frage nach dem Wechselverhältnis von Kontrolle und Machine Learning hat Mackenzie (2017) geleistet. In Anlehnung an Deleuze spricht er von einer Assemblage, die Menschen gemeinsam mit den Machine Learnern bilden. Nur auf der Ebene der Assemblage sei das Phänomen des Lernens adäquat verstehbar. Es – mit anderen Worten – allein auf der Ebene der künstlichen neuronalen Netze ergründen zu wollen, verkennt, wie viel an den Vorgängen in der Trainingsphase dem menschlichen Zutun geschuldet ist. Selbst bei den Schematisierungsleistungen des Machine Learners sind für Mackenzie menschliche Ideen, Gedanken und Empfindungen eingegangen. Es wäre also falsch, anzunehmen, die beteiligten Technologien würden grundlegend anderen Prinzipien folgen als diejenigen der Menschen. Das Gegenteil ist der Fall: Der Machine Learner ist für Mackenzie durch und durch Ausdruck des menschlichen Willens, Wissen und damit Macht zu erlangen. Die Besonderheit dieser Technologien liege darin, dass sie in Bereiche vorzustoßen vermögen, die bislang für den Menschen verschlossen waren. In riesigen Datenmengen spürten sie Muster und Schemata auf und würden damit überhaupt erst komplexe Phänomene kontrollierbar machen, was zu neuen Formen der Akkumulation von Machtwissen führe.

Mackenzie vermutet sogar hinter den Lernprozessen des Machine Learners eine »Technologie des Selbst«; denn die Mustererkennungen und Schematisierungen dienten ja nur dazu, um sich noch besser an die Gegebenheiten der Umwelt anzupassen. Nichts Anderes seien gemäß Foucault »Technologien des Selbst« (1993), die im Rahmen einer Assemblage auch bei den Menschen, die sich auf Deep Learning einlassen, anzutreffen seien: Sich als Subjekt zum Objekt zu machen, Erwartungshaltungen des sozialen und gesellschaftlichen Umfelds so zu internalisieren, dass sie ununterscheidbar von »ur-eigenen« Bedürfnissen und Strukturen werden, kurzum: sowohl als Subjekt als auch als Objekt von Machine Learning in Erscheinung zu treten. Bereits heu-

te ist für Mackenzie sichtbar, dass uns dabei die Machine Learner verändern. Ihre ubiquitären Verwendungsmöglichkeiten führten im Ergebnis dazu, dass es letztendlich ununterscheidbar werde, was auf unseren eigenen Gedanken fuße und was durch Machine Learning »errechnet« sei. Aber genau hierin sei die Wirkmacht der »Technologie des Selbst« zu suchen: Die komplexer werdende Gesellschaft kann für Mackenzie nur gemeistert werden, wenn sich der Mensch auf die Machine Learner einlässt, was zwangsläufig zu einer Veränderung seiner selbst führt. Mackenzie spricht von einer Vorwärtsregelung, die in der Trainingsphase des Learners zu beobachten ist. Hierbei gehe es weniger um Verbesserung der Erfahrung, denn um das tentative Auffinden neuer Relationen, um drängende Realitäten verstehbar und bearbeitbar zu machen.

Dabei steht nach Mackenzie Machine Learning für ein komplett anderes Verständnis, wie Computeroperationen wirkungsvoll zu kontrollieren sind – nämlich jenseits der klassisch-deterministischen Programmierung. Denn Machine Learning sei nur im Horizont einer Krise der Kontrolle, die durch das exponentielle Wachstum des Internets der vergangenen Jahrzehnte verursacht worden sei, zu verstehen. Die operative Macht des Machine Learnings ergibt sich für Mackenzie aus dem Versprechen, im Wildwuchs der Datenströme und der digitalen Kommunikationsprozesse Kontrolle wiederzuerlangen. Dabei sei Machine Learning eine verschachtelte Form von Kalkulation, um zukunftsorientierte Entscheidungen zu ermöglichen.

Mackenzie fordert von den Sozialwissenschaften in diesem Zusammenhang, auf dem gleichen Abstraktionsniveau wie Algorithmen zu forschen, um die Vielfalt der Kategorisierungs- und Klassifikationsformen bei Machine Learning in den Blick zu bekommen. Denn die Differenz der Praktiken wirke sich in der im vorhergehenden Absatz beschriebenen Weise auf das Welt- und Selbstverständnis des Menschen aus. Erst die genaue Kenntnis dieser Praktiken ermögliche es, ihnen kritische Praktiken und Experimente entgegenzusetzen, um die Veränderbarkeit dieser abstrahierenden Prozesse, Kalkulationen und Automationen unter Beweis zu stellen und daran anschließend revolutionäre Institutionen aus Daten und Codes zu schaffen. Denn – gemäß Mackenzies sozialutopischer Vision – ist der Mensch als Element dieser soziotechnischen Assemblage stets in der Lage, das Kollektiv – sprich: die Assemblage selbst – zu verändern.

Ganz im Sinne dieses Assemblage-Gedankens verteilter Kontrollkonstellationen erkundet Engemann (2018) die Rolle des menschlichen Körpers bei ML. Er sei indexikalischer Zeichengeber, der vor allem sich bei der Erstellung der Trainingsdaten und dem sog. Labeling bemerkbar mache. Hier

fungiere der Mensch als Relais, der zwischen Datum, Label und Wirklichkeit einen Zusammenhang herstelle. Im haptischen Sinne erfolge dies durch millionenfaches Clickwork, also den menschlichen Handlungsvollzügen des Bestätigens via Mausclick. Erst dadurch entstehe der Konnex zwischen Zeichen und Referenten, der so bedeutsam für die Güte der Datensätze in Bezug auf die ML-Lernzwecke und damit für die Wirksamkeit des ML-Verfahrens in der Praxis seien. Zwar wird – so Engemann – fieberhaft an unsupervised-Verfahren geforscht, die Realität sieht aber nach wie vor gänzlich anders aus: Nur durch das menschliche Zutun, werden die Verfahren in die Lage versetzt, überhaupt so etwas wie Präzision zu entwickeln.

Abkehr von deterministischen bzw. kausalistischen Verfahren

Ein wichtiger Topos der sozialwissenschaftlichen Befassung mit Machine Learning- und Deep Learning-Verfahren bildet der Aspekt, dass sie weder deduktiv noch induktiv vorgehen. In diesem Sinn bildet etwa für Parisi (2018) Machine Learning eine grundlegend andere Herangehensweise als die bis dahin vorherrschenden algorithmischen Strategien der Erzeugung von Künstlicher Intelligenz. Denn konstitutiv für das »neue« maschinelle Lernen sei, dass das Unvollständige, Unbestimmte, Undeterminierte und damit letztlich Unberechenbare eine konstitutive Rolle für die internen Lernvorgänge spielten. Im Kern geht es – gemäß Parisi – um abduktiv gewonnenes Wissen, wobei abduktive Verfahren der Erkenntnisproduktion nicht etwa im Laufe der Zeit zur Auflösung dieser Unbestimmtheit führt, sondern im Gegenteil zu dessen Bewahrung. Statt wie früher durch Top-Down-Programmierung Ansatzpunkte für maschinelles Lernen zu suchen, seien es nunmehr automatisierte Systeme, die sich bei der »Trial-and-Error-basierten Datengewinnung durch schnelle, unbewusste und nicht-hierarchische Befehle von Entscheidungsprozessen auszeichnen« würden (vgl. ebd.: 99).

Diese nicht-deduktiven Verfahrensweisen öffneten das Tor zu Vorhersagen und Mustererkennungen. Diese Verfahrensweisen bildeten mit statistischen Kalkulationen zusammen eine Synthese, bei der die digitalen Prozeduren lernen, ihre Pfade durch die Datensuche zu verändern. Dabei werde gerade anhand von kontext-spezifischen Dateninhalten und dem Scharfstellen auf deren Besonderheiten sowie auf Überraschungsmomente gelernt. Für Parisi geht es um die Generierung hypothetischer Bedeutung, bei der die Undeterminiertheit ein aktiver Teil der Berechnung geworden ist. Damit lernten die Maschinen das Lernen selbst, als eine Unwissenheit bewahrende Aktivität, welche

weitere Lernkurven ermögliche. Denn die Lückenhaftigkeit der Berechnung eröffne die Welt der Zufälligkeit innerhalb der algorithmischen Abläufe und damit den Weg zu neuen Ideen, Verhalten Mustern und Regeln.

Bächle et al. (2018) verweisen hierbei auf den »Differentiable neural Computer« (DNC) von DeepMind/Google, dessen künstliches neuronales Netz Besonderheiten aufweist, was ihm die Kennzeichnung eines »externalisierten Gedächtnisses« eingebracht hat. So würden Informationen von DNC nicht eindeutig, sondern in Form einer Verteilung gespeichert, auch schwach ausgeprägte Ähnlichkeiten berücksichtigt, aufgrund dieser beiden Besonderheiten konkrete Problemstellungen sowie die Größe des dann faktisch in Aktion tretenden künstlichen neuronalen Netzes festlegen und Inhalte selbst nach ihrer Abspeicherung veränderbar seien – um nur die wichtigsten Besonderheiten zu nennen. Mit diesen Besonderheiten solle das technische System in die Lage versetzt werden, selbsttätig Abwägungen und Beurteilungen vornehmen zu können sowie Informationen kreativ für die Entwicklung eines Lösungspfad für konkrete Probleme zu entwickeln, sich diesen zu merken und für ähnlich gelagerte Problembearbeitungen generalisieren zu können. Darin äußert sich – so die Autor:innen – ein »diagrammatisches Denken« im Sinne Charles S. Peirce³, der diesem Denken spezifische Transfer- und Übersetzungsfähigkeiten zuspricht.

Die Hinwendung zu Realexperimenten

Mit einer Veränderung des logischen Schließens geht auch eine Veränderung des Zugriffs auf die empirische Wirklichkeit einher. Unter dem Topos Realexperimente wird auf den Aspekt der neuen Künstlichen Intelligenz scharfgestellt, dass es sich um lernfähige und damit stets in Optimierungsschleifen befindliche Systeme handelt. Stilgoe (2018) wählt zur Exemplifizierung dieses Sachverhalts das autonome Fahrzeug Modell S von Tesla, welches im Mai 2016 einen tödlichen Unfall produziert hat, da es einen weißen Laster an einer

3 Das Diagramm bildet für Peirce eine besondere Klasse ikonischer Zeichen. Sie weisen nämlich – so Peirce – die nur ihnen zukommende Charakteristik auf, eine »intellektuelle Verbindung zwischen bildlichen und symbolischen Elementen herzustellen. Das, was sie repräsentieren, verdeutlichen sie mit anderen Worten bildlich, aber auch zeichenhaft. Insofern korrespondieren Diagramme für ihn mit dem abduktiven Denken, insofern von einer Beschäftigung mit einem Diagramm mehr gelernt werden kann, als in ihre Verfertigung eingeflossen ist (vgl. Peirce 1976: 47–54).

Straßenkreuzung übersehen hat. Ausgerechnet technische Unfälle ermöglichen, den Forschenden und Entwickelnden die Kontrolle über das Realexperiment zu entziehen und auf die mehr oder weniger impliziten Unsicherheiten scharfzustellen. Denn gerade, weil die neue Künstliche Intelligenz dort eingesetzt werde, wo klassische Programmierung scheitere – nämlich bei komplexen Aufgaben, bei denen es zu viele mögliche Situationen geben könne, als dass diese im klassischen Sinne programmiert werden könnten –, gerade deshalb befänden sich die Machine Learning-Verfahren in einem andauernden »work-in-progress«- bzw. »Beta«-Zustand. Ja, nur der Einsatz in realen Situationen ermöglichte derartigen Systemen, im Hinblick auf die komplexe Aufgabe, angemessen zu lernen.

Für Stilgoe führen die Begriffe »selbstfahrend« oder »autonom« in die Irre. Vielmehr handle es sich um Autos, die lernen würden, zu fahren – und dies mitten unter uns. Insofern sei es nicht verwunderlich, dass die Entwickler:innen von Tesla die Antwort auf den Crash in der Fortsetzung der Forschung an den ML-Verfahren sähen. Jedoch, so Stilgoe, bezieht Tesla damit gleichzeitig eine Position, der gemäß das Lernen dieser Systeme im privaten Hoheitsgebiet von Tesla liegt, was hochproblematisch ist. Vielmehr müsse das Lernen umfassender gedacht werden: im Sinne eines sozialen Lernens, das nicht nur die technischen Systeme betreffe, sodass sie sich besser in einer sozialen Welt zurechtfinden, sondern darüber hinaus wie alle Anspruchsgruppen im Hinblick auf ihren alltäglichen Umgang mit der neuen Technologie und wie die Gesellschaft als Ganzes im Sinne einer Governance neuer Technologien lernten. So gesehen, gäbe es weitere Optionen, produktiv an den Crash anzuknüpfen – etwa mit der Governance-Option, Konzerne wie Tesla dazu zu verpflichten, ihre Daten zu teilen, sodass nicht jedes Unternehmen selbst solche schmerzhaften Erfahrungen machen und Schäden produzieren müsse. Eine Technologieentwicklung, die auf ein so verstandenes soziales Lernen abhebt, ist Stilgoe zufolge resilienter und kann aus Krisen sogar gestärkt hervorgehen. Gerade Machine Learning biete hierzu eine hervorragende Möglichkeit, einen Take off des (sozialen) Lernens zu vollführen: denn prinzipiell brauche nur ein einziges Auto eine bis dahin für das System unbekannt Situation zu lernen, sodass alle lernfähigen Autos über Auto-zu-Auto-Kommunikation (Stichwort: Internet der Dinge) davon lernten. Entsprechend mündet Stilgoes Argumentation in der Forderung nach einem »social ML«, also lernende Systeme in ihre soziale Welt zurückzubetten und Technologieentwicklung als einen ganzheitlichen Prozess des sozialen Lernens zu reframe.

Anzeichen einer postdigitalen Ära

Mindestens ebenso oft wie auf die Betaisierung technischer Systeme wird in der sozialwissenschaftlichen Forschung auf dessen postdigitale Aspekte abgehoben. Prominent vertritt diese These Sudmann (2018), indem er auf das in Deep Learning-Verfahren bedeutsame Backpropagation hinweist, bei dem es im Kern um die Anpassung von Gewichten zwischen Verbindungen im betreffenden künstlichen neuronalen Netz geht. Dabei betont auch Sudmann (wie auch Parisi, s.o.) eine Bottom-up-Strategie in der neueren KI, also auf digitalem Wege ein Gehirn nachzumodellieren (im Unterschied zu der alten KI mit ihren Top-down-Strategien, wie dem Anlegen großer Wissensspeicher). Nunmehr würde die Nachbildung des biologischen Gehirns das vormals vorherrschende kognitivistische Verständnis verdrängen.

Der Fokus auf künstliche neuronale Netze stellt für Sudmann

»in mindestens zweifacher Hinsicht ein Gegenmodell zur Funktionsweise digitaler Computer gemäß der seriell organisierten Von-Neumann-Architektur dar: (1) Die Gewichtung der Aktivität zwischen den Neuronen, d.h. die Stärke ihrer Verbindungen, wird bei neuronalen Netzen zumeist durch Fließkommazahlen, also quasi-analog, repräsentiert. (2) Die massenhaft miteinander verbundenen Neuronen feuern gemeinsam bzw. parallel, und bilden auf diese Weise ein komplexes emergentes System, das letztlich die Diskretheit der Elemente, aus denen es besteht, fundamental aufhebt.« (Ebd.: 66)

Entsprechend bilde die neue KI auch ein Gegenmodell zur Welt der klassischen Algorithmen, in der durch serielles Operieren innerhalb eines wohldefinierten Problemraums alle auftretenden Probleme logisch gelöst werden konnten. Demgegenüber steht die neue KI für ein konnektionistisches Paradigma: In Entsprechung zum grauen Rauschen im Gehirn liegt mit einem künstlichen neuronalen Netzwerk ein »Chaos« an (potentiellen) Verknüpfungen vor, innerhalb dessen Muster und Sinnzusammenhänge als emergente Verkopplungsphänomene erscheinen – so Sudmann. Bedeutung ergebe sich also aus spezifischen Systemzuständen des künstlichen neuronalen Netzes. Dieses Netz »ist demzufolge ein Unschärfesystem mit probabilistischen Resultaten, dessen Operationen eher als analog denn als digital zu beschreiben sind« (ebd.: 67). In letzter Konsequenz erwiesen sich dadurch auch die bisherigen Computer, die weitestgehend nach der Von-Neumann-Architektur aufgebaut sind, als unzu-

reichend, weil sie gerade auf das serielle Operieren spezialisiert seien. Demgegenüber stünden Technologien wie die Tensor Processing Units (TPUs), die Neurocomputer sowie die Quantencomputer für aussichtsreiche Forschungsaktivitäten, die den Weg zum ›analogen Rechnen‹ und damit zu einer postdigitalen Ära bahnten.

Opake Systeme

Schließlich besteht ein weiterer bedeutsamer Topos in der sozialwissenschaftlichen Forschung zur neuen Künstlichen Intelligenz in ihrem opaken, intransparenten Charakter. Gerade bei Deep Learning-Verfahren bleibt es in der Regel unklar, wie sie zu einem Ergebnis gelangen – und dies selbst dann, wenn man die Inputdaten genau kennt (was selbst schon oftmals nicht der Fall ist). Dies ist insofern für die Sozialwissenschaften von besonderer Relevanz, da diese Verfahren zum Beispiel für die Berechnung der Kreditwürdigkeit einer Person eingesetzt werden. D.h. der Einsatz dieser Verfahren zeitigt heute schon erhebliche soziale und gesellschaftliche Konsequenzen, ohne dass deren Ergebnisproduktion verstanden, nachvollziehbar und damit legitimierbar gemacht werden kann. Burrell (2016) differenziert dabei drei Formen der Undurchsichtigkeit dieser Verfahren:

1. Sie kann zum einen daraus resultieren, dass sie einer organisationalen oder staatlichen Geheimhaltung unterliegen. Man denke an NSA oder die großen Plattformkonzerne, deren Geschäftsmodell nicht zuletzt gerade darin besteht, immer ausgeklügeltere Big Data-Analyseverfahren zu entwickeln, um zum Beispiel an tiefere Schichten des menschlichen Verhaltens heranzukommen, um es zu manipulieren.
2. Die Undurchsichtigkeit von Deep Learning-Verfahren kann darüber hinaus auch schlicht daran liegen, dass wir nicht über den nötigen technischen Sachverstand verfügen, um ihre Operationsweise zu verstehen.
3. Undurchsichtigkeit kann aber auch dadurch entstehen, dass die durch diese Verfahren beschrittenen mathematischen Lösungswege in Big Data-Kontexten nichts mehr mit dem zu tun haben, wie menschliches Verstehen vorstangeht. Sie folgen mit anderen Worten ganz anderen Anforderungen als diejenigen, die menschliche Einsichten und semantische Interpretationen ermöglichen.

Unter Rückgriff auf eine Forderung von Pasquale (2015) sieht Burrell die erste Form von Undurchsichtigkeit durchaus als auflösbar an: Denn dass Organisationen und staatliche Institutionen Geheimhaltung in der angesprochenen Form betreiben können, hat ja letztlich mit einem Vakuum an Regulation zu tun. Wären diese Organisationen gesetzlich dazu verpflichtet, ihre Datenanalyseverfahren offenzulegen – etwa durch Etablierung staatlicher Prüfstellen und Auditoren, welche zu weitgehende Manipulationen oder Ausspionierungen unseres Verhaltens ahnden können.

Auch die zweite Form von Undurchsichtigkeit lässt sich gemäß Burrell gesellschaftlich abmildern, wenn nicht gar heilen: Zum einen könnte die Ausbildung und Erziehung einen wesentlich stärker ausgeprägten Fokus auf Computer-Skills, insbesondere Programmierfähigkeiten, für alle legen. Zum anderen könnte man die Programmierer:innen dazu verpflichten, Entschlüsselungshilfen für ihre Codes mitzuliefern, sodass sie nicht nur für Computer lesbar, sondern auch für Menschen nachvollziehbar werden, was unter dem Begriff *explainable AI* verhandelt wird.

Von besonderem forscherschem Interesse ist jedoch die dritte Form von Undurchsichtigkeit für Burrell: Denn selbst wenn die Datensätze, mit denen der Learner trainiert werde, nachvollziehbar und dessen Code klar geschrieben seien, führe das Wechselspiel zwischen beiden zu nicht mehr nachvollziehbaren Komplexitäten. Dies sei bei der automatischen Spam-Erkennung beispielsweise der Fall. Diese operiere dadurch, dass jedes Wort in einer E-Mail daraufhin bemessen werde, inwieweit es mit Spam assoziiert sei. Kleine Unterschiede und Schlüsselwörter bilden – so Burrell – dann den Gradmesser dafür, ob eine E-Mail in den Spam-Ordner wandert oder nicht. Da bei diesem Filterungsprozess keine semiotische, geschweige denn eine semantische, also auf Bedeutungen abzielende Analyse, und schon gar nicht eine Narrationsanalyse zur Anwendung komme, die zum Beispiel Rückschlüsse auf die Intentionen des Verfassers der betreffenden E-Mails vornehmen würde, operiere die automatische Spam-Erkennung keineswegs für Menschen eingängig. Dies werde an der Liste derjenigen Wörter deutlich, die am stärksten mit einer Spam-E-Mail assoziiert seien: »our (0.500810), click (0.464474), remov (0.417698), guarante (0.384834), visit (0.369730), basenumb (0.345389), dollar (0.323674), price (0.268065), will (0.264766), most (0.261475), pleas (0.259571)« (Burrell 2016: 8). Unklar bleibe insbesondere die Gewichtung harmloser Wörter wie »visit« oder »want« als Indikatoren für Spam. Wenn also ein Machine Learning-Verfahren konsequent seine eigene Repräsentation der Klassifikationsentscheidungen aufbaue und Gewichtungen zwischen den

selbstgenerierten Klassifikationen so lange manipulierte, bis die Lerndaten die gewünschten Ergebnisse hervorbrächten, geschehe dies ohne Rücksicht auf die menschlichen Verstehensmöglichkeiten. Deshalb spricht sich Burrell generell dafür aus, Machine Learning nicht bei kritischen Anwendungen zu verwenden, und dort, wo Machine Learning zur Anwendung kommt, stets zu prüfen, ob dadurch unmittelbare oder mittelbare soziale Diskriminierungen aufgelöst werden.

Einen anderen Aspekt von Opakheit thematisieren Bechmann und Bowker (2019), wenn sie auf die klassifikatorische Arbeit rund um die neue KI abheben. Sie differenzieren hierbei drei Schichten dieser Arbeit: (1) Die Datensammlung selbst kann durch Datenklassifikation mitgeprägt sein; (2) bei der Datenreinigung findet offensichtlich eine Klassifikation in brauchbare und nicht-brauchbare Daten statt; sowie (3) das Trainieren des Learners kann ebenfalls als klassifikatorische Arbeit begriffen werden. Bei standardisierten Klassifikationssystemen würden so beispielsweise kulturelle Unterschiede mit zum Teil fatalen Folgen vernachlässigt – etwa, wenn Minderheitskulturen benachteiligt würden; so geschehen bei einer ML-Anwendung zur Bilderkennung, welche People of Color fälschlicherweise als Gorillas klassifiziert habe. Gerade durch die Hidden Layers bei Deep Learning werde es schier unmöglich, die vorgenommenen Kategorisierungen logisch aufzuschließen. Dieses Problem spitzt sich für die Autor:innen bei Unsupervised Machine Learning zu, da dort scheinbar induktive Klassifizierer zur Anwendung kommen. So bilde die hier verwendete Latent Dirichlet Allocation (LDA) ein Unsupervised Machine Learning-Verfahren, das insbesondere für semantische Analysen von Textdaten eingesetzt werde. Genau besehen, stelle sich dieses sogenannte unsupervised Modell als ausgeprägt supervised heraus, da angefangen von der Festlegung der Topics, über die Datenreinigung bis hin zu einem Vorverständnis für als bedeutungsvoll gehaltene Cluster ein hoher Grad menschlicher Kontrolle in das Verfahren Eingang gefunden habe. Durch derartige klassifikatorische Arbeit jedenfalls würden durch Machine Learning-Verfahren regelmäßig falsche Exkludierungen genauso wie falsche Inkludierungen stattfinden – mit den oben beispielhaft näher umrissenen diskriminierenden Konsequenzen für Minderheiten und andere durch Klassifikationen marginalisierte Gruppen.

Dieser Ausblick auf diese fünf zentralen Topoi ist natürlich keineswegs erschöpfend. Auch wäre ein systematisches Anknüpfen an all diese Punkte jenseits der Möglichkeiten eines Sammelbandes. Aus der White'schen Perspektive sind ohne Frage die Aspekte der Kontrolle und der Opakheit das Offensichtlichste. Der Themenbereich um das Verhältnis von Kontrolle und Machi-

ne Learning trägt sogar einen von Whites Grundbegriffen in der Bezeichnung, auch wenn Mackenzie aus einer anderen theoretischen Richtung argumentiert. Die Opakheit der KI-Systeme stellt ein anderes aus soziologischer Perspektive enorm relevantes Phänomen dar. Mit Blick auf die empirischen Gegenstände wird aber auch deutlich, dass der Realexperimentcharakter der Anwendungen nicht zu übersehen ist.

4. Die Beiträge des Bandes

Dementsprechend schließt der Sammelband an einige der vorher skizzierten Problembereiche explizit an und widmet sich anderen eher indirekt. Im ersten Text von Roger Häußling und Marco Schmitt wird für den Gegenstand eine relationale theoretische Perspektive im Anschluss an White entwickelt, einschließlich der Klärung zentraler Begriffe, ihres Zusammenhangs und ihre Funktion im Anwendungskontext der Künstlichen Intelligenz.

Im Anschluss daran fokussiert der Sammelband auf einzelne Themen, die Künstliche Intelligenz als gesellschaftliche Herausforderung betrachten. Zu diesem Zweck ziehen die jeweiligen Autor:innen eine Reihe von zwischen 2019 und 2021 geführten Expert:innen-Interviews mit einschlägigen Forscher:innen an der RWTH Aachen und darüber hinaus für ihre Texte zu rate.

Bei den behandelten speziellen Themen erfolgt der Anschluss insbesondere an die Topoi der Kontrolle und der Opakheit. Ein Text widmet sich der Frage nach den Vorgängen, die sich eher der Beobachtung entziehen und solchen die aktiv ins Licht gerückt werden. Marco Schmitt und Christoph Heckwolf diskutieren in ihrem Text »KI zwischen Blackbox und Transparenz« der wichtige Widerspruch zwischen »blackboxing« und »explainability« als unterschiedliche Kontrollprojekte. Sie unterscheiden im Anschluss an White zwischen Kontrolle, Kontrollversuchen und Kontrollprojekten und zeigen die Transparenzproblematik empirisch auf unterschiedlichen Ebenen (Daten, Trainingssetting, Personen und mathematische Grundlagen) auf. Hierbei greifen sie zur Interpretation auch auf die White'sche Systematik von Getting Aktion und Blocking Aktion zurück, um die Strategien der Forschenden und Entwickelnden zu fassen.

Darüber hinaus nimmt ein Text den Forschungs- und Arbeitsprozess mit diesen Algorithmen in seiner kulturellen Verfasstheit in den Blick. Wie wird hier gelernt bzw. wie wird das Lernen verteilt und wie werden die Vorgehens-

weisen gerahmt. Dieser Frage gehen Claudius Härpfer und Nadine Diefenbach in ihrem Text »Zur Kunst des Lernens« nach, indem sie in Anlehnung an Herbert Simon und White ein Modell künstlichen Lernens entwickeln. Dieses Modell wenden sie auf die beiden Lernformen, das Supervised Learning und Unsupervised Learning, an, um unter Rückgriff auf das Interviewmaterial zu zeigen, wie und unter welchen Voraussetzungen sich Lernstories entwickeln können. Wobei am Ende deutlich wird, wie viel menschliches Bauchgefühl und Hintergrundrauschen in den Lernvorgängen trotz aller Metrisierung vorhanden ist.

Dem folgt die Frage nach der Macht der Daten, also konkret die Frage nach der Kontrolle, die hier ausgeübt wird bei der Formierung der Daten ebenso wie beim Einsatz der Ergebnisse. Um dieser Frage techniksoziologisch gehaltvoll nachzugehen, greifen Philip Roth, Matthias Dorgeist und Astrid Schulz in ihrem Text »Deep Learning Techniken als Boundary Objects zwischen Entwicklungs- und Anwendungsfeld« auf den Begriff der Boundary Objects zurück, als die sie Deep Learning Technologien begreifen. Diese Grenzobjekte haben die Eigenschaft, in Entwicklungsvorgängen strukturierend zu wirken und so das Verhältnis von Entwickler:innen und Anwender:innen dieser Technologien zu beeinflussen. Um aus dieser Perspektive das Wechselspiel von Entwicklung und Anwendung anhand zweier Projekte aus den Lebenswissenschaften zu skizzieren, stützen sie sich auf Feld- und Praxistheoretische Konzepte, um diese im Nachklang am White'schen Ansatz zu messen.

Schließlich werden drei zentrale Anwendungskontexte mit erheblichen gesellschaftlichen Implikationen in den Blick genommen. Hierbei zeichnet sich deutlich insbesondere der Realexperimentcharakter der Machine Learning-Anwendungen ab. Da ist zum einen der Bereich von Medizin und Gesundheit, in dem man sich einerseits erhebliche positive Effekte vom Einsatz künstlicher Intelligenz erhofft, und andererseits gleichzeitig einem Bündel an ethischen und rechtlichen Bedenken gegenübersteht. Dieses Spannungsfeld betrachten Dhenya Schwarz und Tabea Bongert in ihrem Text »Über Identitäten und Selbstverständnis medizinischen Personals in Zeiten künstlicher Intelligenz und Algorithmierung« aus der Anwender:innenperspektive, indem sie das Selbstverständnis von Ärzt:innen und deren Verhältnis zu Patient:innen im sich wandelnden klinischen Arbeitsalltag in den Blick nehmen. Unter Rückgriff auf die Theorie Whites und Arbeiten Andrew Abbotts rekonstruieren sie die Institution Medizin im Spannungsfeld jener stabilisierenden und innovierenden Faktoren und geben mit Hilfe von einigen Interviews Einblicke in den Wandel der Ärzt:innenrolle.

Als zweiter Anwendungskontext dient die Perspektive der technischen Start-up-Kultur, die sich um neue datengetriebene Plattform-Konzepte dreht, ohne die Künstliche Intelligenz in ihrer aktuellen Form nicht möglich wäre und für die Künstliche Intelligenz dementsprechend ein verheißungsvoller Zukunftsmarkt ist. Tim Franke, Niklas Strüver und Sascha Zantis betrachten in ihrem Text »Plattform und jetzt?« die Gründungstories digitaler Start-up-Unternehmen. Hierzu nutzen sie von theoretischer Seite Whites Story-Begriff und Gedankenfiguren des Solutionismus, der sich in jener schnelllebigen Startup Welt als Rechtfertigungsordnung einer Polis der Solution etabliert hat, die Start Up-Gründer:innen dazu bringt, sich den Unsicherheiten und Widrigkeiten einer Gründung auszusetzen. Als empirische Basis dienen acht Interviews, die im Rahmen des BMBF-geförderten Projekts INDIZ erhoben wurden.

Den Abschluss der Aufsätze machen Axel Zweck und Thomas Werner, die in ihrem Beitrag das Thema Künstliche Intelligenz aus Perspektive der Zukunftsforschung beleuchten. In einem Abriss der historischen Entwicklung künstlicher Intelligenz zeigen sie, wie diese ihre Rolle ändert. Vom Beginn als große Utopie hin zur schwachen KI, die Insellösungen bietet und mit der Zeit immer leistungsfähiger wird, um schließlich nicht mehr bloß reines Instrument zu sein, sondern zu einer Art Akteur:in im Forschungsprozess aufzusteigen und die Art zu Forschen verändert. Letzteres gilt insbesondere für die Zukunftsforschung, deren Prognoseverfahren mit den neu entstandenen Möglichkeiten einerseits ausgebaut werden können, andererseits stellt sich die Frage der Qualitätssicherung.

Schließlich greifen die Herausgeber den Faden nochmals auf und ziehen ein systematisierendes Resümee um anschließend – wie bereits erwähnt – den aktuellen Entwicklungen rund um ChatGPT Rechnung zu tragen und den Chatbot selbst in einem Interview zu Wort kommen zu lassen.

Dieser Sammelband ist im Arbeitsbereich des Lehrstuhls für Technik und Organisationssoziologie der RWTH Aachen University aus einem gemeinsamen techniksoziologischen Interesse an den Entwicklungen der neueren KI heraus entstanden. Die ersten, oben erwähnten Interviews wurden in diesem Zuge 2019 geführt. 2021 erschien das erste gemeinsame Working Paper, das versucht hat, die White'sche Perspektive auf ein Recommender-System anzuwenden (Häußling et al. 2021). Im Anschluss daran, nahm dieser Band Form an. Wir danken allen Beteiligten Autor:innen für ihre Diskussionsfreudigkeit und ihre Bereitschaft, tief in ein Thema abzutauchen, das mit dem Tagesgeschäft nicht immer einfach in Einklang zu bringen war. Darüber hinaus un-

terstützten uns bei der Durchführung der Interviews und in der konzeptionellen Phase des Bandes Annika Fohn, Julia Kolb und Jacqueline Lemm. Von Seiten des Verwaltungsteams seien Christa Siebes und Marlon Steffens genannt, die zu unterschiedlichen Zeitpunkten wertvolle Beiträge lieferten. Finja Bersch und Josias Bruderer schließlich gingen uns bei der Formatierung und Finalisierung für den Druck zur Hand.

5. Literatur

- Bächle, Thomas C./Ernst, Christoph/Schröter, Jens/Thimm, Caja (2018): »Selbstlernende autonome Systeme? Medientechnologische und medientheoretische Bedingungen am Beispiel von Alphabets Differentiable Neural Computer (DNC)«, in: Christoph Engemann/Andreas Sudmann (Hg.), *Machine Learning. Medien. Infrastrukturen und Technologien der künstlichen Intelligenz*, Bielefeld: transcript, S. 167–192.
- Bechmann, Anja/Bowker Geoffrey C. (2019): »Unsupervised by any other name: Hidden layers of knowledge production in artificial intelligence on social media«, in: *Big Data & Society*, S. 1–11.
- Beher, Hannah/Braun, Matthias (2023): »Wissen und nicht wissen. ChatGPT & Co und die Reproduktion sozialer Anerkennung«, in: *Forschung & Lehre* 30, S. 260–261.
- Burrell, Jenna (2016): »How the machine ›thinks‹: Understanding opacity in machine learning algorithms«, in: *Big Data & Society* 3, S. 1–12.
- Engemann, Christoph (2018): »Rekursionen über Körper. Machine Learning-Trainingsdatensätze als Arbeit am Index«, in: Christoph Engemann/Andreas Sudmann (Hg.), *Machine Learning. Medien. Infrastrukturen und Technologien der künstlichen Intelligenz*, Bielefeld: transcript, S. 247–268.
- Foucault, Michel, (1993): »Technologien des Selbst«, in: Luther H. Martin/Huck Gutman/Patrick H. Hutton (Hg.), *Technologien des Selbst*, Frankfurt a. M.: S. Fischer, S. 24–62.
- Häußling, Roger/Franke, Tim/Härpfer, Claudius/Roth, Philip/Schmitt, Marco/Strüver, Niklas/Zantis, Sascha (2021): »Mendelian und das Erklärungspotential der Theorie von Identität und Kontrolle. Ein techniksoziologischer Blick auf Recommender-Systeme«. Working Paper des Lehrstuhls für Technik- und Organisationssoziologie 02, Aachen: Lehrstuhl für Technik- und Organisationssoziologie, Institut für Soziologie, RWTH Aachen. [<https://publications.rwth-aachen.de/record/822003/files/822003.pdf>]

- Heiberger, Raphael (2022): »Applying Machine Learning in Sociology: how to Predict Gender and Reveal Research Preferences«, in: *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 74, S. 383–406.
- Mackenzie, Adrian (2017): *Machine Learner. Archaeology of a Data Practice*, Cambridge: MIT.
- Mainzer, Klaus (2019): *Künstliche Intelligenz. Wann übernehmen die Maschinen?*, Berlin: Springer.
- McCarthy, John/Minsky, Marvin/Rochester, Nathaniel/Shannon, Claude E. (2006): »A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. August 31, 1955«, in: *AI Magazine* 27 (4), S. 12–14.
- Mökander, Jakob/Schroeder, Ralph (2021): »AI and social theory«, in: *AI & Society* (2021), <https://doi.org/10.1007/s00146-021-01222-z>
- Molina, Mario/Garip, Filiz (2019): »Machine Learning for Sociology«, in: *Annual Review of Sociology* 45, S. 27–45.
- Nehlsen, Johannes/Fleck, Tilmann (2023): »Zulässige Hilfsmittel für Hochschulprüfungen? Rechtliche Aspekte von ChatGPT«, in: *Forschung und Lehre* 30, S. 262–264.
- Parisi, Luciana (2018): »Das Lernen lernen oder die algorithmische Entdeckung von Informationen«, in: Christoph Engemann/Andreas Sudmann (Hg.), *Machine Learning. Medien. Infrastrukturen und Technologien der künstlichen Intelligenz*, Bielefeld: transcript, S. 93–113.
- Pasquale, Frank (2015): *The Black Box Society: The Secret Algorithms that Control Money and Information*, Cambridge/MA: Harvard University Press.
- Peirce, Charles S. (1976): *The New Elements of Mathematics: Vol. 4: Mathematical Philosophy*, The Hague: Paris Mouton Publishers.
- Rosengrün, Sebastian (2021): *Künstliche Intelligenz. Zur Einführung*, Hamburg: Junius.
- Stilgoe, Jack (2018): »Machine Learning, social learning and the governance of self-driving cars«, in: *Social Studies of Science* 48, S. 25–56.
- Sudmann, Andreas (2018): »Zur Einführung. Medien, Infrastrukturen und Technologien des maschinellen Lernens«, in: Christoph Engemann/Andreas Sudmann (Hg.), *Machine Learning. Medien. Infrastrukturen und Technologien der künstlichen Intelligenz*, Bielefeld: transcript, S. 9–23.
- Zweig, Katharina (2019): *Ein Algorithmus hat kein Taktgefühl. Wo künstliche Intelligenz sich irrt, warum uns das betrifft und was wir dagegen tun können*, München: Heyne.

Mustererkennung in Datennetzwerken

Zu einer Relationalen Soziologie neuer Formen »Künstlicher Intelligenz«

Roger Häußling, Marco Schmitt

1. Einleitung

Die Neue Künstliche Intelligenz – und hier insbesondere Deep Learning-Verfahren – stellt die Gesellschaft und damit auch die Soziologie vor neue Herausforderungen! Diese Verfahren stehen für einen neuen Typ von Technologie: sie sind selbstlernend dank künstlicher neuronaler Netze. Ein Pionier der ersten Stunde, der schon Ende der 1940er Jahre mit solchen neuronalen Netzen experimentierte, war Alan Turing. Er brachte bereits 1950 das Spezifische dieser neuen Technologie auf den Punkt: »Erlernete Prozesse führen nicht zu Ergebnissen, die hundertprozentig sicher sind; wenn sie es täten, könnten sie nicht verlernt werden.« (Turing 1950) Techniksoziolog:innen sprechen von höheren Freiheitsgraden, die dazu führen, dass derartige Technologien zu Interaktionspartner:innen des Menschen und mehr und mehr zu Entscheidungsträger:innen werden. Entsprechend findet man den Einsatz dieser neuen KI bereits heute bei selbstfahrenden Autos, bei der Krebsdiagnostik, im Börsenhandel, im US-amerikanischen Wahlkampf und im so genannten Predictive Policing zur Aufdeckung von kriminellen Netzwerken – um nur einige besonders prominente Anwendungsfälle zu nennen.

Offensichtlich werden immer mehr gesellschaftliche und soziale Entscheidungsprozesse in automatisierte Kalkulationen verlagert. Der große Vorteil dieser Technologie, Muster jenseits menschlicher Wahrnehmungsmöglichkeiten in riesigen Datensätzen zu erkennen, bildet gleichzeitig auch das Problematische an der neuen KI: Wie und warum sie zu bestimmten Ergebnissen gelangt, bleibt oftmals unklar. Dieser Intransparenz steht eine normative Macht des scheinbar objektiv Berechneten gegenüber. Sich gegen

eine KI-basierte, medizinische Diagnose beispielsweise zu entscheiden und stattdessen auf menschliche Erfahrung der entsprechenden Expert:in zu setzen, wird immer schwieriger.

Aber der Einfluss dieser neuen KI ist bereits heute noch weitreichender, was man an den so genannten Recommendation-Systemen der globalen Handelsplattformen des Internets verdeutlichen kann: Bedürfnisse und Interessen werden von automatisierten Kalkulationen unterlaufen, sodass am Ende nicht mehr unterscheidbar ist, was auf ureigene Intentionen und was auf subtil agierende KI fußt.

Als Techniksoziolog:in fragt man sich aber auch, warum ausgerechnet jetzt und ausgerechnet in einer Gesellschaft, wie der unsrigen, so sehr die Verheißungen der neuen KI auf offene Ohren stoßen. Eine mögliche Antwort könnte darin liegen, dass mit dem Internet und all unseren smarten mobilen Geräten eine Datenflut einsetzte, die gleichzeitig für eine Krise der Kontrolle steht (vgl. auch Mackenzie 2006). Machine Learning verspricht hier eine Wiedergewinnung der Kontrolle, aber – so die kritische Frage – zu welchem Preis und wer kontrolliert wie wen? Ein anderer Pionier der Künstlichen Intelligenz, Norbert Wiener, hat hierzu die denkwürdigen Worte formuliert: »Wenn wir Maschinen programmieren, um einen Krieg zu gewinnen, müssen wir gut nachdenken, was wir mit ›gewinnen‹ meinen.« (Wiener 1952)

Wenig soziologische Theorieangebote scheinen geeignet zu sein, die Besonderheiten des Phänomens Deep Learning sachadäquat erfassen zu können und für soziologische und gesellschaftstheoretische Perspektiven fruchtbar zu machen. Eine Ausnahme bildet Harrison Whites Theorie von Identität und Kontrolle (White 1992 und 2008). Im vorliegenden Beitrag soll aufgezeigt werden, dass zentrale soziologische Aspekte von KI durch diese Theorieangebote sehr grundlegend beschrieben und erforscht werden können. Hierzu soll Problem- und Gegenstands-zentriert vorgegangen und die bisherigen soziologischen Antworten auf zentrale soziale und gesellschaftliche Herausforderungen durch die neue KI vergleichend dargelegt werden. Die gute Passung zwischen dem Gegenstand der neuen KI und einer an White orientierten Relationalen Soziologie resultiert daraus, dass sowohl das soziologische Theorieangebot von White, als auch die einflussreichen Ansätze des Deep Learnings die Mustererkennung als zentralen Aspekt ansehen. Hierbei soll der vorliegende Beitrag einerseits die relationalen Konzepte am Phänomenbereich selbst schärfen und andererseits ihre Reichweite für die Erklärung und Interpretation der gesellschaftlichen und sozialen Dimensionen der neuen KI ausloten.

Ein relationaler Ansatz impliziert eine Perspektive, die stets die Einbettungskonstellationen miterfasst. Insofern ist KI nicht nur technisch zu begreifen, sondern immer auch sozial – und zwar sowohl was die Herstellung der sog. Learner anbelangt, als auch die Trainingsphase – insbesondere, wenn sie supervised ist – wie auch schließlich in der Anwendungsphase eines trainierten künstlichen neuronalen Netzes, bei denen automatisierte Kalkulationen auf soziale und gesellschaftliche Entscheidungsprobleme treffen.

Zunächst starten wir mit dem Problem des *Kontextwechsels* (oder *switching* bei White), das auch avancierte Verfahren künstlicher Intelligenz immer wieder vor große Probleme stellt. Hieran wollen wir deutlich machen, welchen Beitrag das theoretische Modell von Harrison White bei der sozialwissenschaftlichen Betrachtung der neueren Entwicklungen auf dem Gebiet der Künstlichen Intelligenz leisten kann. Im Anschluss sollen dann aber auch andere Konzepte aus der White'schen Theorie eingebracht werden, die ein großes Potenzial für die Analyse und Untersuchung der gesellschaftlich relevanten Aspekte des Phänomens der neuen KI besitzen. Dabei ist vor allem auf die Begriffe *Netdom* und *Stil* einzugehen, die beide Möglichkeiten bieten, Kontexte in einer Weise zu typisieren, die Übergänge einfacher markieren bzw. deutlich machen, welche konkrete Änderungen zu Kontextwechsel führen. Schließlich arbeitet sich der Beitrag zu den Grundbegriffen *Kontrolle* und *Identität* vor, an denen sich dann auch gleich zahlreiche gesellschaftlich relevante Problemlagen entzünden. So sind die Kontrollverstrickungen in verschiedene Richtungen verfolgbar, weil es zum einen um die Kontrolle über Lern-Daten ebenso geht, wie um die Kontrolle der Ergebnisse, welche Phänomene überhaupt von den Verfahren kontrolliert werden sollen, wie dies passiert und an welchen Stellen dann wiederum die Kontrolle anderer Akteur:innen eingebaut wird. Gleichzeitig muss auch auf die Konstruktion sozialer Einheiten scharf gestellt werden, die zunehmend unter Mitarbeit von Verfahren Künstlicher Intelligenz erzeugt werden oder diese selbst beschreiben. Schließlich bietet der Vorschlag von White interessante Möglichkeiten, auch mit den unterschiedlichen Skalierungsproblemen umzugehen, die ebenso für KI-Verfahren relevant sind, deren Reichweite sich verändert.

Diese theoretischen Vorschläge einer soziologischen Grundierung des Phänomens der neueren KI als erfolgreiche Transformation des technologischen Feldes einerseits und als die Gesellschaft insgesamt transformierende Entwicklung andererseits, ermöglichen eine Analyse der Veränderungspotenziale unterschiedlicher gesellschaftlicher Domänen und der sie tragenden Prozesse.

2. Zu den soziologischen Herausforderungen der neuen KI

Machine Learning ist – so zeigt es auch der gegenwärtige sozialwissenschaftliche Diskurs (siehe Einleitung in diesem Band) – ohne Einbeziehung seines soziotechnischen Umfelds nicht sinnfällig soziologisch analysierbar. Dabei sticht ein soziotechnischer Aspekt besonders ins Auge: Die digitale Datenfülle, welche die Bedingung der Möglichkeit für Machine Learning-Prozeduren überhaupt erst schafft. Wenn man die Geschichte der Digitalisierung bzw. präziser die Geschichte der Computerisierung¹ in den Blick nimmt, kann man für die ersten Jahrzehnte das komplette Gegenteil diagnostizieren: Es herrschte zunächst ein Datenmangel vor – Großrechner etwa, die ganze Hallen füllten, und doch nur ein Bruchteil der Rechenleistung besaßen wie ein herkömmliches Smart Phone heute. Und als dann die ›Datenlast‹ zunahm, etwa mit dem Aufstieg des kommerziellen Internets, geriet die sich immer weiter digitalisierende Gesellschaft mehrmals in ein Load Faktor-Dilemma (vgl. Hughes 1979): zu viel Daten für zu wenig Rechenleistung. WWW wurde damals ironischer- und zugleich treffenderweise als »word wide waiting« übersetzt. Dass heute große Serverfabriken den gesamten Traffic des Internets abzusaugen in der Lage sind, konnte man damals kaum erahnen, geschweige denn, dass man in quasi Echtzeit Berechnungen vornimmt, um aus diesen Milliarden von Daten Muster in dynamischen Prozessen zu identifizieren.

Von Manovich (2001: 223) stammt das Postulat, dass Datenumfang und Komplexität der diese Daten bearbeitenden Softwaresysteme in einem engen Zusammenhang stehen, nämlich dergestalt, dass je weniger Daten vorhanden sind, die Softwareprogramme aus umso mehr Programmzeilen bestehen. Wohingegen komplexe Datenbanken einfache Algorithmen nach sich ziehen. Und tatsächlich bestehen die Learner von ML-Verfahren oftmals nur aus ein paar hundert Zeilen, wohingegen die Datenmengen, auf die sie zugreifen, riesig sind. Doch dazu gleich ausführlicher! Für Manovich ist das Internet – und aus heutiger Warte sollte man ergänzen, der allumfängliche Einsatz von

1 An anderer Stelle wurde der Vorschlag unterbreitet (vgl. Häußling 2019: 325f.), zwischen Digitalisierung und digitaler Technologie zu unterscheiden. Dies erlaubt, besser auf technologische Entwicklungen der vergangenen rund 20 Jahre scharf zustellen, die nicht zuletzt dafür verantwortlich sind, dass der Begriff Digitalisierung zu einem Buzzword wurde. Digitale Technologie gibt es hingegen schon seit geraumer Zeit, etwa wenn man an Zuses Z3 denkt.

Smart Devices – eine Sollbruchstelle für das Selbstverständnis der Programmierung und des Datenhandlings in der digitalen Sphäre. Denn das Internet »vergleichültige« alles: Eine Internetseite kann sich beliebig aus Text-, Bild- und/oder Videodateien zusammensetzen. Sie kann sich auch permanent in ihrer Ausgestaltung ändern und wird auch zunehmend auf die einzelnen Bedürfnisse jeder Nutzer:in hin individualisiert. Mit anderen Worten: Aus einer schier unendlich anmutenden Menge vorliegender Daten setze sich eine solche Internetseite kontingent zusammen. Sie sei also nichts anderes als eine Ansammlung ebenso heterogener wie austauschbarer Elemente. Letztere seien Computerfiles, die stets ediert werden können und weniger auf Vollständigkeit denn auf Erweiterbarkeit und partielle Ersetzbarkeit angelegt seien. Kurzum: Für Manovich ist jede Internetseite ein in-Erscheinung-tretender Ausschnitt aus Datenbanken und das Web insgesamt ein sich fortwährend wandelnder Datenkorpus (vgl. ebd.: 225). Ja, das Internet dürfte sogar das komplexeste Datenbankphänomen unserer heutigen Zeit sein. Dies wird – wie gesagt – durch den massen- und flächendeckenden Einsatz von Smart Devices nur noch auf die Spitze getrieben, da unzählige situierte Daten durch diese Geräte anfallen, die für unterschiedlichste Apps, für Profiling-Zwecke und für vieles andere mehr verwendet werden. Auch hier steht die Datenfülle im Vordergrund. Selbst vermeintliche Datenabgase können für unerwartete Zwecke verwendet werden. Etwa: Wenn Smartphones Profile des Gehens anlegen und aus diesen anfallenden Daten vor jeder möglichen ärztlichen Diagnostik eruiert werden kann, ob eine Parkinson-Erkrankung bei der Person vorliegt, die das Smartphone mit sich führt (Belic et al. 2019). An diesem Beispiel wird auch deutlich, in welchem umfassenden Ausmaß von den anfallenden Daten aus gedacht und engineert wird.

Entsprechend liegt bei Machine Learning-Verfahren das Hauptaugenmerk auf der Güte der Daten. Sie werden unentwegt generiert, gereinigt, organisiert und indiziert. Für Manovich besitzen Daten Objektcharakter, wohingegen Algorithmen für (kausalistische) Prozesse stehen. In vielerlei Hinsicht könnten Computerspiele als ein Paradebeispiel des alten algorithmusgetriebenen Paradigmas der digitalen Sphäre gewertet werden. Sie folgten einer narrativen Logik, bei der der Reiz am Spiel nicht zuletzt darin besteht, die Logik des Spielalgorithmus zu entschlüsseln. Eine Internetseite folge demgegenüber einer komplett anderen Anordnungslogik der Elemente. Hier werden instantan beliebig austauschbare Elemente zu einem Ganzen zusammengefügt. Das Computermodell hierfür ist demzufolge ein gänzlich anderes als beim Computerspiel: Geht es bei letzterem um die Erzeugung eines direktionalen Handlungs-

stroms, bei dem das gerade Ablaufende als Wirkung des Vorausgegangenen inszeniert wird, geht es bei der Internetseite um die Herstellung eines beliebig zusammenstellbaren Contents. Datenbanken repräsentieren für Manovich die Welt als nicht-hierarchische, oder sonst wie vorab geordnete Liste von Werten, die beliebig arrangiert werden können. Wenn man so will, koexistiert alles in Datenform auf gleicher Augenhöhe: Belangloses und Wichtiges, Bedrohliches genauso wie Rettendes, Wahrheiten und Lügen, Reales und Fiktionales etc.

Als Objekt rücken Datenbanken für Manovich auch immer stärker in das Zentrum der kreativen Bemühungen, der digitalen Sphäre Neues hinzuzufügen. So sind eine Vielzahl von Apps eines Smartphones nur deshalb im Einsatz, weil sie um die anfallenden GPS- oder Bewegungsdaten herum entwickelt wurden. Ohne diese Daten würde keine Navigation glücken, oder uns dazu anleiten, auf den gesundheitlichen Sollwert an Schritten pro Tag zu gelangen. Datenbanken sind in der digitalen Sphäre nicht nur das wichtigste Gut geworden, sondern, neue digitale Dienste zu entwickeln, bedeutet im Kern, eine Schnittstelle zu Datenbanken zu konstruieren. Die unterschiedlichsten Trajektorien können aus ein und derselben Datenbank emergieren, die – gemäß Manovich – wie Hypernarrative funktionieren. Ganz so, wie bei einem chronologischen Verlauf der Internetnutzung einer Person, würde die kontingente Nutzung einer beliebigen Datenbank sein eigenes Ready Made-Narrativ erzeugen. Dieses Vorgehen gleicht der Montagetechnik in der Bildenden Kunst, bei der Photoschnipsel mit gemalten Elementen und/oder Alltagsobjekten ›zusammengewürfelt‹ werden.

In Weiterführung dieses Gedankens von Manovich ist es nicht verwunderlich, dass sich gerade in dieser Phase der Digitalisierung, bei der alles datenmäßig vergleichgültigt wird, das Phänomen der Filterblasen und Echokammern breitmacht. Selbst unter Rückgriff auf ein und dieselbe Datenbank lassen sich nämlich antagonistische Ready-mades erzeugen. Echokammern machen aber einen Prozess sichtbar, den Manovich in seiner Argumentation weniger stark akzentuiert: nämlich nicht nur die aktive Verknüpfung von Elementen ist bedeutsam für die post-narrative, datenbankgetriebene Phase der Digitalisierung, sondern auch das Abschotten und Kappen möglicher Verknüpfungen. Genau das leisten eben Machine Learning-Verfahren auch. Sie sind im Kern Erzeuger von Netzwerken, die neben Konnektivitäten eben auch strukturelle Löcher aufweisen. Und das ist nicht nur in dem Sinne zu verstehen, dass sie künstliche neuronale Netze in den Hidden Layers entstehen lassen, die wie große Filter und Absorber funktionieren. Sondern es ist auch auf einem abstrakten datentheoretischen Sinn zutreffend, dass nur eine

bestimmte Auswahl an Daten relationiert und von dem Rest der Datenwelt abgeschottet wird. Aus der Netzwerkforschung ist bekannt, dass nicht nur das Vorhandensein einer Relation aussagekräftig ist, sondern eben auch das Nicht-Vorhandensein einer Relation. Um genau diese inkludierende und gleichzeitig exkludierende Grenzarbeit geht es beim Netzwerkknüpfen der Machine Learning-Verfahren.

Ein schier unendliches Meer an Daten liegt ohne jegliche prästabilisierte Ordnung vor. Diese Daten stehen für nichts Geringeres als die Welt selbst, also alles, was wir denken, fühlen und spüren können. Genau diese Welt ist zerlegt in ein Chaos – im doppelten Wortsinn – gleichgültiger Letztelemente – Bits und Bytes. Diese von Daten repräsentierte und gleichzeitig vollzogene Welt kennt keine Tiefe mehr: Alles ist auf der gleichen Oberfläche, keine Hierarchie von oben und unten, keine Himmelsrichtungen, die uns einen Weg vorgeben könnten. Aus diesem Fundus aus Allem lassen sich nun beliebige Assemblagen bzw. Netzwerke formieren, die nicht nur Reales in beliebigen Auswahlen repräsentieren, sondern dieses Reale auch gleichzeitig formieren. Auf einer sehr abstrakten Ebene ist genau das die Funktion von Machine Learnern. In Anlehnung an Galloway und Thacker könnte man Machine Learner als Protokolle bezeichnen, welche die Bedingungen der Möglichkeit für Sinnerzeugung und Kontrolle in der digitalen Sphäre schaffen. Vielleicht trifft der Begriff lokales Programm die Funktionsweise von Machine Learning-Verfahren noch besser. Denn der Ausschnitt an Daten, den ein Machine Learning-Verfahren bearbeitet, ist trotz des Begriffs Big Data und in ironischer Abkehr zu diesem Begriff verschwindend gering angesichts der in Daten zerlegten Welt. Lokal meint hier natürlich nicht eine geographische Gegebenheit, sondern eine gegenstandsgetriebene Auswahl im Heterogenen – Aufnahmen von krebskranken und nicht-krebskranken Gehirnen im Unterschied zu allen Daten, die für anderes stehen, etwa. Relationiert und entkoppelt wird hier nicht nur auf der harten Ebene dichotomer Unterschiede (wie etwa krank versus gesund), die bei supervised-Verfahren in der Lernphase von Menschen vorgegeben und in der Anwendungsphase eigentätig vom Machine Learning-Verfahren vollzogen werden. In den Hidden Layers werden kleinste Details relationiert, wie etwa bestimmte Schattierungen an einer bestimmten Stelle des Bildes vom Gehirn, und andere Details diskriminiert. Diese kleinsten Details sind ihrerseits wiederum in Bits und Bytes zerlegt, die unzweideutig von ihren benachbarten Bits und Bytes getrennt sind. Selbst also auf der Ebene der digitalen Elementarteilchen findet nichts Anderes statt als ein Knüpfen von Netz-

werken, wenn man ein solches Knüpfen immer auch als ein Kappen anderer möglicher Relationen begreift.

Lokal sind diese transklassischen Programme deshalb, weil sie keinem vorab existierenden Narrativ oder keiner vorab als gültig ausgewiesenen Theorie folgen, sondern sich als versuchsweise einsetzbare Filter im Meer der Daten zeigen, deren emergente Ergebnisse dann darüber entscheiden, ob sie brauchbar sind oder nicht. Deepl.com verdeutlicht augenscheinlich, dass es keiner präzisen Theorie der Übersetzung bedarf – die es trotz unzähliger Bemühungen seitens der Linguistik und Sprachwissenschaften auch nicht gibt –, um überraschend gute Übersetzungen mittels Machine Learning-Verfahren zu generieren. Die Filterung setzt bereits da an, wo es um die Fütterung des Learners mit präzisen Exemplaren der Ziele der zu entwickelnden maschinellen Diagnostik geht; also Beispiele für gute Übersetzungen im gerade benannten Fall. Noch abstrakter betrachtet, handelt es sich um Verfahren, die tentativ Verknüpfungen und Abkopplungen im opaken Meer der Daten vornehmen und bei denen das evolutionäre Prinzip, Survival of the Fittest im Sinne eines good enough bzw. fit enough (Simon 1990), vollumfänglich umgesetzt worden ist.

Um dieses Weben und Umweben von Datennetzwerken zu realisieren, bedarf es einer anderen Strategie als diejenige, die sich in den klassischen Algorithmen manifestiert. Letztere waren nichts Anderes als riesige logische Verzweigungsbäume, die kausale Routen durch den Code legen; Routen im Sinne von linearen Narrativen, um den Gedanken von Manovich noch einmal aufzugreifen. Ein Netzwerk ist alles andere als linear zu begreifen. Rückkopplungen, Schleifen, Interferenzen sind vielmehr seine Markenzeichen. Insofern sind die Resultate der Machine Learner im Grunde keine Codes im klassischen Sinne mehr. Deshalb sprechen wir von transklassischen Programmen. Sie nähern sich nicht nur – wie weiter oben dargelegt – nicht-binären, analogen Herangehensweisen, sie versuchen darüber hinaus das in eine Sprache zu überführen, was der Logik genau dieser Sprache zu tiefst zu wieder läuft. Der Output des Learners ist ein in Codes überführtes künstliches neuronales Netz. Backpropagation im oben dargelegten Sinn gehört also genauso zum Markenzeichen von Machine Learning-Verfahren, wie das, was Sudman das »Unschärfesystem mit probabilistischen Resultaten« (s.o.) genannt hat. Gleichsam ein emergent aus Konnektionen und Abkopplungen resultierender Filter, ein Netzwerk, das wie ein Fischernetz bestimmte Dinge auffängt und andere durch seine Maschen gleiten lässt.

Insofern findet eine Umkehrung dessen statt, was als mittelbar und was als unmittelbar wirksam wahrgenommen wird. War es früher (also in Daten-armen Zeiten) der Algorithmus, der wahrnehmbar seine Wirkungsketten entfaltete, wohingegen die Daten und Datenbanken nur latent als ›Futter‹ für den Algorithmus aufgefasst wurden, kehrt sich das Verhältnis nunmehr um: Die Datenbank als digitale Versammlung von Welt bildet den wirklichen Bestand, der von flüchtigen und unscheinbaren Codes gefiltert wird. Und es kann in gleichsam beliebiger Form gefiltert werden. Ein Zusammenhang, eine Bedeutung ist dabei ein Resultat des konnektionistischen Paradigmas und nicht mehr länger eine in den Code hineingelegte Prämisse.

Eine solche transklassische Programmierung im Sinne eines soziotechnischen Tatbestands stellt die Gesellschaft und damit auch die Sozialwissenschaften vor grundlegend neue Herausforderungen. Sie sind nicht nur weit in die Wirklichkeit hineingreifende soziotechnische Phänomene, sie stehen auch für ein neues Selbst- und Weltverständnis des Menschen, das sich deutlich von den bis dahin gültigen Verständnissen unterscheidet. Entsprechend ist es dringend geboten, ein genaues Bild dieser Machine Learning-Verfahren zu gewinnen. Gerade die elaborierte netzwerktheoretische Begrifflichkeit Harrison Whites erscheint uns als besonders geeignet, auf Spezifika dieser soziotechnischen Verfahren scharf zu stellen, um ein genaueres Bild von ihnen techniksoziologisch zu zeichnen. Genau diesem Unterfangen widmen sich die folgenden Unterkapitel.

3. Herausforderungen der neuen KI und der Beitrag der soziologischen Netzwerktheorie von Harrison White

Der soeben dargelegte konnektionistische Ansatz von Machine Learning-Verfahren steht in unseren Augen – wie so manches mehr in unserer gegenwärtigen Zeit – für ein Denken und Agieren gemäß einer genuinen Netzwerklogik. Der Begriff künstliche neuronale Netze sagt hierzu im Grunde schon alles. Um Relationen und deren Gewichtungen geht es im Backpropagation – der zentralen Prozedur des Deep Learnings, wie dargelegt. Ein Learner ist eine Maschine, die Netzwerke erzeugt. Regelmäßigkeiten, Strukturen und Muster sind die einzig verbliebenen Orientierungsgrößen, um sich in einer Welt des Heterogenen zurechtzufinden, in der es im Grunde nur Einbettungskonstellationen gibt – also weiter diversifizierende Netzwerkkonstellationen. Diese essentiellen Eigenschaften von Machine Learning-Verfahren verweisen auf

das ihnen zugrundeliegende Netzwerk-Paradigma. Insofern sind wir davon überzeugt, dass allen voran netzwerkforscherische Ansätze zu einer sachadäquaten sozialwissenschaftlichen Erschließung der neuen KI beitragen können. Was liegt also näher, den bislang elaboriertesten Ansatz der Netzwerktheorie und -analyse, nämlich denjenigen von Harrison White, heranzuziehen, und daraufhin zu überprüfen, welches Erkenntnispotenzial in ihm steckt? Genau dies soll nun im Folgenden geschehen. Hierbei stellen wir zentrale Herausforderungen der Machine Learning-Verfahren in den Fokus, verdeutlichen, inwiefern diese jeweiligen Herausforderungen Herausforderungen sind, die sich aus der den Verfahren innewohnenden Netzwerklogik ergeben, und konfrontieren diese mit den theoretischen Einsichten Whites, die er – völlig unabhängig von Überlegungen zu KI – für die strukturelle Bestimmung von Netzwerken im Allgemeinen vorgenommen hat. Dabei ist es schon Whites Ausgangspunkt der Theoriebildung den Ausgangspunkten der neueren KI-Entwicklungen verwandt. White geht von zwei Grundphänomenen aus, die Mustererkennung zentral stellen, Selbstähnlichkeit und Verteilung. Wiederholungen ähnlicher Verteilungen, also Muster, sind der Ausgangspunkt seiner Konzeptbildung für soziale Strukturen.

Das Problem des Kontextwechsels in der neuen KI als Switching in soziotechnischen Netzwerken

Mit dem Wechsel von Repräsentationsmodellen von Wissen, wie sie die Expertensysteme der klassischen KI kennzeichnen, hin zu Lernmodellen auf der Basis von Mustererkennung (zunehmend angelehnt an die Netzwerkstrukturen von Gehirnen) hat sich auch die Anforderung an diese ML-Verfahren grundlegend verändert. Da sie in der Lage sind, zu lernen, also aus der Beobachtung von Regelmäßigkeiten Entscheidungsregeln abzuleiten, entsteht die Hoffnung einer erhöhten Anpassungsfähigkeit dieser ML-Verfahren. Dies wird einerseits als wichtiger, wenn nicht gar zentraler, Schritt auf dem Weg zu einer Allgemeinen Künstlichen Intelligenz (AIG) angesehen, führt aber zur Flexibilisierung der Anwendungsgebiete, wie wenn das AlphaGo nicht nur Schach und Go spielt, sondern in kurzer Zeit alle möglichen Brettspiele erfolgreich spielen kann; oder wenn ein ML-Verfahren für Autonomes Fahren unterschiedlichste Merkmale auf der Fahrbahn erkennen kann und daraus Entscheidungen ableitet. Dennoch bleiben die Anwendungsgebiete noch immer sehr spezialisiert. Selbst ein Wechsel zwischen unterschiedlichen Kulturkontexten kann diese ML-Verfahren immer noch vor große Probleme

stellen. Damit ist das Problem des Kontextwechsels direkt angesprochen, dass auch in sozialen Situationen einschlägig ist und verschiedene Dimensionen des Sozialen tangiert. Hier ist auch der Hinweis auf die Wirksamkeit der Daten kennzeichnend, denn die Daten sind häufig spezifisch und, um besseres Lernen zu ermöglichen, auf einen Kontext zugeschnitten. Für den Wechsel müssen zusätzliche Muster gelernt werden, die den Prozess des Wechsels steuern. Der abduktive Sprung, der häufig bei neueren KI-Ansätzen unterstellt wird, bleibt häufig noch aus, wobei im Spezialfall durchaus überraschende Ergebnisse erzielt werden können – besonders wenn die Struktur der Daten nicht bekannt ist.

In der Netzwerktheorie ist das gerade dargelegte Problem ebenfalls bedeutsam, weil Netzwerke zunächst als grenzenlos betrachtet werden können. Fragt man nur nach Beziehungen im Allgemeinen, dehnen sich Netzwerke immer weiter aus, ohne wirklich abgeschlossen zu werden. Zur Untersuchung von Netzwerken müssen hier jedoch Grenzen eingeführt werden. Nach White geschieht dies vor allem durch Festlegung der Beziehungstypen (types of ties) und Identitätsformen. Dabei unterscheidet er verschiedene Komplexitätsstufen solcher Identitäten, die zwar in spezifischen sozialen Situationen entstehen, aber durch Erzählungen in andere Kontexte wandern können. Deshalb sind diese Grenzen in ständiger Bewegung, wenn man sich soziale Situationen ansieht. Daher führt White im Nachgang zur ersten Auflage von *Identity & Control* den Begriff des Switching (Kontextwechsels) ein (White 1995), der deutlich machen soll, dass es kommunikative Marker für Übergänge zwischen Kontexten gibt, die dann jeweils einen Kontextwechsel einerseits ankündigen, ihn damit wahrscheinlicher machen und auf diese Weise Veränderungen an den beteiligten Netzwerken, Beziehungs- und Identitätsformen vornehmen. Gleichzeitig – und dies ist für White sehr wichtig – werden damit diese Kontexte auch aneinandergeschnitten, d.h. man kann der Spur solcher Wechsel folgen und so komplexere Identitäten erkennen und auch Spillover-Effekte (Padgett & Powell 2013)² nachvollziehbar machen. Die Modellie-

-
- 2 Unterscheidet man verschiedene Netzwerkdomänen, kann es durch die über Personen oder andere Knoten vermittelten Beziehungen zwischen diesen Netzwerkdomänen zu einem Spillover kommen. D.h. nach Padgett und Powell, dass Veränderungen in einer Netzwerkdomäne auch zu Änderungen in anderen Netzwerkdomänen führen können, etwa wenn es einer Person gelingt, einen Zugewinn an Zentralität in der Familie zu nutzen, um auch im Netzwerk von politischen Ämtern mehr Gewicht zu erhalten. Ähnliches kann jedoch auch für die Diffusion von Ideen angenommen werden.

rung des Problems ist allerdings nicht ganz einfach, da ein ML-Verfahren die Kontextmarker ebenso erkennen müsste, wie auch die für den jeweiligen Kontext zentralen Muster. Dies gelingt auch menschlichen Akteur:innen nur mit entsprechendem Kontextwissen, welches erlaubt, Marker und Muster leicht zu verbinden; oder durch Rückfallpositionen, die von den Akteur:innen eingenommen werden, wenn der unterstellte Kontext offensichtlich nicht passt. White geht davon aus, dass Sprache hier eine zentrale Rolle spielt, ebenso wie Stile, die ebenfalls Marker und Muster koppeln und so dabei helfen, Kontexte schnell zu differenzieren. Andererseits hebt White hervor, dass es gerade auch eine Stärke sozialer Situationen ist, hier klare Differenzierungen zumindest eine Zeit lang im Ungefähren zu lassen (Mische/White 1998), sodass Optionen offengehalten werden und man nicht zu schnell auf klare Muster scharf stellt, die unangebracht sein könnten.³ Diese Vermeidung des Scharfstellens ist dabei als eine der großen Herausforderungen für KI-basierte ML-Verfahren zu sehen, die eine Allgemeine Künstliche Intelligenz anstreben. Der stetige Umgang mit Kontingenzen und eine Offenheit für vage Konstellationen stellt diese ML-Verfahren immer noch vor große Probleme. Dies liegt auch daran, dass ihre Aufgabenstellung bislang häufig anders aussieht, weil sie Klarheiten produzieren sollen. Techniksoziologisch ist an der Herangehensweise von White allgemein zu beachten, dass auch Materialitäten als solche Wechselmarker dienen können. Häufig sind es Räume oder Gegenstände, die, ähnlich wie Sprache, solche Wechsel nahelegen. Das Konzept des Switchings (und auch des Stitchings)⁴ könnte also auch in anderen Analysen von Technik Verwendung finden, die sich mit Wandel durch die Einführung neuer Artefakte in bestehenden soziotechnische Netzwerke beschäftigen, oder dem Wandern von unterschiedlichen technologischen Entwicklungen in neue Netzwerkkonstellationen.

Dabei tritt die technologische Neuartigkeit der neuen KI gerade am Switching-Thema deutlich zu Tage: Wie beispielsweise Halfmann (1996) herausarbeitet, ist ein durchgängiges Kennzeichen »klassischer« Technik, dass

-
- 3 Hier sehen wir eine Parallele zu dem Unschärfeaspekt bei der neuen KI, den Sudmann (2018) herausarbeitet.
- 4 White erwähnt diese Stitchings als Kehrseite seines Switchingsbegriff im Interview in Schmitt & Fuhse 2015 und meint damit, dass ein Kontextwechsel immer auch eine verfolgbare Spur zwischen den Kontexten hinterlässt, etwa in der Form von am Gespräch beteiligten Personen, oder einem Streit, der von einem Thema zum nächsten springt.

sie weitgehend dekontextualisiert funktioniert. Nur elementare ›harte‹ Kontexte, wie derjenige der Schwerkraft zum Beispiel, sowie ebenfalls ›harte‹ infrastrukturelle Kontexte, wie etwa die Stromversorgung bei Maschinen, bilden Bedingungen der Möglichkeit eines Funktionierens. Für Soziolog:innen besonders bedeutsam, sei jedoch Technik gerade von der sozialen Kontextspezifität weitgehend entbettet. Ein Wecker beispielsweise funktioniert unerbittlich auf jedem Kontinent, in jedem sozialen Milieu und in jeder denkbaren, gerade ablaufenden Interaktionssituation. Diese Entkopplung vom Mitteilungsaspekt, wie Halfmann in Anlehnung an den Kommunikationsbegriff Luhmanns es nennt, ist nicht zuletzt dafür verantwortlich, dass alles Mögliche dann je nach konkreter Situation in die Technik hineinprojiziert werden kann: Das Auto als Fortbewegungsmittel, als Statussymbol, als Schlafstätte, als Schadstoffquelle etc. Mit der neuen KI bekommen wir es quasi zum ersten Mal mit einer Technik zu tun, die in Umkehrung zur ›klassischen‹ Technik einen genuin kontextsensitiven Fokus aufweist. Auch hierin äußert sich also ihr transklassischer Charakter.

Das Problem der Grenzziehung durch neue KI als Phänomene der Bildung von Nettdoms und Stilen

Das Problem der Grenzziehung kann man direkt an das Problem des Kontextwechsels anschließen, da es auch hier um die Klärung von Kontext geht. In KI-Modellen, die auf Deep Learning bauen, sind diese Grenzen zunächst durch die Trainingsdaten gesetzt. Die Trainingsdaten sind letztlich der Kontext, auf dessen Basis das ML-Verfahren lernt. Danach wird dieser Kontextrahmen jedoch erweitert, da das lernende Netzwerk nun auf neue Daten losgelassen wird. Interessant ist hier auch wiederum die Vorgehensweise beim AlphaGo-ML-Verfahren, welches die zu lernenden Spiele einfach immer wieder gegen sich selbst spielt und dabei einen riesigen Korpus an Vergleichsdaten erzeugt, indem es das Spiel simuliert. Die Grenzen sind hier jedoch durch die Spielregeln der verschiedenen Brettspiele klar definiert. Ein solches Selbstlernen durch Simulation ist also in gewisser Weise vor allem dann erfolgreich, wenn die Grenzen klar bestimmbar sind, oder mit Winograd/Flores (1989) formuliert, wenn es sich um einen wohldefinierten Problemraum handelt. Es geht hier also um Kontrollgewinne, die auf der Möglichkeit der ML-Verfahren beruhen in umgrenzten sozialen Bereichen selber weitere passende Daten zu generieren und so quasi experimentell lernen. Daten werden im Verlauf von diesen ML-Verfahren immer weiter erzeugt und teilweise

gehört es zur Selbstverbesserung dieser ML-Verfahren, mit Simulationen zu arbeiten, um auf mehr Daten zu Lernzwecken zurückgreifen zu können. In vielen sozialen Kontexten ist diese Klarheit jedoch nicht gegeben und so geraten die lernenden Verfahren an Grenzen, wenn ihnen keine klaren Grenzen ihres Anwendungshorizonts vorgegeben wurden bzw. sie keine eindeutigen Grenzen aus dem Datenmaterial herausfiltern können.

Dies ist nun ein Problem, mit dem sich auch menschliche Akteur:innen herumschlagen müssen, und damit ein Phänomen, dem sich auch alle soziologischen Differenzierungstheorien widmen. In der relationalen Theorie Whites spricht man seit der Neuauflage von *Identity & Control* (2008) von Netzwerkdomänen, oder kurz: Netdoms. Ein Netdom kombiniert spezifische Typen von Verbindungen und zugehörige Typen von Identitäten (auf einer kulturell-definitiven Ebene) sowie eine ebenfalls spezifische Netzwerkstruktur (auf einer klassisch netzwerkanalytischen Ebene). Ein Beispiel für eine Netzwerkdomäne wäre ein Freundschaftsnetzwerk innerhalb einer Firma, das man unterscheiden kann von einem Ratgeber-Netzwerk etc. Der Begriff bietet, wie die meisten Begriffe bei White, eine große Anwendungsbreite und ist praktisch beliebig skalierbar. Er kombiniert kulturelle Schablonen mit sich dynamisch aufbauenden Beziehungsstrukturen. Schwierig ist auch hier die Erkennbarkeit, da Netdoms nicht immer durch klar abgrenzbare Begrifflichkeiten vorstrukturiert sind. Eine weitere Begrifflichkeit Whites kann hier Abhilfe schaffen, die sich explizit an die Erkennbarkeitsproblematik koppelt: Stile.

Stile kennzeichnen nach White flexible Sozialstrukturen (ebenfalls skalierbar und mit extensiver Anwendbarkeit), die genau auf der Basis wiederkehrender, selbstähnlicher Muster Signalfunktion und Sensibilisierung verknüpfen. In einem Stil – entwickelt von White am Beispiel von Kunststilen (White/White 1993) – wird ein selbstähnliches Muster benutzt, um eine komplexe Identität zu signalisieren, die eine trainierte Beobachter:in (Expert:in) schon an kleinen Ausschnitten des Musters erkennen kann. KI-Verfahren könnte man an dieser Stelle als trainierte Beobachter ansehen, die genau aus solchen Daten Profile erstellen, um dann auf Identitäten zurückzurechnen. Problematisch ist hieran, dass diese Grenzkonstruktionen der KI kombiniert mit den Schwierigkeiten solcher Verfahren mit dem Wechseln von Kontexten (s.o.) zu einer problematischen Verhärtung von Mustern und damit sozialen Identitäten führen kann, die in anderen sozialen Situationen für gewöhnlich länger offengehalten würden. Diese vorschnellen Abgrenzungen könnten dann durchaus sozial schwerwiegende Folgen haben. Es geht hier tatsächlich um die Übergabe von Kontrolle bei der Identitätskonstruktion von Personen und Gruppen

an Technik und damit auch um einen Kontrollverlust dieser sozialen Entitäten bei den entsprechenden Grenzziehungen. Stile halten diese Verdichtungen auch länger offen, indem sie zwar aus Ausschnitten Strukturen erkennbar machen, aber auch hier Flexibilität und Möglichkeiten zum aktiven Stillbruch implizieren. Da Stile als Signale verwendet werden (auch durchaus bewusst), muss man hier auch die Möglichkeiten des strategischen Einsatzes von Markern einkalkulieren.

Techniksoziologisch ist hier zu beachten, dass Grenzziehungen für den Einsatz von Technologien extrem wichtig sind. Häufig geht es darum, das Funktionieren durch Ausgrenzung von Kontexten und Komplexität sicher zu stellen. Hier kann der Bezug auf den Stilbegriff extrem hilfreich werden. Wie sich Identitätskonstruktionen in soziotechnischen Netzwerken ändern, ist häufig genau an diesen Ausgrenzungen besonders gut zu erkennen, etwa auch an der transklassischen Form neuerer KI-Verfahren, die eben eine Kontextsensibilität markieren sollen und daher die klassische Grenzziehung von Technik als entbettet verlassen. Markierungen dieser Grenzen durch Räume, Einrichtungen und Interfaces sind daher von großem Interesse und können bei der Analyse von Transformationsprozessen helfen.

Das Problem der Steuerung von und mittels neuer KI als soziotechnische Kontrollprojekte

Das Problem der Steuerung schwebt über der gesamten Diskussion um den Einsatz künstlicher Intelligenz, vor allem, wenn man von ethischer, rechtlicher und politischer Seite auf das Thema schaut. Aber auch ökonomisch ist die Steuerung von bestimmten Abläufen, vor allem die Geschwindigkeit dieser Steuerung ein großes Thema. Die Frage ist hier stets, wer steuert wen und an welcher Stelle werden dann Entscheidungen von erheblicher Tragweite getroffen. Hier werden gleich mehrere Themenkomplexe aus der Literatur angesprochen. Zum einen natürlich das Verhältnis von Kontrolle und Opakheit der ML-Verfahren (s. Einleitung im vorliegenden Band). Dabei ist KI, wie oben schon erwähnt, zum einen als Antwort auf immer komplexere Entscheidungssituationen zu verstehen, wo die Datenlage für menschliche Beobachter zu unübersichtlich wird. Kontrolle wird hier abgegeben, weil man Kontrolle schon verloren hat, oder anders formuliert: es geht darum Kontrolle durch technische Unterstützung erst überhaupt wieder zurückzugewinnen. Dabei ändern sich jedoch die Bedingungen der Ausübung von Kontrolle, weil die Empfehlungen der KI zunehmend nicht mehr hinterfragt werden können, oder man die Ant-

wort auf diese Frage nur sehr bedingt versteht bzw. nachvollziehen kann. Besonders problematisch wird dieser Kontrollverlust natürlich vor dem Hintergrund der ebenfalls diskutierten Betaisierung dieser ML-Verfahren (s. ebenfalls Einleitung). Wenn Empfehlungen aufgrund von Verfahren getroffen werden, deren Datengrundlagen eventuell noch nicht ausreichen, stellt das die Benutzer:innen vor neue Herausforderungen.

Wenn man Steuerungsversuche als Kontrollprojekte versteht und sich die Frage stellt, wo sich welche Kontrollansätze wie miteinander verknüpfen, bietet sich der Kontrollbegriff von Harrison White an, der Kontrolle zur Grundlegung jeder Art von Relationierung macht. Beziehungen zwischen Entitäten sind in dieser Sichtweise Kontrollverschränkungen, die hier unterschiedliche Bindungen stabilisieren, von denen Steuerung als sehr direkte Durchgriffsmöglichkeit nur eine Möglichkeit von vielen Kontrollmöglichkeiten darstellt. Kontrollprojekte lassen sich hierbei natürlich auf unterschiedlichen Ebenen identifizieren, da auch Erzählungen (oder Stories bei White) als solche Kontrollprojekte eingesetzt werden können.

Daher spielt die Debatte um Steuerung durch die neue KI eine wichtige Rolle, da hier der Möglichkeitsraum für den Einsatz und die Reichweite dieser Verfahren unterschiedlichen soziotechnischen Konstellationen ausgearbeitet wird. Dabei wird vor allem relevant, wie technologische Aspekte in soziale Prozesse eingebettet werden und auf welcher Ebene der Kontrollversuch unternommen wird. Neben der Kontrolle von Entscheidungen auf der Output-Ebene, die verschiedene Formen zwischen Empfehlungen und direkten Eingriffen annehmen kann, gibt es auch eine Kontrolle auf der Input-Ebene, bei der zum einen Daten ausgewählt werden können, von denen dann gelernt wird, und zum anderen auch durch Einstellungen der Parameter in den neuronalen Netzen (Stichwort: Backpropagation). Dies geht bis hin zu einer nachträglichen Kontrolle der Ergebnisse im Hinblick auf Kriterien die man erfüllt sehen möchte, zum Beispiel im Rahmen einer Qualitätskontrolle.

In dieser Komplexität der Kontrollverschränkungen im Einsatz von Machine- und Deep Learning-Verfahrenen zeigt sich dann auch wieder das Problem der Opakheit bzw. der mangelnden Transparenz dieser Verfahren. Es wird zu Teilen schwierig für jemanden sein, herauszufinden, warum ihm die KI einen spezifischen Vorschlag macht. An dieser Stelle wird dann auch die Verantwortungsproblematik deutlich, da Kontrollverschränkungen gekoppelt mit Intransparenz dazu führen, dass Resultate solcher soziotechnischen Netzwerke letztlich nicht mehr klar zurechenbar sind und daher Regeln dafür getroffen werden müssen, wie sich die Verantwortung im Netzwerk verteilt.

Unter Netzwerkbedingungen – netzwerktheoretisch gesprochen – ist sowieso die klassische Vorstellung von Verantwortung abzulehnen und zugunsten einer dynamisch-rückgekoppelten Sichtweise auf mögliche Quellen des Bewirkens und tatsächlich eintretender Wirkungen zu ersetzen (vgl. hierzu auch die »Agency of Assemblages« bei Bennett (2005)). Zentraler Bezugspunkt ist nunmehr die Frage nach der Bildung welcher Relationen auf der Basis welcher Kontrollprojekte, die schließlich das gesamte soziotechnische Netzwerk auszeichnen. Diese Verstrickung von Kontrollprojekten und deren Vermittlung über technische Artefakte kann als ein zentrales Thema jeder Techniksoziologie herausgestellt werden. Wenn Kontrollprojekte materialisierte Formen annehmen, kommt es damit zu einer spezifischen Verhärtung. Dies bedeutet nicht, dass es nicht Möglichkeiten gibt, diese Verhärtung zu umgehen und/oder sie für andere Formen von Kontrolle zu nutzen, es bedeutet aber, dass die Kontrolle hier eine andere Ausdrucksform angenommen hat, die sich durch veränderte Merkmale (Dauerhaftigkeit, Widerstand gegen andere Kontrollversuche etc.) auszeichnet.

Das Problem der Zuschreibung innerhalb der und durch die neue KI als relationale Identitätsbildungsprozesse

Schließlich ändern sich auch die Identitätsbildungsprozesse von Personen und Organisationen, wenn Algorithmen wesentliche Beiträge hierzu leisten. Dies kann man schon an der zunehmenden Dominanz von Rankingsystemen bei der Bewertung dieser Identitäten sehen (Espeland/Sauder 2007; Sauder/Espeland 2009). Hier kommt es auch, wie im chinesischen Social Scoring-System, oder auch nur bei der Bewertung als Top Rezensent bei Amazon, zu einer letztlich stark datengetriebenen Identitätskonstruktion, bei der man sich nicht mehr an direkten Beobachtungen oder Hörensagen (Stories) orientiert, sondern an abgeleiteten Bewertungssystemen, bei denen man weder genau weiß, auf welche Daten sie sich stützen, noch wie der Algorithmus daraus eine finale numerische oder andere Bewertung generiert. Durch die Vermischung unterschiedlicher Identitätsfacetten in den Datengrundlagen können dann auch neue Informationen gewonnen werden, die in der Art eines Profiling über diverse Daten hinweg operieren und damit eine stärker integrierte Identität erzeugen, als es die Person selbst eventuell tun würde. Es kommt hier dann zu Realexperimenten mit Personen und ihrer jeweiligen Identität, über deren Konstruktion sie ein weiteres Stück ihrer Hoheit verlieren. Tatsächlich kommt es hier dann durch die KI-Verfahren auch zu abduktiven Sprüngen,

etwa wenn jemandem aufgrund spezifischer, ihr oder ihm vielleicht nicht mal bekannter Eigenschaftskonstellationen ein Kredit verweigert wird. Dies ermöglicht zum einen forensische Ansätze der Datenanalyse, also eine überzeugende Rekonstruktion von Identitäten aus verteilten Spuren (hierbei kann auch das Identitätskonzept von White (s.o.) helfen, siehe außerdem McFarland et al. 2016), zum anderen aber auch die Erzeugung einer sensiblen Maschine, die auf bestimmte Spurenelemente besonders stark reagiert. So lassen sich aus unterschiedlichen Datenkorpora von Positionen (wo war eine Person, oder welche anderen Personen hat sie getroffen), über Erzählungen (was hat sie dabei gesagt oder geschrieben) bis hin zu komplexen Spuren der Bewegung zwischen unterschiedlichen Kontexten eine umfassende Identitätskonstruktion ableiten. Hier erscheint dann auch die Kontrollproblematik in neuem Licht; denn wer hat hier jetzt noch welche Kontrolle über die so entstehenden Identitäten. Die freie Verknüpfung unterschiedlicher Datenströme ermöglicht dabei Konstruktionen, die der Person oder der sozialen Gruppe, auf die sie sich beziehen, gar nicht zugänglich sind. Die Daten werden hierbei in einer Weise wirksam, die ganz neue, sozial hoch relevante Zuschreibungsprozesse möglich macht und diese der Kontrolle entzieht.

Der Ansatz von White, soziale Identitäten auf unterschiedlichen Komplexitätsformen zu beschreiben, ist geeignet, um hier die entsprechenden datengetriebenen Varianten zu analysieren. Die Theorie bietet dabei Anschlussmöglichkeiten, da sie klar zwischen positionalen und narrativen Elementen unterscheiden kann, diese jedoch in Beziehung zueinander setzt. Damit ist die Relationierung zwischen unterschiedlichen Datenmaterialien schon integriert. White unterscheidet prinzipiell fünf Formen von Identität mit steigender Komplexität und größerer Reichweite. Während Identitäten sich situativ durch eine wechselseitige Positionierung bilden und damit die einfachste Form von Identität realisieren, können diese Positionen in einer zweiten Stufe in der Form einfache Narrative transportierbar gemacht werden. Diese Möglichkeit kann dann eine Positionierung über mehrere Kontexte hinweg bewirken; und dieser Weg von Positionen durch verschiedene Kontexte ist dann selbst auch eine Identität auf der dritten Komplexitätsstufe. Diese Formen können ihrerseits wiederum in der Form von Geschichten verdichtet werden (z.B. als Karriere) und bilden erneut eine transportable Form von Identität. Schließlich unterscheidet White auch noch eine fünfte Komplexitätsform, bei der es um eine Integrationsleistung geht, die in der Lage ist, alle diese Identitätsformen in einem übergreifenden Profil zu integrieren. Dies gelingt zum Beispiel in der Form der Person. Allgemeiner veranschlagt White

hier den Begriff des Stils, der ein stochastisches Profil beschreibt, dass in der Lage ist, solche Integrationsleistungen zu vollbringen.

Für die Analyse der gesellschaftlichen Auswirkungen neuerer Formen der KI sind dabei wichtig, welche Arten von Spuren sie nutzt, was sie daraus konstruiert und was davon wieder sichtbar gemacht wird, um etwa als transportable Einheit benutzt zu werden. Dabei ist bedeutsam, dass die KI-Verfahren auch Positionierungen eigentätig vornehmen, selbst Positionen einnehmen und Verdichtungen anbieten. Damit verschieben sich Kontrollverstrickungen und bestehende Positionen von Identitäten, wie z.B. professionelle Identitäten von Ärzt:innen etc. (siehe den Beitrag von Bongert & Schwarz im vorliegenden Sammelband). KI-Verfahren sind dabei in diese Kontrollprojekte unterschiedlicher, sozial relevanter Identitäten involviert und bestimmen diese zu weiten Teilen mit. So sind mittlerweile Kaufprofile fester Bestandteil sozial relevanter Identitäten, die aggregiert und über Datenhändler weiterverkauft werden. Solche Einheiten bieten jeweils auch neue Kontrollchancen, die man nutzen kann. Grundsätzlich ist in Anlehnung an Whites Theorie davon auszugehen, dass neue Verdichtungen von stochastischen Prozessen, auf welcher Grundlage sie auch immer vorgenommen werden, immer auch neue Wege der Kontrolle etablieren und sich so mit anderen Identitäten verstricken und damit, streng relational gedacht, diese verändern. Ist das Kaufprofil einer Person verfügbar und mit anderen Daten verknüpfbar, ergeben sich neue Wege diese Person anzusprechen, oder mit diesem Profil zu arbeiten. Diese datengetriebene Konstruktion von Identitäten ermöglicht also neue Arten von Kontrolle, indem diese Daten in unterschiedliche soziale Praktiken und Selbst- und Fremdbeschreibungen integriert werden.

Auch an dieser Stelle kann die Grundidee der Rolle von Erzählungen und Zuschreibungsprozessen bei der Konstruktion sozialer Relevanz auf andere Technologien ausgeweitet werden. Diese Technologien gehen immer in Identitätskonstruktionen ein, da man keine Kranführer:in ohne Kran sein kann. Andererseits werden Technologien jedoch auch immer selbst Ziel von Erzählungen und ihnen werden spezifische soziale Relevanzen zugeschrieben, wie man es in bedeutendem Umfang auch im Diskurs um die neuen DL-Verfahren beobachten kann.

Die Skalierungsproblematik in der KI als Reichweitenfestlegung in Einbettungskonstellationen

Schließlich müssen wir uns der Frage der Skalierung dieser KI-Verfahren zuwenden. Skalierung kann hier in zwei Richtungen gelesen werden: Zum einen geht es um die Reichweite der angebotenen Lösungen zwischen Spezialproblemen und einer allgemeinen künstlichen Intelligenz (AIG), zum anderen geht es dabei um immer umfassendere Datensätze oder gesellschaftliche Anwendungsgebiete, die für Speziallösungen zur Verfügung stehen. Die Reichweite ist dabei eine wichtige Bezugsgröße, da White ein Angebot macht, zu erklären, über welche Wege die Skalierung sozialer Formationen erfolgt. Situative stochastische Verteilungen verdichten sich zunächst noch in einer sozialen Situation zu Positionierungen, die sprachlich zu erzählerischen Einheiten gebündelt und damit in andere situative Kontexte überführt werden können. Gelingt mittels dieser Überführung eine Verknüpfung von Kontexten, ergibt sich eine kohärente soziale Spur, die selbst wieder zu einer Einheit verdichtet werden kann. Durch solche gelingenden Verdichtungen mit Verknüpfungen ergeben sich Reichweitengewinne.

Schaut man sich jetzt die Anwendungsfelder der neueren KI an, wird schnell deutlich, dass sie zwar vielfältige neue Einheiten produziert, also Verdichtungen vornimmt, die Verknüpfung allerdings noch häufig außerhalb ihrer Aufgabengebiete liegt. Dennoch existiert schon eine Reihe von Versuchen, auch die Verknüpfung zu automatisieren und so große Reichweiten produzieren zu können. An den beiden einleitend erwähnten Reichweitenproblemen von spezifisch zu allgemein und von Nischenanwendungen zu umfassenden Anwendungen ist jeweils eine spezifische Form dieser Kopplung von Verdichtung und Verknüpfung zu beobachten. Diese Reichweitenvergrößerung lässt sich ebenfalls häufig in der Form von Realexperimenten beobachten. Die ML-Verfahren laufen einfach und lernen (etwa AlphaZero auf immer mehr Arten von Spielen) und es wird beobachtet, wie weit sie damit kommen, immer neue Daten zu integrieren oder selbst zu erzeugen.

Aus netzwerktheoretischer Sicht ist für den ersten soeben dargelegten Fall vor allem das Einbettungsproblem ein aussichtsreicher Erklärungsansatz. Eine Einbettung wird nicht ohne weiteres aus Lerndaten direkt ersichtlich und sie ersetzt häufig Informationen, die nicht in diesen Daten auftauchen. Damit ist im Anschluss an die Diskussion zur Identitäts- und Kontrollproblematik zu beachten, dass solche Verfahren immer in schon existierende soziale Kontexte eingeführt werden. Es gibt also schon vorgängige Verstrickungen, die in unter-

schiedlichen Ausmaßen in die Spezialanwendung integriert sind bzw. deren Funktionsfähigkeit bedingen. Hierbei ist jeweils zu bedenken, in welcher Weise die Lerndaten diese Kontexte spiegeln, oder integrieren können. Die vielfältigen Diskussionen um die Probleme von Bias oder das Scheitern von Anwendungen bei leicht veränderten oder neuen Kontexten deuten bereits auf das Problem hin, dass die Einbettung in der soziotechnischen Konstruktion nicht ausreichend reflektiert worden ist. Hieraus ergibt sich dann auch für die zweite Reichweitenproblematik eine konstitutive Vertrauensproblematik, für die Lösungen angeboten werden müssen. Weiterverwendbare Einheitenbildung muss ihre Fähigkeit zur Verknüpfung diverser Kontexte erst zeigen und nachvollziehbar unter Beweis stellen, um größere gesellschaftliche Reichweite zu erlangen.

Reichweite und Skalierbarkeit sind in der Techniksoziologie sehr grundlegende Themen, wenn es zum Beispiel um die Diffusion von Innovationen geht, oder wenn sich ganze gesellschaftliche Formationen auf bestimmte technologische Grundlagen beziehen, die etwa den Energiehaushalt einer Gesellschaft (Kohle, Öl, Gas etc.) bestimmen. Whites Netzwerktheorie bietet hier ein offenes Skalierungsmodell mit mehreren Konzepten, mit denen ein solcher Reichweitengewinn beschrieben und erklärt werden kann. Festzuhalten bleibt hier, dass sich auch technisch realisierte Kontrollprojekte gegenüber anderen Projekten behaupten müssen, genauer: sich mit diesen verknüpfen, oder von ihnen ablösen müssen, um ihre Reichweite auszudehnen und dabei durch Switching und Stitching neue Kontexte erobern.

4. Schluss

Die neuen ML- und DL-Verfahren, die der Erforschung künstlicher Intelligenz einen neuen Aufschwung beschert haben, sind durch ihre gesellschaftliche Relevanz ebenso Thema für die Soziologie, wie auch durch ihren transklassischen Charakter eine theoretische Herausforderung. Auf diese theoretische Herausforderung sollte in diesem Beitrag scharf gestellt werden. Dabei sollte deutlich geworden sein, dass netzwerktheoretische Konzepte, wie die hier vorgestellten aus der Theorie von Harrison White sehr geeignet sind, um mit diesen theoretischen Herausforderungen umzugehen. Dies liegt daran, dass die Konzepte zwei Merkmale aufweisen, die eine gewisse Verwandtschaft zu den Mustererkennungsprozessen der DL-Verfahren, die hier im Fokus stehen, aufweisen. Das erste Merkmal ist die explizite Inkorporation von Unsicher-

heit und Chaos in die Konzepte selbst. Alle Konzepte weisen Unschärfen auf, die Momente von Chaos und Ordnung zusammen denken. An jeder Stelle können Muster nur gehalten werden, wenn Kontrolle sich verschränkt und gerade aus dieser Verschränkung ergeben sich stets neue Lücken, aus denen Veränderungen erwachsen können. Das zweite Merkmal bezieht sich auf die Skalierbarkeit der Konzepte. Die Konzepte bezeichnen Muster, die man auf unterschiedlichen sozialen Skalen beobachten können. Ein explorativer Stil kann etwa in einer Einzelsituation als Strategie einer Teilnehmer:in beobachtet werden, aber auch als Stil einer Person über verschiedene Situationen hinweg und als Stil einer Organisation, einer wissenschaftlichen Produktionsgemeinschaft, oder einer ganzen Nation. Zusätzlich haben wir diese Wahlverwandtschaft zwischen techniksoziologischem Anwendungs- und Theoriebezug auch genutzt, um das Theorieangebot von White techniksoziologisch in ersten Schritten zu erweitern. Die Perspektive von Identität und Kontrolle bringt hier Möglichkeiten ein, von soziotechnischen Netzwerken nicht nur als Metapher zu sprechen, sondern mit dem umfangreichen Arsenal netzwerkanalytischer Methoden zu arbeiten.

5. Literatur

- Belić, Minja/Bobić, Vladislava/Badža, Milica/Šolaja, Nikola/Đurić-Jovičić, Milica/Kostić, Vladimir S. (2019): »Artificial intelligence for assisting diagnostics and assessment of Parkinson's disease – A review«, in: *Clinical neurology and neurosurgery* 184.
- Bennett, Jane (2005): »The Agency of Assemblages and the North American Blackout«, in: *Public Culture* 17(3), S. 445–466.
- Espeland, Wendy N./Sauder, Michael (2007): »Rankings and reactivity: How public measures recreate social worlds«, in: *American journal of sociology* 113(1), S. 1–40.
- Halfmann, Jost (1996): *Die gesellschaftliche »Natur« der Technik: eine Einführung in die soziologische Theorie der Technik*, Opladen: Leske+Budrich.
- Häußling, Roger (2019): *Techniksoziologie*, Opladen/Toronto: utb.
- Hughes, Thomas P. (1979): »The electrification of America: the system builders«, in: *Technology and Culture* 20(1), S. 124–161.
- Mackenzie, Adrian (2006): *Cutting code: Software and sociality* (Vol. 30), New York: Peter Lang.

- Manovich, Lev (2001): *The language of new media*, Cambridge/Massachusetts: MIT Press.
- McFarland, Daniel A./Lewis, Kevin/Goldberg, Amir (2016): »Sociology in the era of big data: The ascent of forensic social science«, in: *The American Sociologist* 47, S. 12–35.
- Mische, Ann/White, Harrison (1998): »Between conversation and situation: Public switching dynamics across network domains«, in: *Social research* 85, S. 695–724.
- Padgett, John F./Powell, Walter W. (2012): *The emergence of organizations and markets*, Princeton/Oxford: Princeton University Press.
- Sauder, Micheal/Espeland, Wendy N. (2009): »The discipline of rankings: Tight coupling and organizational change«, in: *American sociological review* 74(1), S. 63–82.
- Schmitt, Marco/Fuhse, Jan (2015): *Zur Aktualität von Harrison White: Einführung in sein Werk*, Wiesbaden: Springer-Verlag.
- Simon, Herbert A. (1990): »Bounded Rationality«, in: John Eatwell/Murray Milgate/Peter Newman (Hg.), *Utility and Probability*, London: Palgrave Macmillan, S. 15–18.
- Sudmann, Andreas (2018): »On the media-political dimension of artificial intelligence«, in: *Digital Culture & Society* 4(1), S. 181–200.
- Turing, Alan M. (2007): »Computermaschinerie und Intelligenz (1950)«, in: Karin Bruns/Ramón Reichert (Hg.), *Reader Neue Medien: Texte zur digitalen Kultur und Kommunikation*, Bielefeld: transcript Verlag, S. 37–64.
- White, Harrison C. (2008): *Identity and Control. How social formations emerge*. Second edition, New Jersey: Princeton University Press.
- White, Harrison C. (1992): *Identity and Control. A Structural Theory of Social Action*, Princeton/New Jersey: Princeton University Press.
- White, Harrison C./White, Cynthia A. (1993): *Canvases and careers: Institutional change in the French painting world*, Chicago: University of Chicago Press.
- White, Harrison C. (1995): »Network switchings and Bayesian forks: reconstructing the social and behavioral sciences«, in: *Social Research* 62 (4), S. 1035–1063.
- Wiener, Norbert (1952): *Mensch und Menschmaschine*, Berlin: Metzner.
- Winograd, Terry/Flores, Fernando (1989): *Erkenntnis-Maschinen-Verstehen: zur Neugestaltung von Computersystemen*, Berlin: Rotbuch.

KI zwischen Blackbox und Transparenz

Das Koppeln und Entkoppeln von Kontrollprojekten

Marco Schmitt, Christoph Heckwolf

1. Einleitung

Der zunehmende Erfolg und Einfluss von KI-Verfahren, die auf Mechanismen aus dem Deep Learning setzen, also auf informatische Modelle, die als neurale Netzwerke dem neuronalen Netz des menschlichen Gehirns nachempfunden sind, hat zu einer stärkeren Diskussion über die damit verbundenen Transparenzprobleme geführt. Kontrolle von gesellschaftlichen Entscheidungsprozessen setzt auf deren transparenter Nachvollziehbarkeit. Dies gilt auch und vor allem hinsichtlich der Prüfung von Technologien. Verfahren künstlicher Intelligenz entwickeln aber ihre eigenen Strukturen der Mustererkennung in einem über Lerndaten realisierten Prozess, auf dem schließlich Entscheidungen basieren. In diesem Prozess kann sowohl über die Daten, als auch über die Konstruktion der Algorithmen eine gewisse Kontrolle über das entstehende Netzwerk der Mustererkennung ausgeübt werden, der Prozess selbst ist aber nicht kontrollierbar. Um die Qualität der Ergebnisse eines solchen Prozesses zu prüfen, kommen dann Hilfsstrategien zum Einsatz, die zusätzliche Informationen hinsichtlich der Richtigkeit der Ergebnisse geben sollen. Der Beitrag soll der Frage nachgehen, inwiefern für diesen Vorgang der Begriff der Blackbox in Anschlag gebracht und welche Kontrollstrategien in auf KI-Verfahren basierenden soziotechnischen Settings dahinterstehen. Eine Blackbox erfüllt ihren Auftrag bzw. ihre Funktion, ohne dass es für einen Beobachter von außen nachvollziehbar wäre, was da im »Innern« passiert (Latour 1994). Kontrollmöglichkeiten werden eingeschränkt (Latour/Woolgar 1986). Dabei geht es zum einen um die Freistellung von strukturellen Kontrollbeziehungen, also den Eingriffen in den ablaufenden Prozess, zum anderen aber auch um eine rhetorische Kontrollstrategie, die hier von der »Last der

Transparenz« befreit. Wenn auch für Konstrukteur:in oder Nutzer:in nicht im Einzelnen nachvollziehbar ist, wie die KI sich selbst strukturiert, dann befreit sie das auch von der Möglichkeit von Nachfragen oder einer auf den Prozess gerichteten Kritik. Hier findet sich dann häufiger eine Gegenkontrollstrategie, die darauf gerichtet ist hier Transparenz einzufordern und unter dem Etikett einer »Explainable AI«, einer erklärbaren künstlichen Intelligenz firmiert (Angelov et al. 2021). Das dies erforderlich wird, hängt mit gesellschaftlichen Transparenzansprüchen zusammen. Sollen Entscheidungen von erheblicher Relevanz an KI-Verfahren ausgelagert werden, muss sichergestellt werden, dass für Betroffene nachvollziehbar bleibt, warum so entschieden wurde. Andernfalls verlieren die Verfahren an Akzeptanz und Legitimität. Dies kann sich jedoch in verschiedenen kulturellen Umfeldern unterschiedlich stark zeigen, da es hier zu jeweils eigenständigen Umgangsformen, bei White (1992, 2008) »Stilen«, mit der Transparenzproblematik kommen kann, die sich dann auch unterschiedlich auf die Adaption und Entwicklung von KI-Verfahren auswirken (Züger/Asghari 2022). An dieser Stelle werden zum Beispiel immer wieder die Unterschiede zwischen China, den USA und Europa betont (Castro et al. 2019; Probst et al. 2018). Der Beitrag nutzt die Begriffe der Kontrolle und des Stils aus der Theorie von Harrison White, um die Transparenzproblematik im Bereich der neueren KI deutlich zu machen und den in der Forschung, wie in der Praxis entwickelten Kontrollstrategien nachzugehen (White 1992, 2008; Schmitt/Fuhse 2015). Dabei werden 19 Experten Interviews mit Forscher:innen und Entwickler:innen herangezogen, welche in den Jahren 2020 und 2021 von Mitarbeiter:innen des Lehrstuhls für Technik- und Organisationssoziologie der RWTH Aachen geführt wurden¹. Des Weiteren werden Zeitungsartikel aus den USA und dem deutschsprachigen Raum aus den Jahren 2018 bis 2021 für weitere Analysen zur sich ausbildenden Diskursarena benutzt.

2. Das Transparenzproblem der neueren KI

In der KI-Forschung wird klassisch zwischen regelbasierten und selbstlernenden Algorithmen unterschieden und im Bereich der selbstlernenden Varianten zwischen klassischen Machine Learning- und sogenannten Deep Learning-Verfahren. Während bei klassischen Machine Learning-Verfahren vor allem

1 Neun dieser Interviews wurden im Rahmen des Projektes »ERS University of Alberta AI« durchgeführt.

die Auswahl und Zusammensetzung der Datenbasis ein Transparenzproblem begründen können, so begründet bei Deep Learning-Verfahren zusätzlich auch die tatsächliche Verarbeitung der Daten, welche in gewissem Sinne opak bleibt, ein Transparenzproblem. Die Erzeugung von verschiedenen »hidden layers« in neuronalen Netzwerken ist nicht wirklich vorhersagbar und ihr Einfluss erst durch nachträgliche Analysen zu bestimmen. Hinzu tritt noch die für Benutzer:innen häufig intransparente Art des Trainings der Systeme durch ihre Konstrukteur:innen, wo Veränderungen an der Datenbasis ebenso auftreten, wie eine Versuch-und-Fehler-Manipulation der Vernetzungsparameter. All dies führt nicht nur zur techniksoziologisch allgegenwärtigen Unterscheidung von Expert:innen und Benutzer:innen. Expert:innen versuchen diese Verfahren von innen heraus zu verstehen und beherrschen ihre Funktionsweise, wohingegen Benutzer:innen diese Verfahren nur anwenden und auf ihr Funktionieren vertrauen müssen. Es kulminiert auch in Verfahren, deren Funktionieren prinzipiell selbst für die Expert:innen immer weniger durchschaubar wird und nur nachträglich und an Ergebnissen gemessen werden kann, wie es sonst nur für die Benutzer:innen gilt. Dazwischen lassen graduell verschiedene Expert:innenpositionen unterscheiden, wie etwa Theoretiker:innen (Expert:innen für mathematische Grundlagen, etwa solche die sich mit neuronalen Netzen beschäftigen), Konstrukteur:innen (Expert:innen die tatsächlich Systeme bauen, etwa hier Software-Ingenieur:innen) und Anwender:innen (die Expert:innen für den Anwendungskontext sind, etwa hier Ärzt:innen). Für jede dieser Expert:innenpositionen fallen unterschiedliche Intransparenzen an, die sie entweder bearbeiten können oder die sie als gegeben hinnehmen müssen.

Damit lässt sich die Transparenzproblematik in mehrere Teile aufgliedern, für die nur in unterschiedlichem Maße Lösungsmöglichkeiten bereitstehen und die natürlich auch im Sinne rhetorischer Kontrollstrategien eingesetzt werden können:

1. *Intransparenz der eingesetzten Lerndaten bzw. ihrer Fehlerquellen*
2. *Intransparenz des Trainingssettings und der vorgenommenen Änderungen am System*
3. *Intransparenz der tatsächlichen Funktion für die an der Konstruktion beteiligten Expert:innen*
4. *Intransparenz auch der mathematischen Grundlagen*

Während sich die ersten beiden grundsätzlich noch auf der Ebene transparenter Expert:innendokumentationen nachvollziehen und daher bearbeiten lassen, stellen gerade die dritte und vierte Quelle von Intransparenz im Kontext einer breiten Anwendung von Deep Learning-Verfahren ein erhebliches neues Problem dar.

3. Kontrolle, Kontrollversuche und Kontrollprojekte

Wir schließen in diesem Beitrag an das Kontrollverständnis der Theorie von Identität und Kontrolle, wie sie der Netzwerktheoretiker Harrison White umrissen hat (White 1992, 2008), an. Hier ist grundlegend zwischen den Begriffen Kontrollversuch, Kontrollprojekt und Kontrolle zu unterscheiden. Kontrollversuche sind dabei jegliche Ausgriffe einer Entität auf ihre Umgebung, die gleichzeitig zu einer Verkopplung mit dieser Umgebung führen und somit Positionierung und damit Identitätsformierung nach sich ziehen. Solche Versuche sind nur bedingt intentional zu lesen, da es praktisch unmöglich ist, nicht mit seiner Umgebung in Wechselwirkung zu treten. Ein Kontrollprojekt wird aus einem Kontrollversuch, wenn Positionierungsintentionen zugeschrieben werden können; so bald also eine Entität Kontrollversuche zu organisieren beginnt, um ganz spezifische Kontrollergebnisse in Bezug auf seine Umgebung zu realisieren. Dabei sind Kontrollprojekte als Möglichkeiten der Stabilisierung von sozialen Identitäten zu verstehen, die ein »social footing« erreichen wollen, um so die chaotische Unsicherheit ihrer Umgebung zu reduzieren (ebd. 1992, 2008; White/Godart 2007). Kontrollprojekte sind jedoch nicht einfach als Handlungen von Individuen im Sinne einer soziologischen Handlungstheorie zu verstehen, da sie in jedem Fall auf andere Kontrollversuche und Kontrollprojekte in ihrer Umgebung treffen und erst in ihrer Verstrickung Relationen bilden, die dann eine Positionierung und damit eine erste Etablierung von sozialer Identität erlauben. Diese Unterscheidung von Kontrollprojekten und Kontrolle ist dabei grundlegend, weil Kontrolle jeweils als Kehrseite der Identitätsbildung zu verstehen ist und nicht durch einzelne Handlungen repräsentiert werden kann. Die Positionierung erfolgt nicht auf der Grundlage des Kontrollprojekts allein, sondern ist Folge der Verknotung von Kontrollprojekten, die letztlich eine Identität auf einer relational definierten Position hält. Kontrolle und Identität sind also Produkte relationierender Ereignisse, bei denen sie sich wechselseitig stabilisieren (White 1992, 2008). Der Begriff des Kontrollprojekts ist dabei näher am Handlungsbegriff als der

Kontrollbegriff selbst. Ein Kontrollprojekt ist als Spiel mit den bestehenden Verstrickungen der Identität zu verstehen, basiert also schon auf erfolgreichem »social footing« und kann auf die Erzeugung von Möglichkeiten oder deren Beschränkung gerichtet sein (White 2008: 1ff.).² Die Begrifflichkeit erlaubt es daher die Etablierung von Identitäten nicht im Sinne einer geplanten Verschwörung zu denken, sondern als komplexes Zusammenspiel von kurzen Kontrollversuchen, mehr oder weniger koordinierten umfassenderen Kontrollprojekten und einer sich dabei zwischen den verschiedenen Identitäten etablierenden Kontrollverstrickung. Dem Verstrickungsverhältnis von reziproken Kontrollprojekten soll nun im Einzelnen nachgegangen werden, um mehrere Problemkreise in den Blick zu bekommen, die sowohl in der wissenschaftlichen und öffentlichen Diskussion um die neuere KI bedeutsam sind als auch in den Interviews immer wieder auftauchen. Da ist zum einen die Diskussion um die Notwendigkeit von Theorie. Wenn uns nur Ergebnisse interessieren, müssen wir dann verstehen, wie sie zustande gekommen sind? Hier wird oft die These in Anschlag gebracht, dass man von Kausalität auf Korrelation umstellen sollte, da uns Korrelationen auch ohne Verständnis Zusammenhänge aufzeigen können, während die Suche nach Kausalitäten uns nur verlangsamen würde. Ohne Kausalität ist aber keine Erklärung möglich und es ist fraglich, woher dann das Vertrauen in die Ergebnisse kommen soll (Schmitt 2018). Zugleich spielt hier auch die Vermenschlichung von Technik und die Diskussion um posthumanistische Zugänge eine Rolle, wenn es etwa darum geht, dass Intelligenz nicht in Abhängigkeit von menschlichen Vermögen diskutiert werden sollte, sondern die menschliche Besonderheit aus dem Konzept entfernt werden muss. Schließlich geht es hierbei auch um die Frage der Komplexität von Zusammenhängen, die sich einmal auf der Ebene der KI-Technologie selbst stellt, dann aber ebenso auf der Ebene der Einbettung in soziotechnische Systeme. Der Zugang über die Theorie von Harrison White ermöglicht hier aber eine Verarbeitung, die durch das Scharfstellen auf Kontrollversuche und Kontrollprojekte sowie deren Verstrickung durchaus weiterführend ist.

2 Bei White wird die Erzeugung von Möglichkeiten oder deren Beschränkung »getting action« bzw. »blocking action« beschrieben (White 1992: 230ff.; White 2008: 279ff.).

4. Blackboxing als Kontrollversuch und als Kontrollprojekt

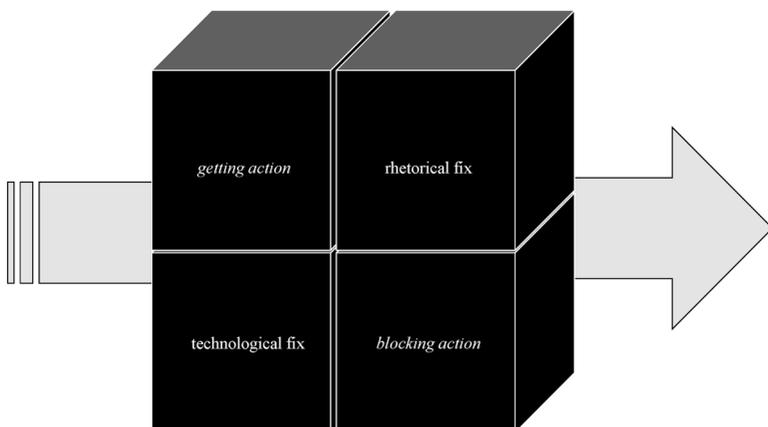
Ausgehend von dieser Beschreibung von Kontrollprojekten und Kontrolle ergibt sich eine doppelte Lesart des Verweises von KI-Forscher:innen auf den Blackbox-Charakter der Deep Learning-Verfahren: Zum einen ist der Verweis auf das Blackboxing als Kontrollprojekt lesbar, dass bestimmten Identitäten Freiheitsgrade bzw. Kontrollmöglichkeiten eröffnen soll, während es versucht die Kontrollmöglichkeiten bzw. -ansprüche anderer Identitäten abzuwehren (Geitz et al. 2020). Gleichzeitig kann man aber in Anlehnung an die Arbeiten der ANT auch davon sprechen, dass Blackboxing eine gelungene Kontrollverstrickung darstellt, die bestimmte Zusammenhänge innerhalb von Netzwerken dem Zugriff entzieht, also eine Realisierung von Kontrolle/Identität darstellt, die Positionierungen in besonderer Weise festigen kann, indem auch entkoppelt wird (Callon/Latour 1981; Latour 1999). Dabei ist auch klar, dass Erklärung der und das Angebot von Interpretationsmöglichkeiten nur eine andere Variante von Kontrollversuchen darstellt, welche die Vertrauensproblematik des Blackbox-Charakters der KI-Verfahren bearbeitbar machen.

Kontrollprojekte zwischen Technologie und Rhetorik sowie zwischen Blocking und Getting Action

Blackboxing soll in diesem Beitrag daher als ein Kontrollprojekt verstanden werden, das es einerseits möglich macht, mit den beteiligten Intransparenzen zu arbeiten, ohne doch erhebliche Vertrauensverluste in die Technologie hinzunehmen. Dafür wird ein diskursiver Rückgriff auf die gängige technologische Intransparenz und das dabei entscheidende Vertrauen in Technik unterstellt. Da das innere Funktionieren der Technik für die Benutzer:in nicht einsichtig ist, muss sie sich auf eine Kontrolle der Ergebnisse verlassen, die aber das Nicht-Funktionieren zunächst einmal in für die Benutzer:in irritierender Weise offenbaren müsste. Blackboxing als Kontrollprojekt erzeugt Freiheitsgrade durch Einsatz von Unterbrechung von Kontrolle. Intransparenz kann hier sozusagen produktiv zum Einsatz gebracht werden, solange das Vertrauen in das technische Funktionieren sichergestellt werden kann. Dies kann genau an der Stelle problematisch werden, wo das Funktionieren der technischen Systeme sowohl für Expert:innen, als auch die Benutzer:in nicht mehr ohne Weiteres festgestellt werden kann. Hier müssen dann eventuell auch weitere Strategien und Kontrollversuche zum Einsatz kommen, um die Vertrauensproblematik, die aus der Intransparenz folgt, erfolgreich zu bearbeiten. Um

diese Problematik detaillierter herauszuarbeiten, wollen wir im Folgenden einen tieferen Blick in die Interviews und auch in die öffentliche Debatte werfen, um die hier eingesetzte Kontrollstrategie zu verstehen und möglichst klar in ihrer erzählerischen Form zu umreißen.

Abbildung 1: Schematische Darstellung der 4 Foki von Kontrollversuche



Es fällt auf, dass es häufig zu einer erzählerischen Kopplung von Intransparenz, Blackboxing und Fragen nach Erklärbarkeit oder Interpretierbarkeit kommt. In der Erzählung bzw. kommunikativen Bearbeitung äußert sich also genau jene Vermengung von Relationen, denen man am besten mit einer netzwerktheoretischen Konzeption begegnen kann. Blackboxing steht damit nicht alleine, sondern muss als ein Einsatz im Umgang mit Intransparenz verstanden werden, der andere Einsatzmöglichkeiten impliziert und in seinem Erfolg von diesen abhängig bleibt. Es ist daher aus unserer Sicht notwendig und gleichzeitig gewinnbringend das Transparenzproblem der KI in eine Kontrollproblematik zu verwandeln und uns die erzählerische Bearbeitung dieses Problems anzusehen. Dabei kann man vier Foki der Kontrollversuche unterscheiden, je nach ihrer Ausrichtung und der Methode, um die Transparenzproblematik beschreiben zu können. Zum einen kann man unterscheiden, ob es sich um einen Einsatz handelt, der Handlungsmöglichkeiten öffnen oder schließen soll (getting action/blocking action) und zum anderen, ob es sich um einen technologischen oder einen rhetorischen Ansatz (technological fix/rhe-

torical fix) handelt (Abb. 1). Geht man von diesen Kombinationsmöglichkeiten aus, ergeben sich kommunikative Versuche der Begrenzung oder Eröffnung von Freiheitsgraden und technologische Versuche der Schließung oder Öffnung. Dabei handelt es sich um jeweils perspektivische Schwerpunktsetzungen und graduelle Verortungen von Kontrollversuchen, die eher in ihrer spezifischen Schwerpunktsetzung unterscheidbar bleiben. Jedes Kontrollprojekt setzt hier erkennbare Schwerpunkte, die beobachtbar sind und anzeigen, wie die Blackbox im jeweiligen Fall eingesetzt wird.

Empirische Darstellung der Transparenzproblematik

Wie wir dargestellt haben, präsentiert sich die Transparenzproblematik der KI auch auf Seite der Expert:innen in unterschiedlicher Tiefenschärfe und damit auf unterschiedlichen Analyseebenen. Dies lässt sich auch in den von uns analysierten Interviews gut nachvollziehen. Dabei stellen wir jeweils dar, worin die Probleme gesehen werden und diskutieren dann an Ankerbeispielen die Formen der Kontrollversuche und -projekte, die in Anschlag gebracht werden, um damit umzugehen. Diese sind dann jeweils im Hinblick auf die eingeführten Schwerpunktsetzungen zu analysieren.

Ausgehend von der Kodierung möglicher Quellen von Intransparenz (Kap. 2) in den Interviews, wurden inhaltsanalytisch Kontrollversuche kategorisiert und zu Kontrollprojekten zusammengefasst/strukturiert. Kontrollversuche sind wie oben beschrieben, als Ausgriffe einer Entität auf Ihre Umgebung zu verstehen, welchen noch keine Positionierungsintention zuzuschreiben ist. Im Rahmen des verwendeten Expert:inneninterview-Korpus, welcher verschiedene Anwendungsszenarien von KI im wissenschaftlichen Kontext der Expert:innen adressiert, verbleiben Kontrollversuche auf der Ebene von technologischen Ansätzen im Umgang mit Quellen von Intransparenz. Die Zuschreibung von Positionierungsintentionen auf Ebene der Kontrollprojekte umfasst dann sowohl technologische als auch rhetorische Ansätze, welche mögliche Quellen von Intransparenz zum Anlass für die Erzeugung von Möglichkeiten (getting action) oder deren Beschränkung (blocking action) nehmen.

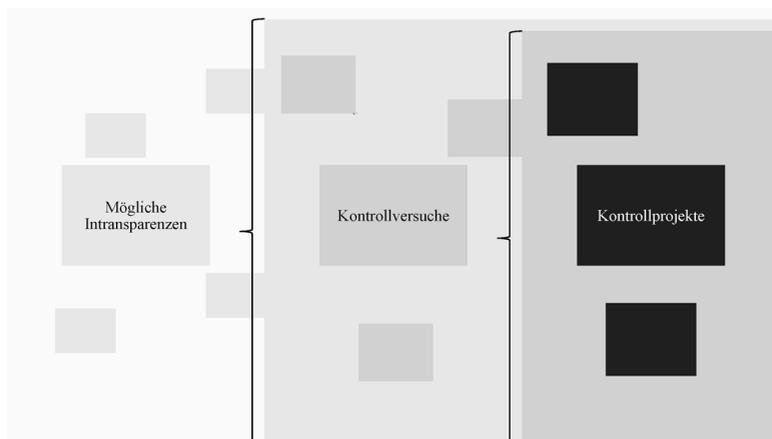
»Es gibt auch noch so ein paar Netzwerkarchitektur spezifische Sachen, also dass man beispielsweise will, dass Gewichte, die innerhalb von diesen Netzwerken benutzt werden, eine bestimmte Eigenschaft haben, zum Beispiel, dass sie nicht beliebig groß werden oder auf einen bestimmten

numerischen Bereich oder sowas begrenzt sind und wenn man quasi während der Modellgenerierung im Prinzip solche Constraints quasi einführt, kann man auch im Prinzip so ein bisschen steuern, in welche Richtung dann das Training läuft. Genau.«

»Ja gut, grundsätzlich kontrolliert man das ja erstmal nur insofern, als dass man die Architektur des neuronalen Netzwerks vorgibt und dann den Lernalgorithmus. Was dann da drin dann genau passiert, das ist eben etwas, was grundsätzlich nicht so genau verstanden wird. Deswegen natürlich die Gefahr [...], dass so ein Algorithmus eben Entscheidungen trifft, die man nicht unbedingt als Mensch nachvollziehen kann. Vielleicht sind das gute Entscheidungen, vielleicht auch nicht; wir können es nicht nachvollziehen.«

Das hier eingeführte Zitat macht deutlich, dass es im Umgang mit den KI-Verfahren sehr stark um Kontrolle geht, also darum, was kontrolliert und vor allem, wie kontrolliert werden kann. Aus den Expert:innen-Interviews geht hervor, dass zunächst an drei Stellen in einem Deep Learning-Verfahren Kontrolle ausgeübt werden kann. Eine (1) spezifische Architektur des neuronalen Netzwerkes kann gewählt werden, genauso wie (2) der Lernalgorithmus. Kontrolliert werden kann auch die (3) Qualität und Quantität der Lerndaten.

Abbildung 2: Von Intransparenzen/Unsicherheiten zu Kontrollprojekten



Wir gehen im Folgenden davon aus, dass sich im Feld der KI unterschiedliche Formen von Intransparenzen/Unsicherheiten ergeben, die in den analysierten Interviews mit unterschiedlichen Kontrollversuchen verknüpft werden, die sich in gelegentlichen Fällen zu Kontrollprojekten verschränken. Wir gehen also den von den Expert:innen selbst hergestellten Verknüpfungen nach, um aufzuzeigen, wie Intransparenzen durch Kontrollversuche und Kontrollprojekte in Blackboxes verwandelt werden bzw. wie versucht wird diese Blackboxes zur Rückgewinnung von Kontrolle wieder zu öffnen.

Intransparenz der eingesetzten Lerndaten bzw. ihrer Fehlerquellen

Die hier bestehende Intransparenz ist auch im öffentlichen Diskurs um die KI-Systeme häufiger angesprochen worden³. Es ist hier nicht klar auf welcher Datengrundlage Ergebnisse eigentlich beruhen und ob diese Datengrundlage auch gut ist, also frei von Vorurteilen oder Rauschen. Aus der Kontroll-Perspektive weisen die Interviews hier auf einige Versuche und Projekte hin, die gängigerweise in Anschlag gebracht werden. Diese werden jedoch häufig nach außen nicht sichtbar, sodass eine Kontrolle dieser Daten von externer Seite häufig nicht gegeben ist.

-
- 3 Bezogen auf die Leistungsfähigkeit von KI-Verfahren, schrieb die Zeit: »Falsch wäre es, nur auf die Software zu blicken oder auf die zugrunde liegenden Algorithmen. Denn meistens sind nicht die Algorithmen der interessante Punkt, sondern die Daten« (Randow 2018). In der Welt wird das Problem mit gebiasteten Algorithmen folgendermaßen beschrieben: »Algorithmen arbeiten mit historischen Daten und erkennen Muster, anhand derer sie Entscheidungen treffen. Gab es in den zugrundeliegenden Daten jedoch rassistisch, sexistisch oder homophob gefärbte Tendenzen, wird der Algorithmus diese übernehmen und auf zukünftige Entscheidungen übertragen. Diskriminierende Effekte werden bei rein maschinell getroffenen Entscheidungen jedoch unsichtbar – gerade, weil wir der Maschine Neutralität zuschreiben« (Lehmann 2021). Zuletzt ergab eine Recherche des Time Magazine, dass OpenAI auf Clickworker aus den globalen Süden zurückgegriffen hat, um ein Sicherheitssystem für ChatGPT zu entwickeln, um zu verhindern, dass diese, nicht wie seine Vorgängerversionen, erlernte anstößige oder missbräuchliche Sprache verwendet: »But it was a difficult sell, as the app was also prone to blurting out violent, sexist and racist remarks. This is because the AI had been trained on hundreds of billions of words scraped from the internet—a vast repository of human language. That huge training dataset was the reason for GPT-3's impressive linguistic capabilities, but was also perhaps its biggest curse. Since parts of the internet are replete with toxicity and bias, there was no easy way of purging those sections of the training data« (Perrigo 2023).

Da die Intransparenz der Daten nicht abschließend zu lösen ist, wird nach Wegen gesucht hiermit umzugehen, um wieder Vertrauen in die Daten herzustellen. In der Regel geht es hier um die Fragen, ob die Verfahren auf einer »guten« oder »schlechten« Datengrundlage basieren, ob diese die Realität außerhalb der Lerndaten einfangen können und ob sie Verzerrungen unterliegen oder gar problematische soziale Strukturen aufgreifen und damit helfen diese zu reproduzieren (Langer/Weyerer 2020). An diesen Leitplanken orientiert lassen sich dann eine Reihe von Intransparenzen unterscheiden. Zunächst können die Daten manifeste Vorurteile enthalten, die sich aus der Datenstruktur oder der Annotationsmethode ergeben. Sie können latente Vorurteile enthalten, die in den Sinnstrukturen selbst liegen, etwa wenn Frauen oder Migrant:innen systematisch bei Bewerbungsprozessen benachteiligt wurden und diese Daten in den Daten abgebildet ist (Lloyd 2018)⁴. Schließlich lassen sich diese Intransparenzen, hinsichtlich der möglichen Fehlerquellen, auf die sie sich beziehen, noch weiter klassifizieren. Hier geht es vor allem um die Repräsentativität der Daten, etwa inwiefern diese ihren Anwendungskontext angemessen repräsentieren, also die in Frage stehende Grundgesamtheit abbilden. Entsprechen die in den Daten anzutreffenden Quantitäten denjenigen, mit denen es der Learner in der erweiterten Umgebung zu tun bekommt? Gibt es positive oder negative Selektivitäten in den Daten? In verschiedenen Anwendungsdomänen, etwa in der Medizin, überwiegt die Anzahl positiver Fälle in gesammelten Gesundheitsdaten (Van Aert et al. 2019). Schließlich können gerade auch annotierte Daten die entsprechenden Experten:innenurteile widerspiegeln, die in sie eingegangen sind und verzerrende Effekte haben.

In den Interviews lassen sich verschiedene Kontrollversuche der Forscher:innen unterscheiden, um mit dieser Inadäquatheit der Daten umzugehen, also »schlechte« Daten trotzdem sinnvoll verarbeiten zu können oder diesen Datensatz zu »bereinigen«. Kontrollversuche umfassen hier unterschiedliche Formen der Qualitätsprüfung, der Normierung und Veredelung, sowie technische Lösungen wie Algorithmen, die ungleichgewichtige Daten ausgleichen oder schlicht über die Kontrolle des Outputs der Systeme. Solche Kontrollversuche verdichten sich häufig zu spezifischen Kontrollprojekten, von denen wir nun einige am Beispiel aus den Interviews aufzeigen möchten.

4 Mittlerweile eines der prominentesten Beispiele für systematische Benachteiligung von Frauen in Bewerbungsprozessen ist ein Experiment Amazons mit einem KI-Rekrutierungstool (Dastin 2018).

Das *Kontrollprojekt* »*Good enough*« (100 % oder Perfektion nicht das Ziel), dass sich dadurch auszeichnet nur eine zufriedenstellende Output-Lösung zu bekommen bzw. die Daten auf einen Stand zu bringen der als akzeptabel angesehen werden kann.

»Man sucht diese Daten in der Regel nicht wirklich aus, sondern man nimmt, was man kriegt.«

»Also im Prinzip, beispielsweise im medizinischen Fall würde man da einfach im Idealfall mehrere Mediziner fragen, die dann eben genau die gleiche Datenmenge annotieren und dann kriegt man so eine Art mittlere Annotierung von verschiedenen Medizinern zum Beispiel und man merkt auch, dass beispielweise jeder Mensch, der irgendwas annotiert, im Prinzip das einen Tick anders macht. Also man hat da auf jeden Fall auch so eine Varianz zwischen den Menschen, die die Annotierungen erstellen und auch innerhalb.«

»Es gibt jetzt für meinen Bereich keinen vernünftig annotierten, großen Datensatz.«

»Also wie muss die Infrastruktur für die Bereitstellung von Daten eigentlich aussehen? Wie kriegen wir die Data Provenienzen? Wie bekommen wir die ethischen Fragen dazu technisch erfüllt oder die ethischen Anforderungen technisch erfüllt und organisatorisch erfüllt, damit wir diese Daten dann auch zum Lernen in neuronalen Netzen nutzen können? Und wie können wir den Aufwand, den man dafür hat, sowas überhaupt dann in ausreichender Anzahl und qualitativ hochwertig zu haben, so hin, dass diese Daten mit der hinreichenden Qualität überhaupt verfügbar sind?«

»Aber was funktioniert schon zu hundert Prozent? Also da ist es natürlich ganz praktisch, wenn man hauptsächlich mathematische Beweise macht, da ist man auf der sicheren Seite. Aber es hilft ja nichts. Das finde ich also ((Pause)) Ich finde das nicht so schlimm, dass Fehler passieren. Selbst wenn diese Fehler manchmal bei technischen Sachen sehr weitreichend sein können. Aber mein Gott, es passieren Fehler. Aber man muss versuchen damit umzugehen und da hat man jetzt noch nicht überall die richtigen Lösungen gefunden. Andererseits ist das Ganze ein Thema ... ja, Fairness vor allen Dingen aber auch das was explainability, also die Erklärbarkeit von den Ergebnissen, was damit zu tun hat, das ist im Moment in der Forschung ein zentrales Thema.«

Das *Kontrollprojekt Naturalisierung (Irren ist menschlich)*, weist Ähnlichkeiten mit »Good Enough« auf, enthält aber gleichzeitig eine implizite oder explizite Rechtfertigungsebene, welche außerhalb der Akzeptabilität der Ergebnisse liegt und damit auch potenziell schlechten und eben auch diskriminierenden Output rechtfertigen kann. Wichtig ist an den Zitaten hier die Rechtfertigungskomponente in Bezug auf natürliche Vorgänge, wie der menschlichen Produktion von Irrtümern und Vorurteilen, die nicht zu verhindern wären. Die Naturalisierung spitzt sich in der Gleichsetzung des Prozesses der Bildung von Vorurteilen mit dem von Erfahrungswerten zu, welche auch in anderen Kontexten zur Rechtfertigung von struktureller Diskriminierung auf Basis von statistischen Erfahrungswerten herangezogen wird. Die Schwelle von Kontrollprojekt zu Institution wird an dieser Stelle allerdings überschritten, und zeigt auf, wie problematisch dieses auf Getting Action ausgelegte Kontrollprojekt ist.

»Da hat man halt Schwierigkeiten [...], dass sich Vorurteile verschieben und verstärken und ähnliches. Das ist schwierig in dem Sinne, weil man das nur sehr schwer kontrollieren kann und das dann auch Dinge sind, die man selber nicht gut versteht. Also das eine was man sagen muss: Natürlich sind Vorurteile immer da, ich glaube auch tatsächlich, dass Vorurteile wesentlich sind, dass wir überhaupt funktionieren als Menschen. Nämlich, dass wir schnell Entscheidungen treffen können, also das ist nicht beides was rein Schlechtes.«

»Und natürlich basieren Entscheidungen immer auf Vorurteilen oder Erfahrungswerten. Je nachdem wie man das sagen will, Erfahrungswerte ist irgendwie positiv. Vorurteil ist negativ, aber es ist genau das gleiche, letzten Endes.«

Diese *Naturalisierung* funktioniert natürlich auch in die andere Richtung (*Irren ist menschlich*), wenn etwa die schlechte Qualität der Lerndaten damit entschuldigt wird, dass diese »menschengemacht« seien.

»[...] das ist immer eine Frage, die man sich stellen muss, was die Datenqualität ist und was man davon erwarten kann; Man kann sich natürlich vorstellen, dass jetzt bei supervised, wo Menschen schon etwas kategorisiert haben, jetzt stärker menschliche Einschätzungen, Vorurteile oder so etwas reinkommen, andererseits gilt das für alle Daten, dass die verfälscht sein können.«

Trotzdem ist *menschenbasierte Supervision* eine weitere Art von Kontrollprojekt, da hier das Gütekriterium der Akzeptabilität von Lerndaten letztlich auf den Menschen verlagert wird. Der Mensch wird als Letztentscheidungsinstanz stark gemacht und die Verantwortung wird auf ihn zurückgespielt, um den technischen Lösungsansatz einerseits zu schützen und andererseits Ängste zu nehmen. Wertmaßstab bleibt der Mensch außerhalb des KI-Verfahrens.

»Also im Endeffekt ist es eigentlich der Mensch, der da so das letzte Gütekriterium da aufstellt. Also, oder, beziehungsweise, man kann es eigentlich nicht wirklich bewerten, wenn es nicht ein Mensch bewertet.«

»Also da gibt's ganz viele große Fehler, die auch passieren würden bei so Trainingsdaten, sammeln, oder ein Bias der im Endeffekt durch die Daten reinkommt, wo auch 'n Mensch notwendig ist, das zu erkennen und das zu überblicken und einschätzen zu können natürlich. Weil die Maschine wird es nicht einschätzen, das ist klar. Die ist komplett abhängig von den Trainingsdaten und von der Art und Weise wie die gefüttert wird. Also an der Stelle ist 'n Mensch notwendig.«

Schließlich lassen sich auch verknottete Kontrollprojekte beobachten, die menschenbasierte Supervision und Naturalisierung kombinieren, die dann auch normativ den Menschen als Kontrollinstanz ins Spiel bringen, da der Algorithmus ja ohnehin auch bei den anderen Aspekten von Menschen kontrolliert wird und man z. B. bei ethischen Fragen, dann auch andere Expert:innen ins Spiel bringen muss als Konstrukteur:innen von Deep Learning-Verfahren.

»How do we control it? We control it with the algorithm, we control it with the parameters we give to the algorithm and we control it with the data that we give to learn. [...] If the AI is biased it's because we give it the biased data; its starts making its own decisions, those are very important decisions that ethicists... people specializing in ethics should make. We shouldn't leave it up to the AI specialist to make that decision.«

Als Vorgriff lassen sich auf dieser Ebene auch schon zwei Stile unterscheiden, die in den Interviews zwischen den deutschsprachigen und den englischsprachigen Forscher:innen zum Ausdruck kommen. Für die deutschen Interviewten ist die Feststellung wichtig, dass mehr Daten nicht unbedingt ein Erfolgsgarant für die Systeme sind, während in den englischsprachigen Inter-

views Probleme mit den Daten als Herausforderung hervorgehoben werden. Wir werden am Ende nochmal auf diese Stilunterschiede zurückkommen.

Es wird also der Lernansatz verändert, um mit diesen Unwägbarkeiten in den Daten *klarzukommen*, aber es kommt nicht zu einer deutlichen bzw. integrierten Kontrollstrategie, da nicht klar ist nach was hier eigentlich gesucht wird. Ein Weg ist, wieder menschliche Beobachter:innen und ihre Gütekriterien hinzuzuziehen, um hier ausgleichend zu wirken, was aber scheitern kann.

Intransparenz des Trainingssettings und der vorgenommenen Änderungen am System

Wenn man sich nicht nur die Daten und deren Ursprung selbst anschaut, wird schnell deutlich, dass auch das Testen und Voreinstellen des Systems beim Training eine wichtige Rolle spielt. Was hier passiert, wird zum einen von den Verfahrensproduzent:innen kontrolliert. Sie drehen an den Schrauben, aber ohne immer ganz genau zu wissen, was sie da eigentlich tun. Parameter werden verändert und es passiert etwas, Kontrolle wird aber nur über die Überprüfung von Ergebnissen ausgeübt.

Auf dieser Ebene lässt sich vor allem das Kontrollprojekt von Versuch und Irrtum in allen möglichen Variationen erkennen, die aber dann auch wieder Intransparenz erzeugt. Das Trainingssetting als Prozess mit konstanten Veränderungsbedarf scheint von vielen Randbedingungen neben den Daten abhängig zu sein. Die Liste dieser Intransparenz erzeugenden Randbedingungen ist lang. Da geht es zum einen um die Ressourcen Zeit und Rechnerkapazität. Wie lange kann gelernt werden und mit welchem Durchlauf? Vorgehensweisen und Verfahren, die die Erklärbarkeit, bzw. die Interpretierbarkeit der Prozesse und Ergebnisse erhöhen können, sind zeitaufwendig, sodass limitierte Zeit- und Rechnerkapazitäten auf Kosten der Erklärbarkeit gehen. Zudem besteht immer die Möglichkeit von zu starker Generalisierung und Übertraining, wenn zu sehr auf das Lernmaterial oder gewünschte Ergebnisse hin angepasst wird. Wie wird mit Ausreißern in den Daten umgegangen? Kann eine Manipulation der Daten, ob intendiert oder nicht intendiert, ausgeschlossen werden. KI-Verfahren erfassen (und verstärken eventuell) Muster, die von menschlichen Beobachter:innen nicht immer nachvollzogen werden können.

Schließlich brauchen die Forscher:innen Erkenntnisse über die Unterschiede zwischen Lerndaten und Anwendungsdaten, um überhaupt interpretieren zu können, was ihr KI-Verfahren hier zu leisten in der Lage ist. In Machine Learning- und Deep Learning- Verfahren unterscheiden sich

meist Lern- oder Trainingsdaten von den Daten, die in der Anwendung dem Verfahren zugeführt werden. Ein KI-basiertes Transkriptionstool, das Sprachaufnahmen verschriftlichen soll, lernt zum Beispiel auf Audiodaten von Nachrichtensendungen oder parlamentarischen Debatten, da diese einerseits öffentlich zugänglich sind und andererseits mit zugehörigen Transkripten veröffentlicht werden. Unterscheidet sich der Sprachgebrauch im Anwendungskontext, in dem etwa wissenschaftliche Interviews automatisiert transkribiert werden sollen, kann die auftretende Fehlerrate eben nur mit Kenntnis der Lerndaten interpretiert werden.

Daraus können dann eine Reihe von Intransparenzen entstehen, da nicht immer klar ist welche Änderungen im Trainingsverlauf vorgenommen wurden und/oder welche Effekte diese Änderungen eigentlich produziert haben. Zusätzlich ist es schwierig zu beurteilen, welche Änderungen hier relevant sind. Es könnten sich auch Rahmenbedingungen geändert haben, die eventuell nur bedingt kontrollierbar sind und deren Einfluss auch nur schwer abzuschätzen ist. Etwa das Verfahren auf einem anderen Gerät (einem schnelleren oder langsameren Computer oder Computernetzwerk) zum Einsatz kommt. Diese Änderungen an den Rahmenbedingungen sind ein fundamentaler Ausgangspunkt für die Intransparenz, die beim Training von Deep Learning-Verfahren bestehen. Im Bereich der Bild- und Signalverarbeitung ist es vorstellbar, dass sich im Trainingsdatensatz nur Computertomographie-Scans eines bestimmten Gerätetyps befinden. Wird das fertig trainierte KI-Verfahren nun auf Scans eines anderen Gerätetyps angewendet, ist es wichtig, inwiefern die Scans sich genau unterscheiden, um etwaige Abweichungen in der Performance des Verfahrens interpretieren zu können.

Die Kontrollversuche drehen sich hier mehr um tatsächliche Eingriffsmöglichkeiten als – wie auf der vorigen Ebene – um Erwartungsmanagement. Es geht um Sensibilitäten im Umgang mit den Lernern. Welche Erfahrungen haben die Forscher:innen gemacht, welche Veränderungen an Parametern sind erfolgsversprechend für bessere Ergebnisse und wie wird mit auftauchenden Problemen umgegangen? Eine Kontrollmöglichkeit liegt in der Regularisierung der Variationen. Wenn man immer wieder dieselben Änderungsschritte vollzieht, kann man einerseits gut nachvollziehen, was getan wurde und es fällt in der Summe über mehrere Lernsettings dann auch leichter Effekte abzuschätzen. Eine weitere Möglichkeit sind Vereinfachungen vorzunehmen, also Variablen und Dimensionalität zu reduzieren. Dadurch gehen eventuell bestimmte Informationen verloren, aber gleichzeitig steigt die Interpretierbarkeit der Ergebnisse durch diese Vereinfachungen. Angewendet wird auch

das sogenannte »fitting«, wobei versucht wird die Fehler zu reduzieren und erwartete Lösungen zu maximieren. Hier ist jedoch die Grenze zur – auch nicht intendierten – Manipulation sehr dünn. Eine Option ist auch immer, mehr Daten und stärker unterschiedliche Daten heranzuziehen. Schließlich gibt es immer die Möglichkeit der Output-Kontrolle.

All diese Versuche können sich wieder in spezifischen Kontrollprojekten verstricken, welche eine gewisse Ähnlichkeit mit jenen aufweisen, die wir schon auf der 1. Ebene der Datenintransparenz gesehen hatten. Konzentrieren wir uns also zunächst auf die auf dieser Ebene neu eingeführten Kontrollprojekte, die spezifisch für die Intransparenz des Trainingssettings sind:

Da ist zum einen der *diskrete Charm der Black Box* (eine Möglichkeit von Getting Action im Sinne Whites), wobei der Verweis auf die Intransparenz gleichsam als magische Fähigkeit Resultate zu liefern angepriesen wird:

»Also der Charme an dem neuronalen Netz ist ja, dass man als Mensch eben nicht so ganz genau durchschauen muss, wie die Muster darin aussehen. Sonst bräuchte ich ja diesen ganzen Lernprozess nicht. Wenn ich das schon wüsste, dann könnte ich das ja direkt schon programmieren, ohne mir diese ganze Mühe mit dem Lernen zu machen.«

Hier wird Erklärbarkeit direkt abgelegt und als Vorteil dargestellt, es gehe eben nicht um Erklärbarkeit.

Dann die auch in anderen Kontexten stark verbreitete Kontrollstrategie von *Trial und Error*, die wir oben auch als grundlegend angesprochen hatten. Sie ist charakteristisch für die Beschreibung des gesamten Trainingssettings:

»Das ist teilweise so ein bisschen Trial and Error. Also man hat natürlich irgendwann so ein bisschen Erfahrungswerte, an welchen Stellschrauben man jetzt quasi drehen muss, um das Ergebnis in die richtige Richtung zu bekommen, aber es ist teilweise schon auch viel so Blackbox-mäßig. Also dass man im Prinzip einfach Trial-and-Error-mäßig Sachen ausprobiert und die einen Sachen funktionieren und die anderen funktionieren nicht. Und oft ist es auch so, dass man sich nicht so hundertprozentig erklären kann, woran es eigentlich liegt.«

»Das könnte auch sein, dass der, wenn man dem jetzt irgendwie 100 neue Daten gibt, dass der plötzlich 'nen riesen Sprung macht und viel besser wird,

aber das ist... Vorher kann man das nicht wissen, also... Kann auch sein, dass es umgekehrt wirkt, ich gebe dem neue Daten, er wird schlechter, überall.«

Schließlich noch die schon bekannte an der Outputkontrolle orientierte Strategie des »good enough« (100 % oder Perfektion nicht das Ziel) (siehe oben).

»Bei mir war es tatsächlich etwas Glück. Ich habe an Parametern rumgespielt, die ich für sinnvoll erachtet habe, wo ich so das Gefühl hätte, ok, hier könnte man was dran schrauben und dann wird es besser. Und das wurde dann tatsächlich auch besser, aber nicht zu hundertprozentig perfekt, sagen wir mal so.«

Hier auch noch in Kombination mit Versuch und Irrtum, wobei aber zentral ist, dass mit einem Maß operiert wird, das der Forscher:in oder Entwickler:in anzeigt, dass eine zureichende Genauigkeit des Verfahrens durch die Veränderung der Parameter erreicht wurde.

Intransparenz der tatsächlichen Funktion für die an der Konstruktion beteiligten Expert:innen

Auf der Grundlage der quantifizierbaren Genauigkeit der Verfahrens-Ergebnisse im Trainingssetting, haben Expert:innen ausgeprägte Kontroll-Möglichkeiten über diesen Teil des Verfahrens. Die *tatsächliche* Funktion bzw. das Funktionieren in Anwendungskontexten hingegen ist nur sehr bedingt kontrollierbar. Neue Kontrollverstrickungen kommen hinzu, die nicht in das Trainingssetting einbezogen werden können. Was hier passiert ist daher nur bedingt nachvollziehbar und kann, wenn überhaupt, nur mit großem Aufwand ex-post aufgearbeitet werden. Der Anker für die Kontrolle liegt hier darin, zu bewerten, ob der Umsetzung eine gute oder schlechte, sinnvolle oder nicht sinnvolle Entscheidung zugrunde liegen, also nach welchen Kriterien hier selektiert wurde.

An dieser Stelle kann von einer Intransparenz der tatsächlichen Funktion oder des Funktionierens für die an der Konstruktion beteiligten Experten:innen gesprochen werden. Diese drückt sich darin aus, dass die verschiedenen Ebenen des neuronalen Netzes zwar identifiziert werden können, aber nicht erklärt werden kann, warum etwas auf dieser oder jener Ebene passiert. Dies bedeutet, dass auch der Entscheidungsprozess in der Anwendung, wie auch schon der Lernprozess nicht gänzlich transparent gemacht werden kann. Die

Grundprinzipien können theoretisch nachvollzogen werden, aber die Komplexität der Verschaltungen kann als Prozessergebnis und in seinem konkreten Ablauf nicht nachvollzogen werden. Durch diese Intransparenz kann es dann dazu kommen, dass zu hohes Vertrauen in Verfahren gesetzt wird, die selbst keine perfekten und auch keine als stimmig überprüfbareren Ergebnisse liefern können. Warum-Fragen kann das Verfahren nicht beantworten und es bietet in der Regel auch keine für menschliche Beobachter:innen nachvollziehbare Prozessdaten an, die hier eine zusätzliche Orientierung anbieten könnten. Es ergibt sich daraus eine Kontrollproblematik, die im Spannungsverhältnis zwischen dem Sinnzwang sozialer Systeme und der menschlich nicht nachvollziehbaren Mustererkennung der KI-Verfahren entsteht und die dann wiederum mit Kontrollversuchen und Kontrollprojekten eingehegt werden muss (Luhmann 1991: 92ff.).

Die Kontrollversuche nehmen hier gegenüber den beiden vorherigen Ebenen neue Wege, da hier die Expert:innen auch die Prozesskontrolle verlieren, über die sie im Trainingssetting noch verfügten. Hier treten dann auch kritisches Hinterfragen und anwendungsbezogene Gütekriterien auf den Plan. Kritisches Fragen richten sich dabei auf die Sinnhaftigkeit der Ergebnisse (nicht auf Erklärungen des Prozesses). Können der menschlichen Evaluator:in die Ergebnisse plausibel vor und kann sie sie außerhalb des Ergebnisses des Learning-Verfahrens als sinnvoll rekonstruieren? Dies wird zusätzlich vereinfacht, wenn das Anwendungsfeld schon Gütekriterien für diese Ergebnisse formulieren kann, an denen man sich dann orientiert. Es wird auch versucht, die KI's nur als Assistenzsysteme zu rahmen, die nur Vorschläge anbieten und nicht selbst entscheiden, um ihre Intransparenz so weniger problematisch erscheinen zu lassen. Es tauchen an dieser Stelle dann aber auch viele Kontrollversuche auf, die mit White als Blocking Action beschrieben werden können, die also die Anwendung von solchen Lernern eher einschränken. Dazu gehören die kritische Stellungnahme, dass eine Erklärung und damit legitime Verteidigung der Entscheidungen dann nicht möglich sei, dass auch eine Fehleranalyse nicht möglich sei und dass der Verzicht auf die Beantwortung der Warum-Frage Manipulationen Tür und Tor öffnet. Eine Möglichkeit damit umzugehen wird in dem Versuch gesehen über viele Anwendungsfälle hinweg eine Reproduzierbarkeit von Ergebnissen sicher zu stellen und damit zu zeigen, dass es sich um ein gutes Modell handelt, dass durch viele Fälle abgesichert wurde.

Auch bei den Kontrollprojekten ist eine stärkere Durchmischung von getting und Blocking Action zu beobachten. Die klarste Positionierung zur Black-

box um Handlungsfähigkeit herzustellen (getting action) ist die Position der Ignoranz. Die praktische erfolgreiche Implementierung setze gar kein rigoroses Verständnis der Abläufe voraus:

»Also kontrollieren können Sie den Algorithmus nicht. Sie verstehen ja auch nicht, was er macht. Das ist ja eine Black Box. Das ist ja sehr, sehr kompliziert, posthoc rausfinden zu wollen, warum der Algorithmus so oder so gelernt hat. Da gibt es gewisse Verfahren, aber die sind so wahnsinnig aufwendig, die interessieren uns momentan auch nicht.«

»Und das ist auch nicht so, dass das so sein, also dargestellt werden müsste, oder dass das irgendwie einzigartig ist oder so, sondern da gibt's unendlich viele Kombinationen, die zum gleichen Ergebnis führen, und das hängt gar nicht unbedingt davon ab, was da intern an irgendeiner Stelle ganz genau passiert. Relativ schwierig zu verstehen. Also wir verwenden das in dem Sinne.«

Dies kann jedoch genauso gut negativ gewendet werden, da aus dem Faktum, dass eine praktische Implementierung erfolgreich auch ohne rigoroses Verständnis erfolgen kann, folgt dann ein hohes Risiko für Manipulation der Ergebnisse (blocking action):

»[...] weil so lange das existiert, und wir nicht verstehen was an Mechanismen wirklich zu den Entscheidung und Ausgaben des Netzwerks führen, man immer das Risiko hat, dass da Dinge passieren, die uns vielleicht nicht gefallen.«

Wir sehen auch wieder die Strategie einer Rückbindung an das Erfahrungswissen der Expert:innen, die intuitiv richtige Anwendungsentscheidungen treffen, weil sie schon häufig mit diesen Verfahren zu tun hatten und dabei implizites Wissen angehäuft haben. Hier sieht man dann eine Begründung von Handlungsfähigkeit aus der Erfahrung der Entwickler:innen und Anwender:innen (getting action):

»Also ich glaube im Moment sind wir im Wesentlichen noch in einem Stadium, wo das erfahrungsgetrieben ist. Also, Leute, die da arbeiten, wissen ungefähr was sie tun können und manche wissen es besser und sind am Ende damit erfolgreicher und das ist zum Teil mysteriös.«

Weit darüber hinaus geht eine weitere Naturalisierungsstrategie, die davon ausgeht, dass was hier Intelligenz genannt wird, immer intransparent ist, dass es sich hierbei sogar um eine definitorische Bedingung für Intelligenz handelt und man deshalb damit leben müsse, dass der Prozess eine Blackbox darstellt (getting action):

»Denn ein intelligentes System ist für mich eigentlich erst intelligent, wenn es schafft, sich selber Regeln zu setzen. Das heißt, wenn ich nicht sage, ich habe vorne Regeln und hinten kommt immer dasselbe raus und ich weiß ja von vornherein, was rauskommt, sondern gerade wenn ich nicht weiß, was das System macht. Also das ist für mich Intelligenz.«

»Also wir verstehen ja auch nicht, wie unser Gehirn so genau funktioniert. Wir verstehen zwar eine ganze Menge, also zumindest manche Leute, aber so ganz genau... naja. (lacht) Weiß ›an's dann doch nicht.«

Demgegenüber stärker zurückgenommen ist die Reduktion der Verfahren auf eine Assistent:innenrolle, eine Absicherung dahingehend, dass Entscheidungsträgerschaft beim Mensch liegen sollte, sodass sichergestellt ist, dass ein Mensch der seine Entscheidung begründen kann hier letztlich die Verantwortung trägt (blocking action):

»I'm much more a proponent of saying let's build systems that are... that are assisting... that are assisting humans, that are enhancing humans, that are not overwriting them, but work alongside them and make them more powerful and show them things that they might have missed.«

Dann gibt es noch die Perspektive des Coping durch den Menschen oder wenn man so will der Koevolution von Mensch und Technik, dahingehend, dass wir mit den Entscheidungen solcher Systeme umzugehen lernen und sie als gegeben hinnehmen, ohne sie zu verstehen (getting action):

»Also ich glaube, das ist ((Pause)) immer ein Phänomen unseres etwas merkwürdigen Umgangs mit der Technik. Dass wir natürlich unser Verhalten darauf anpassen, dass wir damit vernünftig umgehen können. Und wenn das jetzt natürlich so ein bisschen abstraktere, von mir aus KI-Anwendungen sind, dann scheint das vielleicht erstmal noch abstruser, aber das ist natürlich auch beim Umgang mit unseren Autos, Fahrrädern oder sonst

etwas so, dass wir uns auf eigenartige Arten den Dingen anpassen und versuchen uns da durchzulavieren.«

Schließlich ist auch die experimentelle Outputkontrolle wieder wichtig, bei der es darum geht, klarzumachen, ob man den Output experimentell sichtbar machen kann, also dazu kommt ein Setting zu entwerfen, indem man den Mustereffekt sehen kann (getting action):

»Natürlich kann ich eben [...], wenn ich diese Muster selber nicht sehe und mir die Maschine sagt, das Muster ist da, dann ist das natürlich problematisch. Wenn ich dann aber ausprobieren kann, ob da was ist, weil ich dann ein Experiment machen kann, dann gehe ich diesem Problem aus dem Weg. Wenn ich mich nur auf das verlasse, was die Maschine erkennt, ist das schwieriger.«

Es ist wichtig zu sehen, dass gerade dieses Feld sich durch hochkomplexe Kontrollverstrickungen auszeichnet und noch keine festen dominanten Kontrollstrategien etabliert sind. Es kommen hier zahlreiche Rhetorical Fixes für das getting, wie das Blocking Action vor und ringen um Dominanz. Technologische Fixes werden durch automatisiertes Erklären versucht, dies steckt jedoch noch ganz am Anfang und bietet nur scheinbare Kontrollmöglichkeiten.

Intransparenz auch der mathematischen Grundlagen

Bei dieser letzten Intransparenz haben wir es mit dem Problem zu tun, dass hier Algorithmen eingesetzt werden, die selbst Kontrollverstrickungen innerhalb ihrer Architektur produzieren (Netzwerkeffekte), die mathematisch noch nicht vollständig verstanden sind. D.h. hier wird etwas produziert, das zumindest von einem wissenschaftlichen Standpunkt aus noch prinzipiell in seinen Ergebnissen nicht vorhergesagt werden kann. Es passieren Dinge, auch Erfolge, deren Zustandekommen man aber nur plausibilisieren kann, aber nicht mathematisch ableiten. Hier stellt sich dann eine abschließende Grenze für das Verstehen und Erklären ein. Diese Ebene der Intransparenz stellt eine sehr grundlegende Frage, ob etwas funktionieren kann, ohne dass man versteht, warum es funktioniert? Es wird damit gesagt, dass letztlich mathematisch nicht berechnet werden kann, was da wie funktioniert. Gerade in der Anwendung werden die Verfahren zu komplex, um mathematisch nachzuvollziehen, was dort passiert. Es ist grundsätzlich nicht möglich, weil

die Komplexität gleich in mehreren Hinsichten zu groß ist. Erstens sind die informatischen Grundprinzipien maschinellen Lernens noch nicht vollständig verstanden, zweitens sind die Abläufe zu komplex, um klare Mechanismen identifizieren zu können und es kommt letztlich zu einer fehlenden theoretischen Fundierung der gesamten Unternehmung. Eine solche mathematische Fundierung ist natürlich eine anspruchsvolle Form der Kontrolle, die zu erreichen in komplexen Systemen eventuell sehr unwahrscheinlich ist. Fehlt diese Kontrollschleife, ist jedoch letztlich eine Erklärung des Erfolgs dieser Verfahren nicht möglich.

Kontrollversuche sind hier sehr schwer zu finden. Natürlich können die fehlenden theoretischen Fundamente durch mathematische Durchbrüche aufgearbeitet werden, aber darauf zu bauen scheint eher eine langfristige Option zu sein. Die Adäquatheit mathematischer Annahmen kann über Experimente versucht werden, zu überprüfen. Eine Realisierung ist jedoch schwierig und aufwendig. Schließlich gibt es die Möglichkeit, diese Art von Kontrolle aufzugeben und sich auf Kontrollversuche zu stützen, die wir schon kennengelernt haben: Zum einen negativ zu konstatieren, dass theoretisches Verständnis nicht maßgebend für den Erfolg ist oder dass man das grundlegende Unverständnis durch menschlichen Umgang mit der Blackbox beheben kann. Letzteres funktioniert dann wiederum über erfahrungsbasiertes Wissen oder Sensibilitäten, die Menschen im Umgang mit Technik entwickeln und auf die man kompensatorisch bauen kann.

Auch hier lassen sich in den Aussagen der Forscher:innen und Entwickler:innen Kontrollprojekte identifizieren, die diese Arten von Kontrollversuchen kanalisieren:

Die fehlende Erklärbarkeit kann in der Praxis zu einem Fall von Blocking Action werden, weil kein Vertrauen in die Technik aufgebaut werden kann bzw. Misstrauen verbreitet werden kann:

»So, for example if you use it in applications like in medicine and the predictive model says you should give this treatment for this patient and the other treatment for the other patient, the doctor wants to say why, and the system cannot explain, because it's a neural network, you can't explain it.«

»Wenn man [...] eine Anwendung hat, die sicherheitsrelevant ist, die ethisch kritisch sind (sic!), ist das natürlich ein No-Go eigentlich, wenn man nicht versteht, was die Risiken sind, der Methode, die man anwendet.«

Eine Möglichkeit hier dennoch Handlungsfähigkeit zu generieren, also KI-Verfahren einzusetzen, besteht darin eine stärkere Einbettung in Praxis-Domäne vorzunehmen und von dieser Erfolgskriterien abzugreifen (getting action). Dies kann einerseits darauf basieren, dass eine stärker ingenieurwissenschaftliche Orientierung angestrebt wird, die ebenfalls wieder auf die Output-Kontrolle setzt. Funktionieren und auch Nutzen ohne tieferes Verständnis ist das Ziel oder man verlegt sich auf die experimentelle Beweisführung innerhalb der Anwendungsdomäne, also ausprobieren:

»Was auch immer ich für ein abstraktes Theorem habe, was ist die Beziehung zu dem, was ich beobachte. Das kann ich natürlich durch Experimente – das war auch schon unabhängig vom maschinellen Lernen so – durch Experimente eigentlich nur belegen, dass wenn ich jetzt vorhersage, dass der Algorithmus meinerwegen schnell läuft, dann kann ich das auch ausprobieren. Läuft das schnell oder nicht?«

Demgegenüber kann auch eine transdisziplinäre Einbettung in die Wissenschafts-Domäne versucht werden, bei der es darum geht Wissen aus allen möglichen relevanten Gebieten für die Modelle zu integrieren und den datengetriebenen Analysen und deren Ergebnissen gegenüberzustellen, um deren Erfolg zu prüfen (getting action):

»Also ganz ganz wichtig, weitere Herausforderung habe ich schon angedeutet, ist auch gelinkt mit dieser Kausalitätsbetrachtung, dass wir versuchen müssen, die... alles was wir wissen über die Welt, und in Modellen und Theorien, seit Jahrzehnten und Jahrhunderten entwickelt haben, das wir das zusammenbringen und, dass diese Datengetriebene Seite, das hat eine komplett andere Richtung, das wir das mit der Experten Sichtrichtung zusammenbringen und da uns irgendwo treffen, und ich glaube, dann kommen wir in die richtige Richtung.«

Schließlich kann man den neueren KI-Verfahren auch gänzlich die Wissenschaftlichkeit absprechen, sodass man sie mit der Homöopathie vergleicht und sagt man könne an ihre Ergebnisse nur glauben, aber nicht wissen, ob sie wirklich funktionieren (blocking action):

»Also da bin ich ein bisschen provokativ und muss sagen, man kann KI so ein bisschen mit Homöopathie vergleichen, wo auch, ich weiß nicht wie viele

Menschen drauf springen. Die Naturwissenschaftler sagen naja, das funktioniert nicht wirklich, weil wir verstehen nicht wie es funktioniert.«

Auch hier ist also wieder das gesamte Spektrum der Blockade bis zur Öffnung des Handlungsspielraums zu beobachten, wobei hier die technischen Lösungen noch stark hinter den rhetorischen zurückbleiben.

Je nachdem, auf welche Ebene sich die Intransparenz der Systeme bezieht, müssen andere Kontrollformen angewendet werden, deren Realisierung, je größer die fundamentale Intransparenz ist (aufsteigend von 1 zu 4), sich immer weniger auf technologische Fixes stützen kann und auch nicht mehr in übergreifende Kontrollprojekte integriert wird.

5. Verstrickungen von Kontrolle: Wie Blackboxing und Explanability bzw. Interpretability aufeinander bezogen werden

Wenn man Blackboxing als Kontrollprojekt liest, stellt sich in den diskutierten Beispielen immer auch parallel dazu die Frage nach möglichen Kontrollprojekten, die hier in die Gegenrichtung arbeiten, also die Öffnung der Blackbox fordern und nicht nur auf Ergebnisse schauen wollen, sondern auch auf ein Verstehen des Zustandekommens pochen, da nur dann eine erfolgreiche Kontrolle gewährleistet ist. Da heute die Vertrauenswürdigkeit intransparenter Prozesse insgesamt als kritikwürdig angesehen wird, erheben sich Forderungen nach Möglichkeiten, die Intransparenz der KI-Systeme stärker in den Blick zu nehmen und Möglichkeiten zu erkunden, wie dies geschehen kann⁵. Interpretierbarkeit und Erklärbarkeit werden hier zu kommunikativen Vehikeln, die unterschiedliche Formen annehmen. Auch hier geht es darum einige Beispiele einzufügen, wie Kontrolle über Erklärung erreicht werden soll, was dabei mit der Intransparenzproblematik passiert und wer hier wie Kontrolle ausübt. Erklärung kann hierbei als »rhetorical fix« angesehen werden, der Kontrolle auf der Erzählebene ausübt. Gleichzeitig wird in der Diskussion jedoch häufig auch ein »technological fix« angestrebt, bei der die Verfahren quasi auto-

5 Oft synonym verwendet, so beschreibt Explainability, ob es für den Menschen nachvollziehbar ist, warum das KI-Verfahren, diese Ergebnisse produziert und nicht andere, Interpretability beschreibt die Vorhersagbarkeit von Ergebnissen, ohne zwangsläufig die zugrundeliegenden Prozesse im Sinne der Explainability verstehen zu müssen. Siehe: Angelov et. 2021, Mittelstadt et al. 2019, Qj 2021, Rudin 2019 und Somani et al. 2023.

matisiert Erklärungen mitproduzieren, wie das Ergebnis zustande gekommen ist. Wie an den zahlreichen Zitaten oben zu sehen ist, wird das Erklärungsproblem im wissenschaftlichen und Entwickler:innendiskurs sehr eng mit dem Problem der Blackbox diskutiert, da vor allem auf den Ebenen der Anwendung und der mathematischen Grundlagen eine sehr basale Kritik an der Blackbox impliziert wird, die man nur schwer einfach bei Seite schieben kann. Eine solche Verstrickung von Kontrollprojekten ist jedoch für die White'sche relationale Lesart der Produktion sozialer Phänomene charakteristisch und in einem umkämpften Diskursfeld, indem Identitäten nicht festgefügt sind auch eher die Norm. Für uns war an dieser Stelle faszinierend, wie stark diese Verstrickung schon selbst im Feld reflektiert wird, sodass hier grundsätzlich ein hohes Problembewusstsein unterstellt werden kann.

Deshalb sollen die Interviewdaten noch kurz einem Topic Modeling der öffentlichen Diskussion über KI in kulturell verschiedenen Weltregionen (Deutschland, USA, Großbritannien und China) gegenübergestellt werden. Dazu wurden Artikel aus überregionalen Tages- und Wochenzeitungen aus den Jahren 2018 – 2021, welche jeweils mind. einmal »künstliche Intelligenz« bzw. »Artificial Intelligence« beinhalten, ausgewählt. Es wird dann eine Word Cloud aus den Topicmodeling-Ergebnissen generiert, da zu jedem Topic eine Wortliste erzeugt wird, deren Gewicht dann in Abhängigkeit zur Worthäufigkeit in den Texten abgebildet wird. Dabei ist zu beachten, dass die Größe der Wörter nicht alleine auf der Häufigkeit in den Topics basiert, sondern auch auf der »Stärke« des oder der Topics, die die Wörter beinhalten⁶. Nun ist eine Interpretation dieser Wortwolken sicher mit Vorsicht zu genießen, da hier vielfältige Auswahlwirkungen greifen. Doch in Anlehnung an Whites Konzept der Stile, das mit selbstähnlichen Verteilungen von Elementen operiert, die eingesetzt und erkannt werden, lassen sich doch grobe Ableitungen im Hinblick auf die wahrgenommenen Problemfelder und deren Relevanz ableiten.

6 Die Topic Models wurden mit der Software ConText erstellt (Diesner 2014; Diesner et al. 2020). Die Topic Models basieren auf 4 Datensätzen, die jeweils über die Volltextdatenbank LexisNexis erstellt wurden. Die Datensätze beinhalten nur Artikel, die im Zeitraum 01.01.2018 – 31.12.2021 publiziert wurden. Im Datensatz »Deutschsprachiger Raum« sind 2029 Artikel aus Neue Zürcher Zeitung (NZZ), Die Welt, Die Zeit und Rheinische Post enthalten. Im Datensatz »China« sind 2766 Artikel der China Daily enthalten. Der Datensatz »Großbritannien & Irland« beinhaltet 2356 Artikel aus The Guardian, The Times, The Irish Times. Der Datensatz »Vereinigte Staaten« enthält 2596 Artikel der New York Times.

schon aus dieser groben Aufschlüsselung lernen kann, ist, dass die Einbettung über die Ausrichtung von diskursiven Kontrollprojekten entscheidet, weil es zu starken Unterschieden im Kontrollstil zwischen kulturell differenten Regionen kommen kann. Rückgewendet auf die Transparenzproblematik kann man dann unterschiedliche Aspekte der Intransparenz hervorheben und ableiten, ob und welche politischen Maßnahmen eventuell auf diese reagieren werden. Hier sieht man dann die weitergehende Kontrollverstrickung, die aus der Einbettung der KI-Systeme folgt und mitbestimmt, welche Grenzen ihnen gesetzt werden.

Bei der gemeinsamen Betrachtung von Interviews aus dem englischsprachigen Raum und der öffentlichen Diskussion im englischsprachigen und dem deutschsprachigen Raum wird deutlich, dass die Wahrnehmung der Bedeutung von Entwicklungen in der KI durch die kulturellen Stile geprägt sind. Dies unterfüttert eine Beobachtung aus dem Vergleich der Interviews zwischen Deutschland und Kanada, die unterschiedliche Problematiken in Bezug auf das Transparenzproblem und das zugehörige Blackboxing fokussieren. Was in den Bereich des sichtbar zu machenden fällt, kann somit sehr unterschiedlich aufgebaut sein und stellt damit auch unterschiedliche Anforderungen an die Entwicklung von KI-Systemen.

6. Stil des Umgangs mit der Transparenzproblematik: Herausforderung oder grundlegender Defekt

Abschließend können unter Rekurs auf die dargestellten Bearbeitungsformen auch unterschiedliche Stile herausgearbeitet werden. Wenn wir nochmal auf die früher vorgestellte Matrix zurückkommen, können Kontrollverstrickungen als dominante Kombinationen gelesen werden. Während in einem Fall rhetorische Mittel benutzt werden, um Handlungsmöglichkeiten zu beschränken, wird in einem anderen Fall versucht, mit einem technischen Ansatz Handlungsspielräume zu gewinnen und umgekehrt. Hier könnte das von White formulierte Konzept des Stils einschlägig genutzt werden. Stile sind selbstähnliche Prozessmuster, die zugleich Signalwirkung haben, also soziale Identitäten anzeigen, aber auch expert:innenähnliche Beobachtungsfähigkeiten, also Sensibilitäten für solche Muster erzeugen (White 2008; Schmitt/Fuhse 2015). Beim Umgang mit Transparenzproblemen, können dabei verschiedene Stile beobachtet werden, die spezifische Kombinationen von Kontrollversuchen aufweisen. Insbesondere sind es hier die Unterschiede

zwischen einer eher mit dem englischsprachigen Raum verbundenen Sicht- und Herangehensweise und an die Blackbox-Thematik und einer dezidiert deutschsprachigen Herangehensweise.

So dominiert in den Interviews und Zeitungen aus dem englischsprachigen Raum eine pragmatische Umgangsweise mit den Problemen. Diese werden keineswegs gelehnet, sie werden jedoch rhetorisch kleiner gemacht und als technisch oder kommunikativ zu lösende Herausforderung angesehen. Damit werden eher Kontrollstrategien angesprochen, die Probleme naturalisieren und auf experimentelles Ausprobieren setzen, auch wenn es letztlich keine Möglichkeit des Nachvollziehens gibt. Probleme mit KI belegen auch im Topic Modeling eher hintere Ränge, wenn sie überhaupt auftauchen, während zum Beispiel Investitionsmöglichkeiten betont werden. In den deutschsprachigen Interviews kommt dieser pragmatische Fokus auch vor, ist aber deutlich stärker mit kritischen Anstrichen durchsetzt und formuliert die Probleme des Nicht-Verstehens deutlicher und sieht sie als schwerer zu lösen an. Zugleich zeigt auch das Topic Modelling einen stärkeren Fokus auf die mit der Technologie verbundenen Probleme. Zusammenfassend können zwei Stile des Blackboxings beschrieben werden:

»Blackboxing als experimentell zu bearbeitende Herausforderung anzusehen (getting action als Fokus) und Blackboxing als grundsätzliches Problem (2), dass zunächst gelöst werden müsste, um weitreichende Einsatzmöglichkeiten zu rechtfertigen (blocking action).«

Auch hier kann man unterschiedliche Herangehensweisen bei den Interviewten in Kanada und Deutschland hinsichtlich der Einbindung von ethischen Entscheidungen wahrnehmen. Während in Kanada die Forderung einer ethischen Überprüfung an den/die Forscher:in selbst gerichtet wird, geht es für deutsche Forscher:innen nur um Qualitätskontrolle, während die ethischen Entscheidungen an Kommissionen ausgelagert werden, sodass der/die Forscher:in weiß, in welchem Rahmen er/sie sich bewegen darf.

7. Ausblick

Die Idee neuere KI-Verfahren als Blackboxes anzusehen, deren Ergebnisproduktion nicht direkt nachvollziehbar ist, ist als Sichtweise bei Forscher:innen und Entwickler:innen weit verbreitet. Auch gibt es ein großes Bewusstsein für

diese Problematik, die aber den öffentlichen Diskurs noch nicht erreicht hat. Es gibt jedoch unterschiedliche Möglichkeiten mit diesen Problemen umzugehen, die man mit theoretischen Konzepten aus der Theorie von Harrison White, wie Kontrolle, Kontrollversuch und Kontrollprojekt sowie Getting Action, Blocking Action und Stil, sehr gut herausarbeiten kann. Wenn man die Entwicklung der neueren KI-Verfahren noch als umkämpftes Feld neuer Technologieentwicklungen fasst, bieten diese Begrifflichkeiten die Möglichkeit Positionen zu identifizieren und dann auch aus dominierenden Kontrollprojekten Prognosen abzuleiten, was die Durchsetzung in unterschiedlichen Ländern und Anwendungsfeldern angeht. Hier liegt auch noch viel Potenzial in weiterer Forschung, die von diesen ersten Einsichten aus starten kann.

8. Literatur

- Angelov, Plamen P./Soares, Eduardo A./Jiang, Richard/Arnold, Nicholas I./Atkinson, Peter M. (2021): »Explainable artificial intelligence: an analytical review«, in: WIREs Data Mining and Knowledge Discovery 11.
- Callon, Michel/Latour, Bruno (1981): »Unscrewing the big Leviathan: how actors macro-structure reality and how sociologists help them to do so«, in: Karin Knorr-Cetina/Aaron Victor Cicourel (Hg.), *Advances in social theory and methodology: Toward an integration of micro-and macro-sociologies*, London: Routledge and Kegan Paul, S. 277–303.
- Castro, Daniel/McLaughlin, Michael/Chivot, Eline (2019): »Who Is Winning the AI Race: China, the EU or the United States?«, in: Center for Data Innovation von August 2019, <https://s3.amazonaws.com/www2.datainnovation.org/2019-china-eu-us-ai.pdf>.
- Dastin, Jeffrey (2018): »Amazon scraps secret AI recruiting tool that showed bias against women«, in: Reuters vom 11.10.2018, <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G>.
- Diesner, Jana (2014): »ConText: Software for the Integrated Analysis of Text Data and Network Data«, in: *Social and Semantic Networks in Communication Research*.
- Diesner, Jana et al. (2020): ConText: Network Construction from Texts, <https://context.ischool.illinois.edu/>.
- Geitz, Eckhard/Vater, Christian/Zimmer-Merkle, Silke (2020): *Black Boxes – Versiegelungskontexte und Öffnungsversuche: Interdisziplinäre Perspektiven*, Berlin: De Gruyter.

- Langer, Paul F./Weyerer, Jan C. (2020): »Diskriminierungen und Verzerrungen durch Künstliche Intelligenz. Entstehung und Wirkung im gesellschaftlichen Kontext«, in: Micheal Oswald/Isabelle Borucki (Hg.), *Demokratiethorie im Zeitalter der Frühdigitalisierung*, Wiesbaden: Springer, S. 219–240.
- Latour, Bruno (1994): »On Technical Mediation. Philosophy, Sociology, Genealogy«, in: *CommonKnowledge* 3(2).
- Latour, Bruno (1999): *Pandora's hope: essays on the reality of science studies*, Cambridge/Massachusetts: Harvard University Press.
- Latour, Bruno/Woolgar, Steve (1986): *Laboratory life: the construction of scientific facts*, Princeton/N.J.: Princeton University Press.
- Lehmann, Katharina (2021): »Wenn der Code schwarz sieht«, in: *Die Welt vom 19.11.2021*, <https://www.welt.de/wirtschaft/better-future/article235149630/Algorithmen-sind-nicht-objektiv.html>.
- Lloyd, Kirsten (2018): »Bias Amplification in Artificial Intelligence Systems«, in: Frank Stein/Alun Preece/Mihai Boicu, *AAAI FSS-18: Artificial Intelligence in Government and Public Sector*, Arlington: Cornell University.
- Luhmann, Niklas (1991): *Soziale Systeme: Grundriß einer allgemeinen Theorie*. 4. Auflage, Frankfurt a.M.: Suhrkamp.
- Mittelstadt, Brent/Russell, Chris/Wachter, Sandra (2019): »Explaining Explanations in AI«, in: *Proceedings of the Conference on Fairness, Accountability, and Transparency*, New York: Association for Computing Machinery.
- Perrigo, Billy (2023): »Exclusive: The \$2 Per Hour Workers Who Made ChatGPT Safer«, in: *Time vom 18.01.2023*, <https://time.com/6247678/openai-chatgpt-kenya-workers/>
- Probst, Laurent/Pedersen, Bertrand/Lefebvre, Virginie/Dakkak-Arnoux, Lauriane (2018): »USA-China-EU plans for AI: where do we stand?«, Brüssel: European Commission.
- Rudin, Cynthia (2019): »Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead«, in: *Nature Machine Intelligence* 1, S. 206–215.
- Schmitt, Marco/Fuhse, Jan (2015): *Zur Aktualität von Harrison White*, Wiesbaden: Springer VS.
- Somani, Ayush/Horsch, Alexander/Prasad, Dilip K. (2023): *Interpretability in Deep Learning*, Cham: Springer International Publishing.
- Van Aert, Robbie C. M./Wicherts, Jelte M./Van Assen, Marcel A. L. M. (2019): »Publication bias examined in meta-analyses from psychology and medicine: A meta-meta-analysis«, in: *PLoS ONE* 14 (4).

- von Randow, Gero (2018): »A wie Algorithmus. Wer über die Digitalisierung spricht, kommt an drei Grundbegriffen der Informatik nicht vorbei. Was bedeuten sie politisch?«, in: Die Zeit vom 22.02.2018, <https://www.zeit.de/2018/06/informatik-roboter-algorithmus-kuenstliche-intelligenz>.
- White, Harisson C. (1992): Identity and Control. A Structural Theory of Social Action, Princeton/New Jersey: Princeton University Press.
- White, Harisson C. (2008): Identity and Control. How social formations emerge. Second edition, New Jersey: Princeton University Press.
- White, Harrison C./Godart, Frédéric (2007): »Stories from Identity and Control«, in: Sociologica 3.
- Züger, Theresa/Asghari, Hadi (2022): »AI for the public. How public interest theory shifts the discourse on AI«, in: AI & SOCIETY 38, S. 815–828.

Von der Kunst des Lernens

Einige Bemerkungen zur Intentionalität von In- und Output

Claudius Härpfer, Nadine Diefenbach

1. Einleitung

Der Begriff des maschinellen Lernens suggeriert, dass Maschinen lernen. Wie auch immer dieser Prozess im Detail aussieht (vgl. z. B. Shalev-Shwartz/Ben-David 2014; Russel/Norvig 2016: 693–859), in den meisten Verfahren sind nicht nur Maschinen involviert, sondern auch Menschen, die die Maschinen anleiten, indem sie ihnen gezielt Lerndaten zur Verfügung stellen und gegebenenfalls Optimierungen an einzelnen Parametern wie beispielsweise der Gewichtung von Verbindungen im künstlichen neuronalen Netz vornehmen. Konstruktionsbedingt kann der die Maschine anleitende Mensch die eigentlichen Vorgänge im inneren des künstlichen neuronalen Netzes nicht überblicken (vgl. z. B. Burrell 2016), sondern entwickelt nur eine mehr oder weniger vage Vorstellung davon, was konkret im Inneren abläuft, auf deren Basis er das Lernen der Maschine zu steuern versucht. Der Mensch hat also eine in irgendeiner Form sinnhafte, metaphorisch aufgeladene Vorstellung von dem, was die Maschine tut, ohne wirklich zu wissen, wie effektiv und zielführend seine Lenkungsversuche des Lernprozesses sein werden, während die Maschine in den zur Verfügung gestellten Daten Muster erkennt und diese ordnet, indem sie neue Verknüpfungen bildet oder bestehende löst. Diesem komplexen soziotechnischen Phänomen wollen wir uns widmen, indem wir die relevanten Abläufe als Netzwerk soziotechnischer Relationen begreifen, deren Elemente sich vermittelt durch das jeweilige Design treffen und – in einem Prozess der begonnenen Übersetzung – voneinander lernen, ohne sich jemals verstehen zu können. Neben einschlägigen Arbeiten Herbert Simons (1984; 1994) greifen wir hierbei in Anlehnung an Häußling (2012; 2016; 2020) auf das Vokabular der Netzwerktheorie Harrison Whites (1992; 2008)

zurück und verstehen die Interaktionen als Verstrickungen wechselseitiger Kontrollprojekte in sich in diesem Prozess bildenden Netzwerkkomplexen.

2. Das Lernen der Maschinen

Zunächst gilt es einen kurzen Blick auf das Lernen der Maschinen zu werfen. Zwar suggeriert das große alte Narrativ einer künstlichen Intelligenz (vgl. z. B. La Mettrie 2001; Riskin 2007; Domingos 2015), dass Maschinen (irgendwann) in der Lage sind, menschenähnlich zu lernen und dem Menschen ebenbürtig (oder gar überlegen) zu sein. In der Praxis gilt seit Turings Arbeiten in den späten 1940er Jahren (2004) jedoch die Maxime kleiner Schritte und den Menschen, die die Maschinen im Lernprozess unterstützen, ist natürlich bewusst, dass maschinelles Lernen auf probabilistischen Modellen beruht, so dass ein Übermaß an Pädagogik und Entwicklungspsychologie hierzu inkompatibel und deshalb fehl am Platze ist. Wie der Blick in die Anwendungsliteratur zum maschinellen Lernen zeigt, sind die faktischen natürlichen Referenzgrößen nach wie vor weniger Menschen als vielmehr das Lernverhalten von Ratten und Tauben (vgl. z. B. Shalev-Shwartz/Ben-David 2014: 1–3).

Der Blick in die soziologisch-kulturwissenschaftliche Literatur wiederum zeigt die Vielfältigkeit der Anknüpfungspunkte und offenbart ein komplexes Lernsetting. Ausgehend von der Annahme, dass menschliche Intelligenz aus einem Arrangement der Einbettung des Menschen in seine jeweilige Umwelt, und dem Zusammenspiel all jener Faktoren hervorgeht (vgl. Taffel 2019), scheint gerade die Auseinandersetzung mit dem maschinellen Lernen in seiner soziotechnischen Vernetzung relevant. Dies umfasst nicht allein die Betrachtung des Begriffs Machine Learning, sondern ermöglicht zudem einen Beitrag zur Reflexion des menschlichen Lernens. Lernen im systemischen Sinn meint nach Ziegler (2007) die umfassende Berücksichtigung verschiedener am Lernprozess beteiligter Faktoren. Darunter sind nicht nur Zeit und Raum zu subsumieren, sondern auch die am Lernprozess beteiligten Akteur:innen. Ganz im Sinne von Reigeluth & Castelle (2021) gehört zum Lernen eine Lehrperson, die anleitet und unterstützt. Die beiden Autoren begreifen maschinelles Lernen nicht allein als die Verarbeitung von bereits entschiedenen Informationen als soziotechnischen Prozess, der auf der Basis verschiedener Algorithmen verläuft. Maschinelles Lernen umfasst vielmehr die »Anleitung«, das Trainieren sowie die Interpretation von Ergebnissen und daraus gefolgerte Erkenntnisse des Menschen, der nicht auf den »Daten- und

Mittelgeber« reduziert werden kann. Maschinelles Lernen wird in ihrem Sinn als »cultural activity«, als »soziale[r] Prozess« gefasst (Reigeluth/Castelle 2021: 105).

Angesichts dieses Auseinanderdriftens der Perspektiven, ist es nicht verwunderlich, dass Herbert Simons technisch gehaltene (und damit vermittelnde) Definition maschinellen Lernens bis heute relevant ist (Kaminski/Glass 2019). Simon fasste Lernen als »any change in a system that allows it to perform better the second time on repetition of the same task or on another task drawn from the same population.« Diese Änderung soll zu einem gewissen Grad dahingehend irreversibel sein, dass der Lerneffekt nicht ohne Weiteres von selbst wieder verschwindet. Besser durchgeführt heißt für ihn, dass die Aufgabe oder eine Aufgabe aus der gleichen Grundgesamtheit vom System beim nächsten Mal »more efficiently and more effectively« erledigt wird (Simon 1983: 28). Diese Definition enthält ein Element der Wiederholung, also der Regelmäßigkeit. Daher ist es folgerichtig, wenn Kaminski und Glass (2019: 130) im Anschluss den Aspekt der Regelbildung betonen, den sie mit Blick auf die Mittel, den Zweck und die Lernintention selbst strukturieren. Das Finden von Regeln ist jedoch soziologisch betrachtet nur eine Dimension, sich Wissen anzueignen, und kann je nach Art der Regel hoch heterogen sein (vgl. Schütz 1972). Simons Definition ist damit aber keinesfalls erschöpft, denn auch die Parameter für die Bildung von Regeln werden benannt. Simon verweist (1) auf eine Zustandsveränderung, die (2) innerhalb eines Systems stattgefunden hat, deren Ergebnis (3) als besser eingestuft werden kann, also auf einen vorhandenen – wie auch immer gearteten – Maßstab (inkl. Effektivitäts- und Effizienzkriterien) übertragen wird und (4) durch den gleichen Vorgang oder bei der gleichen Population angewandt wird.

Wenn wir versuchen, uns dem Phänomen soziologisch zu nähern, bietet sich der Rückgriff auf die White'sche Theorie von Identität und Kontrolle an. White baut seinen Blick auf die Welt (1992; 2008) um das Wechselspiel der Begriffe Identität und Kontrolle auf. Identitäten streben nach Kontrolle und die Kontrollversuche der Identitäten dienen dazu, andere Identitäten auf die eigene Prozessualität zu verpflichten. Dahinter stecken ein komplex werdender, mehrstufiger Identitätsbegriff und eine prozessuale relationale Perspektive. Der Identitätsbegriff ist dazu in der Lage, auch nicht-menschliche Einheiten adäquat einzubeziehen, gepaart mit der prozessualen Perspektive erwächst daraus eine techniksoziologisch gehaltvolle Perspektive, die wir im Folgenden am Lernbegriff skizzieren wollen.

Any change ...

Bei seiner Beschreibung von Prozessualität verstrickt sich Simon an anderer Stelle in metaphorisch aufgeladene systemtheoretische Verallgemeinerungen über Evolution und Auslese in Systemen (Simon 1994: 149–162). Soziologisch gehaltvoller scheint es uns hier, den Lernprozess als Wechselspiel von Kontrollprojekten zu verstehen. Liest man Whites Theorie systematisch, kann man darin drei unterschiedliche Formen dieser Kontrollprojekte finden (vgl. Häußling 2012: 269–272). Erstens die sogenannte »social ambage« (White 1992: 106–107), also die direkte oder indirekte Einflussnahme auf das Gegenüber mittels bestehender oder möglicher Relationen. Zweitens die »cultural ambiguity« (White 1992: 103–106), also die Kontextabhängigkeit der Deutung von Tatbeständen, was einerseits Offenheit hinsichtlich möglicher Deutungsmöglichkeiten bedeutet, andererseits aber auch Abhängigkeit von bestehenden Deutungsmöglichkeiten. Drittens das »decoupling« (White 1992: 12–13; 111–112), also das Vermeiden oder Lösen von Relationen. Whites Kontrollbegriff steht sein Identitätsbegriff gegenüber (White 2008: 9–12), mit dem wir in der Lage sind, auch Maschinen als Identitäten zu fassen, ohne sie mit menschlichen Identitäten gleichsetzen zu müssen.

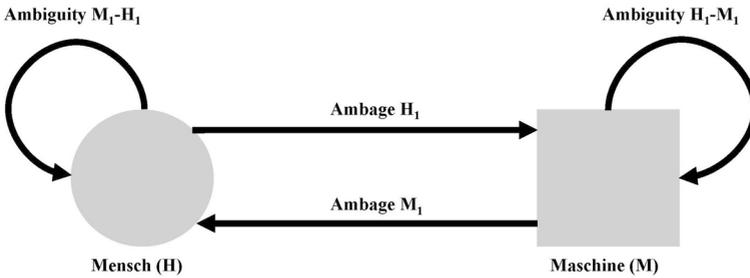
Bezogen auf unser Problem haben wir Formen der Einflussnahme (ambage) durch den Menschen auf die Maschine durch die Eingabe von Daten. Diese Daten interpretiert die Maschine in Abhängigkeit der bisher vorhandenen Daten (ambiguity). Die Maschine wiederrum gibt Informationen über die erkannten Muster aus (ambage), die dann der Mensch auf Basis des ihm bekannten interpretiert (ambiguity) und mit seinen der Maschine gesteckten Lernzielen abgleicht.

Daneben steht die Form der Einflussnahme des Menschen durch Änderung einzelner Parameter, wie der Gewichtung von Verbindungen im neuronalen Netz (ambage). Auf diese Veränderungen »reagiert« die Maschine durch Neuinterpretation der bereits vorhandenen Daten. Ob die Interpretation der Maschine hier nun die Intention des Menschen trifft, ist offen (ambiguity). Auch hier gibt die Maschine wiederrum Informationen über die erkannten Muster aus (ambage), die dann der Mensch auf Basis des ihm bekannten interpretiert (ambiguity).

Das in beiden Fällen Elemente – seien es Datensätze oder Muster – nicht oder nicht mehr berücksichtigt werden (decoupling), liegt in der Natur der Sache. Die Eingabe von Daten seitens des Menschen wollen wir im Folgenden als

Training der Maschine bezeichnen, das Ändern einzelner Parameter hingegen als Lehren bzw. Lernen.

Abbildung 1: Grundschemata der Kontrollprojekte



Die Betrachtung der wechselseitigen Bezugnahme der an der Schnittstelle vollzogenen Interpretations- bzw. Lernvorgängen durch die beschriebenen Kontrollversuche hilft dabei, die Kunst des maschinellen Lernens in ihrer soziotechnischen Konstruktion differenziert zu betrachten. Werfen wir also einen genaueren Blick auf den Rest von Simons Definition.

... in a system ...

Wenn Simon von einem System spricht, so ist klar, dass er keine soziologische Systemtheorie, sondern eine allgemeine Systemtheorie im weitesten Sinne vor Augen hat (Simon 1994: 144; 1956: 74–79). Er vermeidet eine formale Definition und spricht stattdessen von einem komplexen System als einem Gebilde, »das aus einer großen Zahl von Teilen zusammengesetzt ist, wenn diese Teile nicht bloß in der einfachsten Weise interagieren. In solchen Systemen ist das Ganze mehr als die Summe der Teile – nicht in einem absoluten, metaphysischen Sinn, sondern in dem wichtigen pragmatischen, daß es keine triviale Angelegenheit ist, aus den gegebenen Eigenschaften der Teile und den Gesetzen ihrer Wechselwirkung die Eigenschaften des Ganzen zu erschliessen.« (Simon 1994: 145) Die Einwirkung der Daten, der Interpretation der Ausgaben der Maschine und der weiteren Arbeit an den Daten und Algorithmen sowie der Austausch in der Fachcommunity durch beispielsweise Veröffentlichungen verändert in der Erweiterung des Netzwerkes des lernenden Menschen das (maschinelle) Lernen selbst. Die aus den Lernprozessen resultierende Entwicklung wird folg-

lich erst durch das komplexe Zusammenwirken von Menschen und Maschinen möglich, insofern scheint es plausibel, dass Simon hier eine Einheit mit komplexem Innenleben denkt.

In Whites Terminologie lässt sich eine derartige Einheit als Netdom begreifen. Netdom ist eine Kombination der Begriffe »Netzwerk« und »Domäne« (White 2008: 7) und benennt jene Einheit, die entsteht, wenn durch die Kontrollbemühungen von Identitäten Netzwerke eines bestimmten Inhaltes, und damit einer abgeschlossenen Sinneinheit entstehen. Dass diese abgeschlossene Sinneinheit innerhalb eines Netzwerkkonzeptes steht, womit es sich natürlich nur um eine relative Abgeschlossenheit handeln kann, sollte angesichts von Simons pragmatischem, relativierten Systemverständnis eher eine Akzentverschiebung als ein gravierender Einschnitt sein, zumal Simons Fokus auf dem liegt, was an den »schmalen Schnittstellen« zwischen den Systemen, als den »inneren und äußeren Umgebungen« passiert (Simon 1994: 97).

Als Netdom »Machine Learning« besteht, so gesehen ein Netzwerk als Strukturform von sozialen und technischen Identitäten des maschinellen Lernens, das zudem kulturell unter anderem darüber bestimmt, wie zum Beispiel Wissen über Trainingsmethoden, Learner oder Datensätze geteilt wird. Netdoms setzen Beziehungen unter einem bestimmten Thema und dessen Verständnis voraus, diese sind jedoch, wie angedeutet, nicht absolut, sondern prozessual und transitorisch. Mit einem besonderen Strukturmuster in Verbindung mit spezifischen kulturellen Formen ist ein Netdom ein spezifischer Ausschnitt mit je eigener Einbettung in ein gesamtes Netzwerk. Dabei kann ein Netdom eine dyadische Beziehung abschließen und sehr klein und vergleichsweise flüchtig sein, oder ein komplexes Ganzes mit einer Vielzahl unterschiedlicher Akteur:innen und Beziehungen abbilden.

Das Netdom des »Machine Learnings« lässt sich somit als die dynamisch wechselseitige Prägung von Identitäten mit ihren je verschiedenen Kontrollmaßnahmen fassen. Diese wirken sich sowohl auf die kulturellen als auch die strukturellen, aber auch die Einbettungszusammenhänge aus, die sie mitformen. Das »Machine Learning-Netdom« umfasst dabei nicht nur kommunikativ viele Bereiche von Forschung oder Entwicklung, sondern fasst zudem spezielle Relationen der soziotechnischen Interdependenzen an der schmalen Schnittstelle zwischen Mensch und Maschine entlang von unterschiedlichen

Lern- und Trainingsformen,¹ die in ihrer je spezifischen Einbettung verschiedene Kopplungs- und Entkopplungsprozesse bedeuten können.

... to perform better...

Simon begreift den Lernfortschritt als Effizienz- oder Effektivitätssteigerung, was bei einem klar definierten Problem einfach zu quantifizieren ist, in einem komplexer werdenden Setting natürlich zunehmend schwieriger zu fassen ist. Er selbst spricht das Problem der Verarbeitung natürlicher Sprachen (als »very annoying part of the task«) und die damit verbundene Komplexitätssteigerung an (Simon 1983: 29f.). Im Falle einer derartigen gegebenen Komplexität und der damit verbundenen unendlichen Verbesserbarkeit, lässt sich jede Optimierung zwar theoretisch ins Unendliche iterieren, kommt aber in der Praxis an ihre Grenzen, an deren Ende eher sich stilistisch unterscheidende erwünschte Varianten stehen als klar hierarchisierbare Alternativen, die besser oder schlechter sind (Simon 1994: 112). Weswegen Simon auch eher von »Satisfizierungs-« als von »Optimierungsfragen« spricht (ebd. 26). Bei satisfizierenden Methoden kommt es nicht darauf an, ob das Optimum definiert oder alle Alternativen ausgeschöpft sind, sondern ob ein mit Blick auf Kosten-Nutzen-Erwägungen befriedigendes Ergebnis vorliegt (ebd.: 95–119).

In unserem Anwendungsbeispiel entwickelt der Mensch neben allen Indikatoren, die ihm zur Verfügung stehen, ein Gefühl dafür, ob das neue Ergebnis nun besser ist als das alte und damit gegebenenfalls ein zufriedenstellendes Ergebnis vorliegt oder nicht, und deshalb eine weitere Iteration folgen muss. Dies geschieht in Whites Terminologie über ein Narrativ, über »Stories«, also kontextuierte Verbindungen (White 2008: 20). Die zustande kommenden Netzwerke werden durch sie umgebende Erzählungen gefestigt, sodass ein Maßstab entsteht, der dem Menschen hilft, die Leistung der Maschine einzuschätzen.

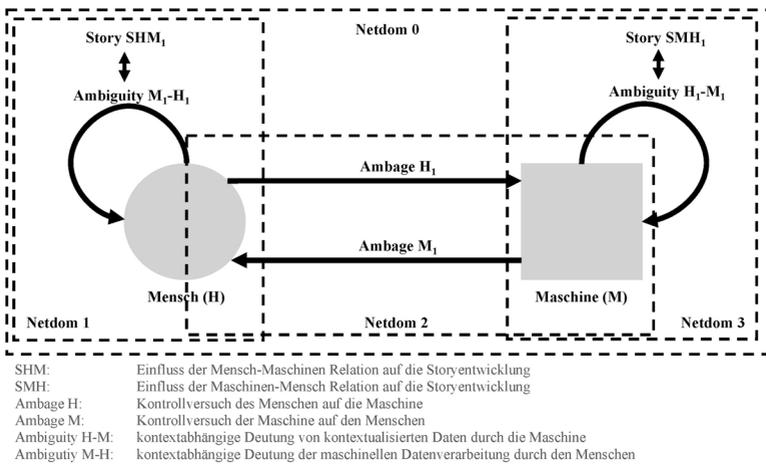
... the second time.

Gleichzeitig ordnen die Stories aber auch die soziale Zeit und grenzen jene Zustände ein, die wiederholt werden. Im Prozess des Vergleichens stellt der Mensch also die Leistung der Maschine zum Zeitpunkt t_1 , der Leistung der

1 Wir folgen hier der Unterscheidung von Lernformen und Lernstrategien nach Kaminski/Class (2019: 130).

Maschine zum Zeitpunkt t_2 gegenüber. Diesen Übergang bezeichnet White als »Switching« (White 2008: 2, 20). Er geht davon aus, dass die Identitäten sich entwickeln, indem sie von einem Netdom zum nächsten wechseln und so eine Spur durch verschiedene Kontexte legen und eine zunehmend komplexere Geschichte schaffen. Bezogen auf den Lernprozess lässt sich das so verstehen, dass unser Modell noch etwas komplexer werden muss.

Abbildung 2: Lernprozess als Switching zwischen Netdoms zum Zeitpunkt t_1



Wenn wir also die vorher kurz beschriebenen Veränderungsprozesse nochmals aufgreifen und uns hier auf den ersten der vorhin beschriebenen Fälle beschränken, heißt das holzschnittartig, dass die Identität Mensch zum Zeitpunkt t_1 im Zuge der Eingabe von Daten das Netdom 1 verlässt und durch die Ambage H_1 eine Relation zur Maschine eingeht, also das Netdom 2 aufspannt. Die Identität Maschine wiederum – die durch Ambage H_1 Teil von Netdom 2 wird, verarbeitet diese Daten unter Rückgriff auf die Story SMH_1 in Netdom 3. Woraufhin sie im Rahmen der Ausgabe der erkannten Muster, also Ambage M_1 wieder eine Relation zum Menschen eingeht und Netdom 2 betritt. Der Mensch nimmt diese Ambage auf, wechselt in Netdom 1 und interpretiert die Muster der Maschine unter Rückgriff auf die Story SHM_1 , wobei er hier nun vergleicht, ob die erzielten Ergebnisse aus seiner Sicht eine Verbesserung gegenüber seiner Referenzgröße, dem Zustand t_0 , darstellen und ob sie bereits

zufriedenstellend sind. Wiederholt der Mensch diesen Vorgang, so erhält er in SHM_n immer mehr verknüpfbare Referenzgrößen, ebenso wie die SMH_n der Maschine.

Whites Theorie spricht von Identitäten und hat dabei ein mehrstufiges Konzept im Hintergrund (White 2008: 10–12), was erlaubt, Mensch und Maschine gleichermaßen im Modell zu integrieren, ohne sie jedoch gleich zu machen. Die Identität Mensch ist fraglos entschieden komplexer als die Identität Maschine in diesem Fall, denn die Software ist für diesen Zweck entwickelt, der Mensch hingegen hat ein Leben darüber hinaus. Auch die Stories, mit denen die beiden die Kontrollversuche des jeweils anderen interpretieren, unterscheiden sich signifikant. Die Maschine arbeitet mit einem System statistischer Zusammenhänge, der Mensch interpretiert mit Blick auf aufbereitete statistische Zusammenhänge und geht im Zweifelsfall mit etwas Bauchgefühl darüber hinweg. Für diese Form des Schließens prägte bereits Hermann von Helmholtz Mitte des 19. Jahrhunderts den Begriff der »künstlerische[n] Induction« (Helmholtz 1896: 171), die überall dort Anwendung findet, wo die Gegenstände so komplex sind, dass sie sich nicht mehr klar überschauen lassen. Der Begriff zielt auf herausragende Kunstwerke, die den Prozess in Reinform veranschaulichen. Er wird aber auch überall dort gebraucht, wo die Schließenden ihre Urteile auf Basis ihres Instinktes oder »psychologischem Tactgefühl« treffen (Helmholtz 1896: 172). In der Folge steht die Frage im Mittelpunkt, wie maschinelles Lernen als Produkt der Relation zwischen Mensch und Maschine im Spannungsfeld zwischen dem Training und Lehren/Lernen mit bzw. durch Daten sowie deren Interpretation zu betrachten ist.

3. Formen des Lernens

Wir wollen nun zwei in der Machine Learning-Debatte prominent diskutierte Trainingsmethoden systematisch entlang des skizzierten Lernprozesses betrachten und aufbereiten: das *Supervised Learning* und das *Unsupervised Learning*. In beiden Lern- bzw. Trainingsformen möchten wir die Interdependenz der soziotechnischen Relationen in der Aushandlung von Kontrolle und Identität differenziert nachvollziehen.

Beide Lernformen sind nicht nur eingebettet in das Netdom »Machine Learning«, sondern bilden auf besondere Weise je eigene Netdoms, in deren Prozess weitere Netdoms aufgespannt werden. Verschachtelt wirken sie in

dynamischer Kohärenz von Struktur und Kultur auch in- und aufeinander ein (Abbildung 2). Hinzu kommt, dass die jeweiligen Identitäten und deren Kontrollbemühungen, wie beschrieben, zwischen verschiedenen Netdoms hin- und herspringen. Diese Switchings (White 2008: 7) wiederum beeinflussen die einzelnen Netdoms.

Für die differenzierte Ausleuchtung der Prozesse greifen wir auf neun von Mitarbeiter:innen des Lehrstuhls für Technik- und Organisationssoziologie der RWTH Aachen University 2020 geführte Expert:inneninterviews zum Thema Machine Learning zurück, in denen die Expert:innen auch Auskunft zu einigen Lern- und Trainingsmethoden und ihren Umgang mit diesen geben. Die Ausrichtung der Interviews ist allerdings allgemeiner gehalten, sodass die Aussagen hierzu eher am Rande der aufgespannten Narrative gefallen sind und mehr exemplarischen Charakter haben. Selbstverständlich fokussiert sich der Blick damit auf die Stories der Menschen.

Um nun die einzelnen Kontrollbemühungen in den jeweiligen Lernprozessen konkreter beschreiben zu können, werden diese zunächst als klar voneinander unterschieden markiert und dargestellt, um ihrer entlang von Identitäten engen und komplexen Vernetzung schließlich wieder Rechnung zu tragen. In der idealisierten Trennung werden wir uns auf den Umgang mit Daten als Interface beziehen. Daten werden in diesem Kontext nicht als Substrat objektiven Wissens (Cardoso Llach 2018) betrachtet (vgl. z.B. ebd.; Häußling 2020). Sie bilden vielmehr eine Schnittstelle zwischen den heterogenen Identitäten in und über die einzelnen Netdoms hinaus. Die (Ent-)Kopplungsprozesse werden im Folgendem entlang dieser Netdoms genauer beleuchtet, um die Trainings- und Lernprozesse in ihrer Komplexität als soziotechnische Aushandlung von Identität und Kontrolle auszuleuchten und so dem Einfluss der heterogenen Entitäten auf die Spur zu kommen. Dabei wird zunächst die Lernform Supervised Learning betrachtet, dem die Ausführungen zur Lernform des Unsupervised Learnings folgen.

Das Netdom »Supervised Learning«

Grundsätzlich geht es beim maschinellen Lernen um die Steuerung von Maschinen durch Daten. Die Maschinen sollen in den Daten Muster erkennen und diese wiederum auf neue, noch nicht bekannte Daten anwenden (vgl. z.B. Lenzen 2018; Sudmann 2018). Supervised Learning umfasst dabei eine Form des Lernens, bei der eine nicht bekannte Funktion von einem Algorithmus erlernt werden soll, die als Regel das »Verhältnis von Eingabe- und Ausgabewerten«

(Kaminski/Glass 2019: 130) aufstellt. Bei dieser Form werden dem Algorithmus bekannte, annotierte Eingabe- und Ausgabewerte zur Verfügung gestellt, aus denen die Maschine schlussendlich eine Hypothese ableiten können soll (ebd.). Ein wichtiger Aspekt dabei ist, dass die Hypothese, die sich dem »Verhalten der gesuchten Funktion« (ebd.) annähert, vom Nutzenden aufgestellt werden muss.

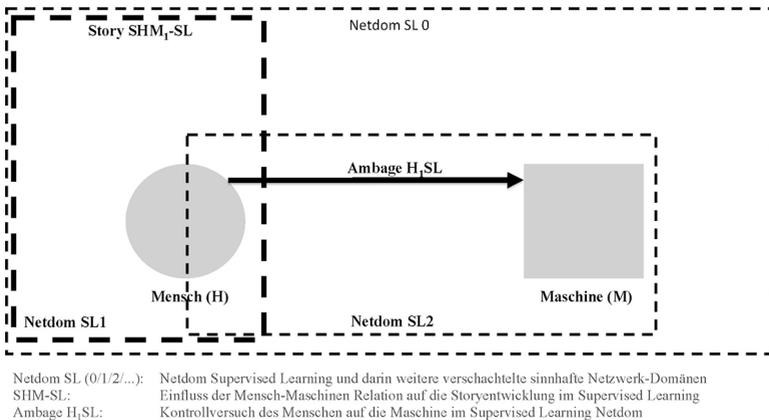
Folglich steht die Frage im Raum, wie groß der Einfluss des Menschen auf den Lernprozess als solchen überhaupt ist und wie stark er das künstlich intelligente Modell durch seine Entscheidungen prägt?

Innerhalb dieses Netdoms (Netdom o/Supervised Learning (SL)) liegen zunächst Entscheidungen für diese Lernform von Seiten der entwickelnden, aber auch anlernenden menschlichen Akteur:innen (vgl. I04) zugrunde. Die Entscheidung für die Lernform des Supervised Learnings hängt dabei zum Teil von pragmatischen Grundlagen ab, die unter anderem auf der Menge der Daten und deren vermuteter Qualität fußen (vgl. I02, Pos. 48f.).

»Also ich würde sagen, wenn man Trainingsdaten verfügbar hat, dann ist es wahrscheinlich schon sinnvoll, erstmal Supervised Ansätze zu probieren, weil die einfach relativ, gefühlt zumindest, robustere Ergebnisse dann direkt liefern.« (I07, Pos. 52–53)

Ist diese Entscheidung getroffen, wird eine nächste Einflussquelle im Netdom Supervised Learning (siehe Abbildung 3) offenbar: Der Mensch muss die Daten, die er dem Learner zur Verfügung stellen möchte, zunächst erst einmal annotieren, um das zu lernende Muster für die »Maschine« offensichtlich vorzuprägen. Die Ambage von Seiten des Menschen (H_1) erfordert in diesem Fall also ein hohes Maß an gezielter Vorbereitung des Lehrenden.

Abbildung 3: Netdom Supervised Learning – SL1 (t1)



Durch diese engere Kontrolle des Menschen kommen im Fall des Supervised Learnings auch stärkere »menschliche Einschätzungen, Vorurteile oder so etwas rein [...], andererseits gibt es das für alle Daten, dass die verfälscht sein können und das ist immer eine Frage, die man sich stellen muss, was die Datenqualität ist und was man davon erwarten kann.« (Io2, Pos. 49)

Durch Daten erlernte, möglicherweise diskriminierende Muster könnten auch durch neue Datensätze wieder verlernt oder Schiefstellungen überschrieben werden (vgl. I15, Pos. 52–54). Im Supervised Learning könne gerade in der Reflexion dieses Umstandes durch die Lehrenden darauf reagiert werden.

»Also einfach Trainingsdaten füttern, die das nicht haben oder man kann natürlich auch versuchen, systematisch gegenzusteuern, indem man systematisch trainiert [...]. Das ist ja bei einem Menschen auch, wenn man etwas falsch gelernt hat. In der Psychologie, wenn Leute irgendwelche Ängste haben, dann sagt man ja auch: Konfrontation mit der Angst und irgendwie darüber lernen.« (I15, Pos. 52)

Die Vorbereitung der Datensätze für die Ambage H₁SL, verbunden mit der Verantwortung, Schiefstellungen zu identifizieren, wird für die Akteur:innen jedoch dadurch erschwert, dass fachfremde Daten nicht leicht zu bewerten sind, gerade wenn diese von Expert:innen anderer Disziplinen annotiert wurden.

»Man sieht es natürlich, wenn man jetzt irgendwie sagt, er [z. B. ein Biologe] soll jetzt die Struktur ganz genau umranden, oder so, und hat dann einfach irgendwie bisschen schludrig dann irgendwie gezeichnet, dann ist das ein relativ offensichtlicher Fehler, den man auch leicht ausmerzen kann. Teilweise ist es auch so, dass die Bilddaten im Prinzip das nicht hundertprozentig hergeben, dass man es wirklich exakt annotieren kann.« (I10, Pos. 71)

Wie hoch die Genauigkeit der Annotation der Daten ist, kann zum Teil nur durch Experten:innen begutachtet werden. Das wäre der »Idealfall, dass man einfach genau das gleiche Bild, die gleiche Aufgabe einfach verschiedenen Experten gibt und dann so eine Art interrater reliability quasi ausrechnen kann, um dann eben zu sehen, wie stark schwankt im Prinzip die Annotierungsgenauigkeit zwischen den Experten.« (I10, Pos. 73)

Dies ist aber »immer nur begrenzt möglich, weil natürlich auch die Arbeitszeit von diesen Experten teuer« (I10, Pos. 73) ist.

Die Datenannotation ist damit Teil der Storyentwicklung (SHM₁-SL) im Netdom Supervised Learning (SL₁), bei der festgelegt wird, mit welchen Daten in welcher Form der Learner trainiert werden soll. Die Verkopplung und Entwicklung einer Story findet dabei nicht nur zwischen den Lehrenden und der »Maschine« statt, sondern wird zum Teil auch in transdisziplinären Teams geschrieben:

»Also gerade in Kooperation mit Biologen, die dann eben die Daten im Labor erstmal aufnehmen. Dann im Prinzip uns die Rohdaten geben [...]. Genau, dann geht es im Prinzip los, dass der Biologe quasi sagt: »Die und die Strukturen interessieren mich«. Und dann basteln wir zum Ersten quasi Tools, mit denen man diese Annotierung überhaupt machen kann [...], sowas wie Paint oder Photoshop in die Richtung. Dass man wirklich auf den Bildern quasi zeichnen kann und dann eben so eine Art Maske sich generieren kann von den Objekten, die man haben will. Genau, das ist so der erste Schritt, dass man einfach mal weiß, was will der Biologe überhaupt haben und dann dafür eben dann diese Label erstellt.« (I10, Pos. 12)

Mit der Einflussnahme des Menschen auf die Daten im Prozess der Annotation ist eine »Verantwortung« (I04, Pos. 57) für den gesamten Entwicklungsprozess verbunden. In dieser Phase scheint der Mensch in der Relation zwischen ihm und der »Maschine« die Prozesse zu dominieren.

Der Einfluss der Daten auf die Mustererkennung insgesamt ist im Sinne der Storyentwicklung von den lehrenden Akteur:innen intendiert, aber was an

Mustern vom Learner erkannt wird, also über das Kontrollprojekt der Ambiguity H1-M1-SL, und wie diese auf andere Daten angewendet werden können, ist gerade nicht ersichtlich.

»Der Algorithmus wird mit Trainingsdaten trainiert, um das zu lernen, was der in den Trainingsdaten sieht. Das ist natürlich nicht unbedingt intendiert, was da rauskommt. Es ist auch nicht intendiert, dass er, sagen wir mal, auch ungewünschte Schiefstellungen in den Trainingsdaten übernimmt. Aber das passiert eben, das ist so. Und das ist auch problematisch.« (I02, Pos. 31)

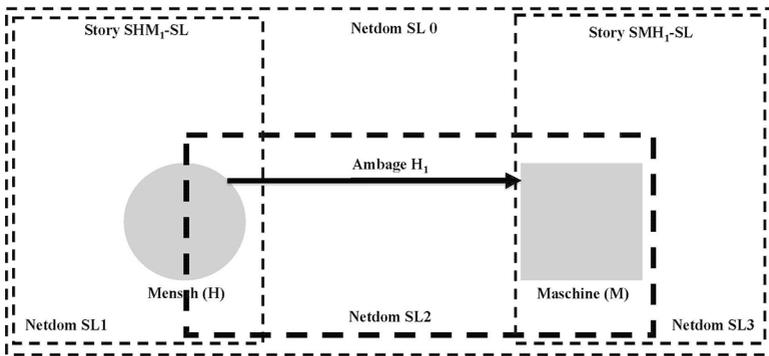
Lehrende Akteur:innen führen damit gerade bei der Datenannotation konsequent auch die Möglichkeit von nicht-intendierten Nebenfolgen mit:

»Da hat man halt Schwierigkeiten [...], dass sich Vorurteile verschieben und verstärken und ähnliches. Das ist schwierig in dem Sinne, weil man das nur sehr schwer kontrollieren kann und das dann auch Dinge sind, die man selber nicht gut versteht. Das ist schwierig, wenn sich diese Sachen tatsächlich verstärken zum einen und dann durch etwas durchgehen, wo man vielleicht eine höhere Objektivität erwartet.« (I02, Pos. 35)

Das Problem des Overfittings durch spezielle Daten(-sätze) und damit die mögliche Verstärkung eines vorhandenen Biases (vgl. I02, Pos. 57) wird bei dieser Lernform erkannt, aber dem soziotechnischen Lernprozess im Supervised Learning förmlich inhärent und als nicht einfach aufzulösen gekennzeichnet. Der Lernprozess im Netdom Supervised Learning kann damit auch als eine Form des Interfacings in Anlehnung an Karafillidis (2018) betrachtet werden, in den sich u.a. die bereits vorab soziotechnisch konstruierten und zu keiner Zeit rohen Daten (vgl. Häußling 2020) als Schnittstellen in die entwickelten Modelle sowie die mit ihnen entworfenen Stories einschreiben. In dieser Daten-Ambiguität und der daran anschließenden tieferliegenden und selbsttätigen Mustererkennung durch den Learner ist der Einfluss des Menschen über die Lernform des Supervised Learnings als Kontrollversuch beschreibbar (Ambage H1 in Netdom SL2). Am Beispiel der Spracherkennung wird beim Supervised Learning der uneinsichtigen Seite des Audiosignals der Spracherkennung durch Vertextung zu begegnen versucht, das heißt, es liegen beim Supervised Learning sowohl die Audiosignale als auch die Transkriptionen dieser vor. Bei der Arbeit mit einem Unsupervised Learner würden eben jene Verschriftlichung fehlen (vgl. I07, Pos. 74).

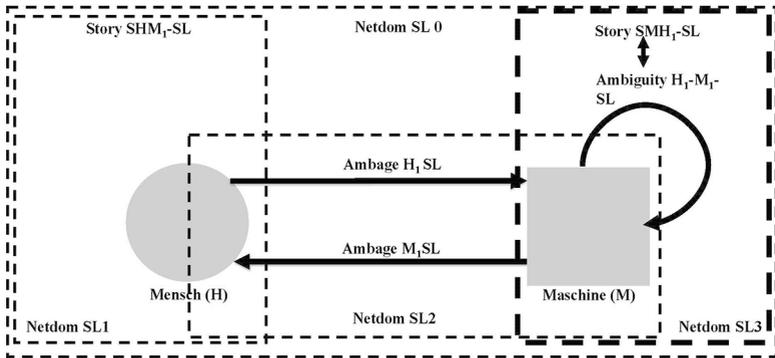
Zurück zu unserer Abbildung. Auf der Basis der entwickelten Story (SHM₁-SL) im Netdom SL₁ wird im Idealfall ein »Trainingsdatenset« für die Entwicklung eines Modells zusammengestellt und dem Learner zur Verfügung gestellt. Damit wird im Prozess des Lernens in das Netdom SL₂ gewischt (Ambage H₁SL im Netdom SL₂).

Abbildung 4: Netdom Supervised Learning – SL₂



Die Ambage H₁ durch den Menschen über die Datensätze im Netdom SL₂ geht über in den Prozess der Verarbeitung der Datensätze, der Mustererkennung, durch den Learner. Dieser Prozess bedeutet ein erneutes Switching in das Netdom SL₃. Hier wird die Verarbeitung der Daten im Sinne der Ambiguity H₁-M₁-SL verortet. Wie und auf welche Weise Muster durch den Learner erkannt werden, bleibt dabei jedoch intransparent. Die Opazität der Datenverarbeitung wird zwar von Seiten des Menschen reflektiert, kann aber nicht aufgelöst werden. Die Datenverarbeitung und Mustererkennung inklusive der Intransparenzen werden folglich durch ein erneutes Switching in das Netdom SL₂ als Ambage M₁SL wirkmächtig. In unserer Abbildung lässt sich der Prozess wie folgt illustrieren:

Abbildung 5: Netdom Supervised Learning – SL3



Der Mensch rechnet gewissermaßen mit der im Netdom SL3 entstehenden Intransparenzen bei der Mustererkennung, welche über die Ambage M_1SL in irgendeiner Art und Weise durch die Datenausgaben wirkmächtig werden.

Die Daten und deren materielle Repräsentation in Datenpaketen, deren Design², werden im Netdom SL1 vom Menschen interpretiert. Dieser Vorgang kann als Ambiguity M_1-H_1-SL beschrieben werden. Die soziotechnisch konstruierten Datenpakete sind in diesem Kontext als Interfaces zu verstehen (i.A. Cardoso Llach 2018; Häußling 2020). Es kann damit zudem eine Diskrepanz zwischen den bereits konstruierten Datensätzen in Abhängigkeit von der Story SHM_1-SL und der Datenannotation durch inter-, trans- oder intradisziplinäre Teams, die zur Visualisierung bereitstehen, entstehen.

»Wir bekommen dann letztendlich die Zeiten ausgegeben [...]. Also das sind dann schon, ja ich nenn sie mal nicht Schaltpläne, die wir dann ausgegeben

2 Design soll in diesem Kontext in Anlehnung an Häußling (2012, 2016) als das »Erfinden neuer Sozialpraktiken« (Häußling 2012: 283) betrachtet werden, weshalb hier von einem Arrangement gesprochen wird (vgl. ebd.: 283). Mit den bei Debray entlehnten Begriffen der »organisierte[n] Materie« und »materialisierte[n] Organisation« (Häußling 2016: 33) kann die Mehrdeutigkeit der Materialität in diesem Prozess differenzierter betrachtet werden. Organisierte Materialität umfasst dabei die Manifestation von Symbolen in Form von Materie und materialisierte Organisation hingegen beschreibt die Übertragung von Ideen durch beispielsweise Technik (vgl. ebd.: 34). Der erste Begriff bildet die Seite der Materialität, die die »Übersetzung der menschlichen Welt für die Technik leistet« (ebd.: 38). Im zweiten Fall wird auf die Einsetzung eines gestalteten Objektes für »menschliche und soziale Belange« (ebd.: 39) verwiesen.

bekommen, aber wir bekommen dann halt ein interpretationsfähiges Bild, mit dem wir dann halt weiterarbeiten können. Und wenn wir das dann wieder auf so einen Schaltplan umrüsten [...] [und] man die Schaltzeiten an den und den Positionen ändert, [dass] ist dann, ja, also ein sehr, sehr interpretationsfähiges Bild, auch mit sehr vielen Diagrammen und Graphen, damit man das dann auch, ich sag mal an die Menschen, die da an den Entscheidungshebeln sitzen, herantragen kann.« (I06, Pos. 50)

Zwischen der Datenausgabe und den daraus gezogenen Informationen entsteht in diesem Fall eine Differenz (Koren/Klamma 2018), die ausgeglichen werden muss, damit Daten interpretierbar bleiben. Es liegt hier also nahe, dass zwischen den Lernern in beiden Nettdoms über (un-)bestimmte Datenpakete zudem weitere (technische) »Übersetzer« wie zum Beispiel technische Analytic Tools (ebd.) stehen, die die (Weiter-)Verarbeitung von Informationen erst möglich machen.

Das Design der Datenpakete nimmt im soziotechnisch konstruierten Supervised Learning aufgrund der Selbstorganisiertheit der Learner in der Verarbeitung der Datensätze, der Ambiguity H_1 - M_1 -SL, einen weiteren Einfluss auf die Arbeit am und mit dem Lernmodell ein und offenbart weitere soziotechnische Ver- und Entkopplungsprozesse, die auf den gesamten Lernprozess wirken können. Auch an dieser Stelle offenbart sich ein Übersetzungsproblem zwischen »Maschine« und »Mensch«, dass sich im Kontext künstlich intelligenter Technologie als ein Problem des Berechnens von etwas schwer Berechenbarem zeigt.

»Also die Idee des ganzen Ansatzes ist gerade die, dass man eben nicht von Anfang an versucht, als Mensch Kontrollstrukturen aufzubauen, weil man sowieso immer etwas vergisst und das einfach nicht hinkriegt.« (I02, Pos. 47)

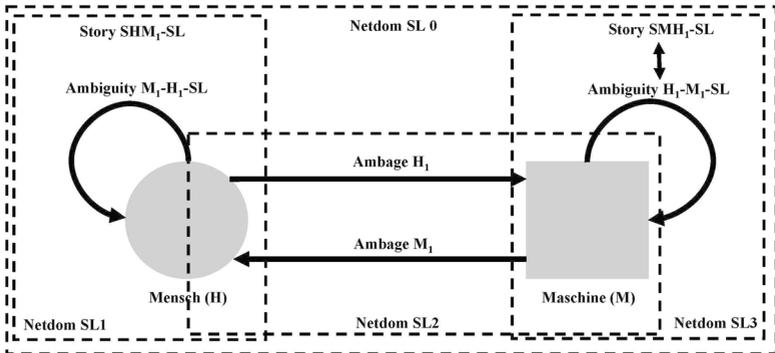
Das Problem des Verstehens, was auf welche Weise berechnet wurde, kann dabei auch im Design des Modells selbst verankert sein (vgl. I02, Pos. 47).

Dieser »zirkuläre Prägungszusammenhang« (Häußling 2012: 265) zwischen den Datensätzen, und den »Ergebnissen« künstlich intelligenter Prozesse ist in Bezug auf künstliche Intelligenz nicht nur, dass durch die Anknüpfung an die interpretierten Daten neue Prozesse gestartet, sondern auch wie diese verstanden und interpretiert werden, um unter anderem KI-Technologie weiterzuentwickeln. Idealerweise können die lehrenden Akteur:innen die Ergebnisse auch durch ein »Validierungsdatenset« prüfen, und

so festzustellen, ob sich das Modell »wirklich auch so halbwegs generalisieren« (I10, Pos. 20) lässt und zufriedenstellende Ergebnisse ermöglicht.

Wieder gibt der Mensch Daten kontrolliert vor, mit deren Hilfe das durch die Maschine entwickelte Modell überprüft werden soll.

Abbildung 6: Netdom Supervised Learning t_2



Hier schließt ein weiterer Zyklus des Lernens an, der über die Switchings in die einzelnen Netdoms im Netdom Supervised Learning (SL 0) zu einem zweiten Zeitpunkt (t_2) führt und so die wechselseitigen Kontrollbemühungen der heterogenen Entitäten umfasst.

Die Lernprozesse werden durch den Menschen im Netdom SL1 zu t_2 über die Ambiguity M_2-H_2-SL erneut interpretiert. Danach schließt sich im besten Fall eine weitere idealtypische »Schleife« der soziotechnischen Konstruktion des Supervised Learnings an. Mithilfe eines unabhängigen »Testsets« im vorgezeichneten Problemhorizont der Story wird eben jenes entwickelte Modell evaluiert und das Ergebnis nach seiner Qualität überprüft (vgl. I10, Pos. 20). Dabei hat der gesamte Lernprozess des Supervised Learnings Einfluss auf die darin fortgeschriebene Story, welche zu jedem Zeitpunkt überprüft und wenn nötig auch angepasst wird.

Die Frage der Kontrolle ist damit eine Frage der Relation und wie bereits angesprochen, keine der Objektivität.

»Wir haben jetzt diese Modelle und wir haben keine Ahnung, wie die funktionieren. Jetzt können wir aber damit, mit den Modellen, auch wieder arbeiten und uns zum Beispiel vorstellen, dass wir Fragen an diesem Modell stel-

len, so wie wir Fragen an eine Datenbank stellen. Zum Beispiel so eine Frage wie: Was passiert, wenn ich die Eingabe hier ein bisschen ändere?« (I02, Pos. 47)

Auf die mangelnde direkte Kontrolle von Seiten des Menschen auf die Datenverarbeitung der Learner, der Ambiguity M_3 - H_3 -SL, werden strategische Versuche gestartet, die im Mindesten Einfluss auf die Ausgaben generieren sollen.

Wenn das Modell nach den Prüf- und Testungen nicht funktionieren sollte, werden Problemanalysen im Netdom SL1 angestellt. Da komme es auf die Expertise der Lehrenden an.

»Wenn man jetzt irgendwie einen Fehler beim Training oder sowas hat, [muss man] eben einfach so ein Gespür zu haben: an welchen Stellschrauben muss ich jetzt drehen, damit ich den und den Fehler beheben kann. Das ist, glaube ich, einfach Erfahrung oder eben auch viel Literaturrecherche. [...] Was haben andere gemacht in der Richtung, was sind vielleicht irgendwie die typischen Probleme, die eigentlich immer bei der speziellen Architektur auftreten.« (I10, Pos. 55)

Bei sehr großen Modellen besteht die Gefahr, dass man »zu viele Stellschrauben hat« (I10, Pos. 22). Das könne zur Folge haben, dass die zur Verfügung gestellten Daten vom Modell schnell auswendig gelernt würden und so nur für das Trainingsdatenset gut funktionieren, »aber nicht auf dem Testset« (ebd.). Eine Möglichkeit, dem zu begegnen, wäre, das Modell anzupassen, es beispielsweise zu verkleinern, damit es »eben gar nicht die Möglichkeit hat, das Trainingsdatenset auswendig zu lernen, sondern gezwungen ist, so ein bisschen allgemeinere Repräsentation der Daten zu finden« (I10, Pos. 22).

Auch die Veränderung der Trainingsdaten kann eine Veränderung des Modells erzeugen:

»Was auch häufig gemacht wird, ist, dass man sogenannte Datenaugmentierung benutzt. Das heißt, man nutzt die Trainingsdaten, die man schon hat und versucht die irgendwie zu variieren, dass man beispielsweise einfach die Bilder rotiert oder irgendwie noch ein bisschen Rauschen hinzufügt oder die Helligkeit ändert und so weiter, dass man quasi verschiedene Variationen, die so in der Realität auch auftauchen können, versucht, mit reinzunehmen.« (I10, Pos. 22)

Der Prozess der Fehlersuche sei jedoch sehr komplex und gleiche oft eher einer Suche über »Trial-and-Error« (ebd., Pos. 24).

»Und oft ist es auch so, dass man sich nicht so hundertprozentig erklären kann, woran es eigentlich liegt. Da diese neuronalen Netze eben so komplex sind mittlerweile, dass man eigentlich fast keine Chance hat, das wirklich im Detail zu verstehen, was da genau passiert.« (Ebd.)

Spätestens an dieser Stelle wird die soziotechnische Konstruktion des Lernprozesses der Form des Supervised Learnings besonders deutlich herausgestellt.

Hinzu kommt, dass Entwickler:innen die Daten mithilfe umfassender Dokumentationen zur Unterstützung der Interpretation der Datenverarbeitungsprozesse durch Dritte einsetzen. Sie sprechen hier von einer Art »Veredelung« (vgl. Io6, Pos. 42) der Daten, die unter anderem verschiedene Datenquellen und Beschreibung der Datenform ausweisen (vgl. ebd., Pos. 70). Damit konstatieren sie erneute Kontrolle im Sinne der Ambage. Entwickelnde Akteur:innen sprechen bei dieser Lernform unter anderem davon, die »Stellhebel in der Hand« (vgl. ebd., Pos. 54) zu halten, sowohl bei der Interpretation der Daten und der Ableitung der Entscheidungen als auch dabei, die daraus generierten Informationen für Dritte, bspw. für eine spezifische Kundschaft, transparent und verständlich zu machen.

Die Ambiguity, welche im Umgang mit den Datenausgaben vorliegt, wird durch die entwickelnden Akteur:innen über die Einflussnahme in Abhängigkeit von den Stakeholdern oder anderen Zielsetzungen vereindeutigt und so wiederum kontrolliert.

Wir können konstatieren, dass der Einfluss der Akteur:innen im Netdom Supervised Learning das Kontrollprojekt Ambage aus Sicht der Akteur:innen prägend und durch die spezifische Datengrundlage mit eindeutigen Mustern dominant wahrgenommen wird. Die Kontrollbemühungen des Menschen an den vielfältigen Stellen im Lernprozess offenbaren damit eine gewisse Eingriffstiefe in die technologischen Abläufe beim Supervised Learning. Dennoch wird offensichtlich, dass der Lernprozess über die verschiedenen Netdom-Switchings als eine Form des Interfacings die soziotechnische Konstruktion vorprägt und diese wechselseitig miteinander verzahnt ist.

Das Netdom »Unsupervised Learning«

Bei der Lernform des Unsupervised Learnings werden die Learner strenggenommen nicht in der oben beschriebenen Weise durch zuvor definierte Ein- und Ausgaben (vgl. Lenzen 2018) angeleitet, wie und auf welche Weise Muster in den Daten erkannt werden sollen. Vielmehr werden Daten durchsucht und der Versuch unternommen, darin »selbstständig« Muster zu identifizieren (vgl. ebd.).

Die Entscheidung für diese oder eine andere Lernform fällt vordergründig auf der Basis der vorliegenden Daten. Liegen große Datenmengen vor, die »nicht irgendwie schon mal von einem Menschen angefasst« (Io2, Pos. 49) wurden, kann der Versuch unternommen werden, mit einem Unsupervised Learning Ansatz aus den Daten »trotzdem etwas« (ebd.) herauszuholen. So können zum Beispiel »sogenannte Auto-Encoder« (vgl. I10, Pos. 51) verwendet werden, um die Netzwerke auf die Lernprozesse inhaltlich einzustellen. Dabei werden zum Beispiel Bilder eingegeben, die es ermöglichen sollen, später »genau das gleiche Bild« (ebd.) wiederzuerkennen. Dieses Training kann dann eine Basis für einen anschließenden Supervised Learning-Pfad bilden.

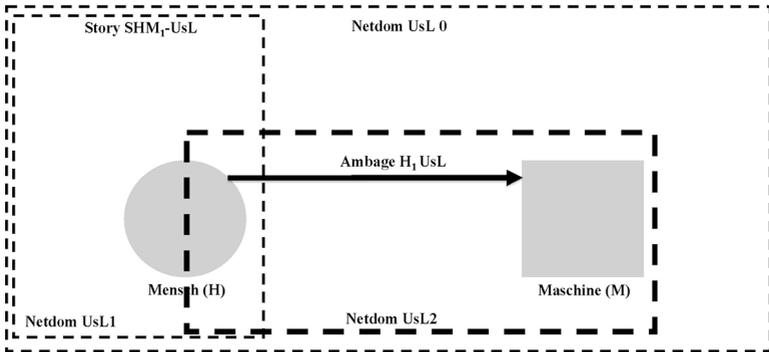
»Man forciert im Prinzip das Netzwerk, dass es eine komprimierte Repräsentation quasi von diesem Bild lernt, und durch diesen Prozess kann man im Prinzip so eine Art Vortraining von beispielsweise einem neuronalen Netz hinbekommen, das man dann wiederum verwenden könnte.« (I10, Pos. 51)

Auch beim Unsupervised Learning steht zu Beginn des Lernprozesses die Entwicklung einer Fragestellung, die als Rahmen für die Story SHM₁-UsL von Seiten der Akteur:innen im Netdom UsL₁ gestellt wird, diese ist aber viel weniger orientiert als beim Supervised Learning. Doch sinkt damit auch der Einfluss des Menschen auf den Lernprozess im Unsupervised Learning?

Denn im Vergleich zum Supervised Learning, bei dem eine vorherige Annotation der verwendeten Daten notwendig ist, wird beim Unsupervised Learning hingegen der Versuch unternommen, direkt etwas aus »unbearbeiteten« Daten zu lernen, die jedoch, niemals in roher Form vorliegen (s.o. vgl. Häußling 2020).

Wie auch schon im Kapitel Supervised Learning dargestellt, findet auch bei dieser Lernform ein Netdom-Switching in das Netdom UsL₂ statt, in dem eine deutliche Ambage (H₁UsL) der Akteur:innen auf die Learner einwirkt.

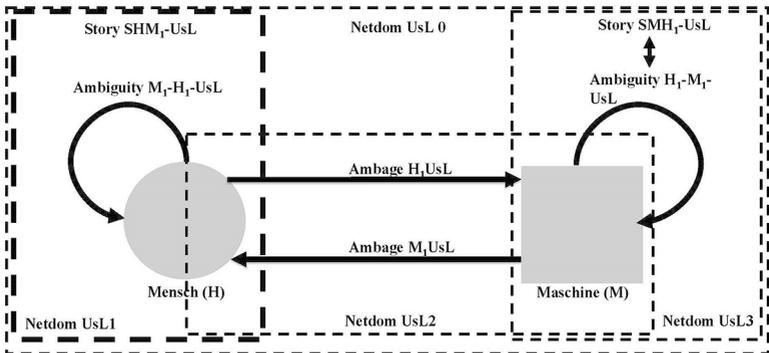
Abbildung 7: Ambage H_1UsL im Netdom UsL_2



Netdom UsL_1 (0/1/2/...): Netdom Unsupervised Learning und darin weitere verschachtelte sinnhafte Netzwerk-Domänen
 SHM- UsL_1 : Einfluss der Mensch-Maschinen Relation auf die Storyentwicklung in der UsL_1 Netzwerk-Domäne
 Ambage H_1UsL : Kontrollversuch des Menschen auf die Maschine im Unsupervised Learning

Eine große, noch nicht klar bearbeitete Datenmenge wird von Seiten der Akteur:innen an die »Maschine« übermittelt, ohne zuvor Muster durch annotierte Daten vorzugeben. Die Learner verarbeiten diese Datensätze und erkennen möglicherweise Muster in den Daten, was als Kontrollprojekt Ambiguity H_1-M_1-UsL im Netdom UsL_3 in unserem Lernmodell beschrieben wird.

Abbildung 8: Ambiguity M_1-H_1-UsL



Diese Ver- und Bearbeitung der Daten und insbesondere die Mustererkennung wird wiederum über die Ambage M_1UsL im Netdom UsL_2 zurück an die Akteur:innen gespielt, welche vom Menschen im Netdom UsL_1 gedeutet wer-

den muss. Die Herausforderung besteht auf Seiten der Menschen gerade darin, dass die durch die Maschine in den Datensätzen erkannten Muster von den anlernenden Akteur:innen selbst nicht mehr erkannt werden können.

»Die Maschine sagt, das Muster ist da, dann ist das natürlich problematisch. Wenn ich dann aber ausprobieren kann, ob da was ist, weil ich dann ein Experiment machen kann, dann gehe ich diesem Problem aus dem Weg. Wenn ich mich nur auf das verlasse, was die Maschine erkennt, ist das schwieriger.« (102, Pos. 17)

Die Ambiguity H_1 - M_1 -UsL im dritten Netdom des Unsupervised Learnings der durch die Learner erkannten Muster, ist folglich eine besondere Herausforderung bei der erneuten Interpretation der Daten durch die anlernenden Akteur:innen, die durch den Prozess der Visualisierung noch einmal verschärft wird, wenn es zum Beispiel um die Eindeutigkeit der Muster geht.

»Man muss immer ein so ein bisschen tricksen, dass man im Prinzip die Sachen dann auch rausfindet, die man haben will. Also beispielsweise, wenn man jetzt eine Clusterung durchführt, weiß man ja im Prinzip gar nicht, welches Cluster ist was, was bedeuten die Cluster und so weiter. Man hat halt irgendeine Gruppierung im Raum von seinen Daten, muss die aber dann später nochmal irgendwie weiterverarbeiten und eben, ja, rausfinden, was was bedeutet.« (110, Pos. 53)

Die über die einzelnen Netdoms im Unsupervised Learning hinweg vorhandene wechselseitige Mehrdeutigkeit und die gegenseitig einwirkende Ambiguität zeigt einen soziotechnischen Lernprozess über verschiedene Netdoms hinweg. Die Kunst des Lernens in der Lernform des Unsupervised Learnings ist damit stark von einer Annäherung durch die Vernetzung von Interpretation gekennzeichnet.

Damit rücken die Auswahl, Interpretation und Verarbeitung der Daten in den Blick, wenn es um die Bedeutung von Intransparenzen bei der Weiterverarbeitung der spezifischen Daten in den verschiedenen Schleifen der wechselseitigen Kontrollprojekte zwischen Mensch und Maschine geht. Es ist nicht nachvollziehbar, was in der Verarbeitung, zum Beispiel der Clusterung von Daten auf der Seite der Maschine passiert, was ein neues Problem der Intransparenz darstellt (vgl. Schmitt/Heckwolf in diesem Band).

»Also vielleicht in ganz kritischen Bereichen kann man die Systeme nur als Zwischenschritt verwenden, das sind auch Möglichkeiten. Dass man am Ende, das was man dann wirklich in der sicherheitskritischen Anwendung einsetzt vielleicht gar nicht mehr so eine Black-Box ist, sondern die Black-Box verwendet, um das andere zu bauen.« (IO2, Pos. 47)

Die Aufnahme der Ambage M_1UsL über das Netdomswitching von $N-UsL_2$ zum Netdom UsL_1 kann im Prozess der Interpretation der Datenausgaben einen weiteren Kontrollversuch (Ambiguity H_1-M_1-UsL) vor der Ambage H_2UsL bedeuten. Da die vorständige Kontrolle der Datenverarbeitung der KI durch den Menschen nicht möglich ist, kann der Versuch unternommen werden, interpretationsfähige Daten darüber zu erhalten, was die KI überhaupt identifiziert hat und worauf sie scharf stellt.

Geht man davon aus, dass künstlich intelligente Modelle über ein Unsupervised Learning Muster in den Strukturen erkennen, besteht die Möglichkeit, systematisch Daten zu manipulieren, und zwar so, dass diese Manipulationen allein vom Modell und nicht für den Mensch wahrgenommen werden können, wie zum Beispiel feinste Farbveränderungen in Bildern. Die Frage ist an dieser Stelle wiederum, welche Auswirkungen dies für die Ambage durch die Maschine (M_nUsL) und den Deutungsversuch des Menschen, der Ambiguity M_n-H_n-UsL hat.

Veränderungen der Daten bewirken klare Veränderungen in den Mustererkennungsprozessen der Learner, die wiederum Datenausgaben weitergeben, welche vom Menschen interpretiert werden müssen. Das Rechnen mit der Intransparenz in der Verarbeitung der Daten durch die künstlich intelligenten Modelle, kommt damit in gewisser Weise einem neuen Kontrollprojekt, dem Blackboxing, gleich (vgl. Schmitt/Heckwolf in diesem Band), das im Kontext der Ambiguity sowohl im Netdom 1 des Unsupervised Learnings (UsL_1) als auch im Netdom UsL_3 wirkmächtig werden kann. Dies ist bei dieser Lernform besonders stark, weil hier keine Datenannotation stattfindet und keine Muster vordefiniert werden, die einen deutlicheren Einfluss durch den Menschen in der Bearbeitung und Orientierung der Daten sowie der Learner bedeutet.

Doch die Reinform des maschinellen Lernens findet in der Forschung und Entwicklung wenig Anwendung. Oft liegen Kopplungen des Unsupervised mit dem Supervised Learning vor, sodass die Kunst des Lernens nicht zuletzt auch eine Kunst des Lehrenden ist, die spezifische Ver- und Entkopplungen der Lernformen miteinander je Kontext anzuwenden.

4. Die Kunst des Lehrens

Die Kunst des Lernens präsentiert sich idealtypisch differenziert, in Abgrenzung der beiden Lernformen zueinander, als einen wechselseitigen Prägungszusammenhang von Kontrollversuchen heterogener Entitäten. Häufig kommt es jedoch zu einer Kopplung der beiden Lernformen. Die beiden Netdoms SLO und UsLO treten dabei in neue Kopplungen, die eine komplexe Verschachtelung der Switchingprozesse in und über die Netdoms hinweg zufolge haben. Um diese komplexen wechselseitigen Ver- und Entkopplungsprozesse der Lernprozesse nachvollziehen zu können, müssen die Entscheidungsprozesse auch im Kontext der Entwicklung der jeweiligen Stories, und damit verbunden für die Beziehung der beiden Lernformen, betrachtet werden. Damit rückt nach der schon beschriebenen Kunst des Lernens die Kunst des Lehrens in den Fokus.

Je nach Forschungsfrage werden im idealen Fall unterschiedliche Lernformen miteinander verbunden, sodass zu Beginn die Storyentwicklung steht, also welche Funktion die künstlich intelligenten Modelle erfüllen sollen. Erst daran orientiert, werden die einzelnen Prozesse und Lernformen je nach Bedarf gekoppelt.

»Das heißt, das wäre im Prinzip so eine Mischung aus beidem. Das man einmal die Unsupervised Methode benutzt, um die Trainingsdaten zu generieren, also um diesen Simulationsansatz quasi durchzuführen und dann diese Trainingsdaten, die wir da generiert haben, diese künstlich erzeugten, dann im Prinzip für einen Supervised Ansatz benutzt und dabei im Prinzip genau das umgedrehte lernt. Also, dass man vom Rohbild auf die Maske kommt, das dann aber wiederum auf die realen Daten später anwendet, um dann aus den realen Daten die Masken rauszubekommen. [...]. Also im Prinzip so ein bisschen vertrackt, also das ist so eine Mischung aus verschiedenen Ansätzen, die aber [...] das eigentliche Ziel eben haben, aus dem Rohbild Daten rauszukriegen.« (110, Pos. 53)

Je nach Fachgebiet wird in der Forschung der Versuch unternommen, das Verhältnis der beiden Lernformen zu erschließen und im Einsatz zu prüfen. In Abhängigkeit zur Fragestellung werden die Lernformen nicht nur ergänzend eingesetzt, sondern auch in grundsätzlicher Beziehung zueinander erforscht. Im Bereich der Spracherkennung wird das »reine« Unsupervised Learning als »kryptographisches Problem« (107, Pos. 74) betrachtet. Dieses Problem wird

theoretisch im Bereich der informatischen Spracherkennung untersucht, aber zumeist durch die Kopplung von Unsupervised und Supervised Learning gelöst (ebd.). Die in diesem Fachgebiet angesprochene Verbindung der beiden Lernformen als mögliche Lösung der Herausforderungen deutet zum einen die technischen Schwierigkeiten des reinen Unsupervised Learnings an. Zum anderen werden auch die geführten wissenschaftlichen Diskurse davon beeinflusst und Theoriebildung dadurch angeregt. Interessant ist gerade die Ergänzung von anderen Lernformen im Netdom. Reigeluth und Castelle (2021) haben bereits darauf hingewiesen, dass diese Lernform den Anschein erweckt, ohne »Lehrende« auszukommen. Die Autoren betonen jedoch die kulturelle Einbettung des Lernens und damit auch die Einbettung von Lernern in sozio-kulturelle Arrangements und konstatieren, dass Lernen ohne Anleitung im strengen Sinn unmöglich ist.

Im Prozess des Anlernens gehe es aber gar nicht um das Verstehen, was der Learner genau mit den Datensätzen macht, wie Muster generiert werden, sondern es gehe vielmehr um die Arbeit mit den entstandenen Modellen (vgl. I02, Pos. 47). Damit werden die strategischen Entscheidungen gestärkt, also welche Sets im Supervised und welche Modelländerungen beim Unsupervised Learning vorgenommen werden müssen.

Wir stehen damit vor der Relevanz der Frage nach dem Vertrauen in einerseits die »richtige« Verarbeitung von Daten in den Modellen und der Bedeutung der Intuition und Erfahrung in der Arbeit an den Modellen:

»Also ich glaube im Moment sind wir im Wesentlichen noch in einem Stadium, wo das erfahrungsgetrieben ist. Also, Leute, die da arbeiten, wissen ungefähr, was sie tun können, und manche wissen es besser und sind am Ende damit erfolgreicher, und das ist zum Teil mysteriös. Und das ist natürlich ein Zustand, der auch nicht so doll ist, ist aber wiederum etwas, worum sich viele Leute Gedanken machen, wie man das vernünftiger machen kann, dass diese Technik letztlich einfacher einsetzbar ist. Da ist man noch ein Stück weit von entfernt.« (I02, Pos. 51)

Andererseits offenbart es auch die Bedeutung der Arbeit mit den Intransparenzen, einem Blackboxing nach Schmitt und Heckwolf (in diesem Band) in der wechselseitigen Aushandlung von Identität und Kontrolle zwischen Mensch und Maschine sowie die Herausforderung der Bedeutung eines grundsätzlichen bestehenden Objektivitätsanspruches von quantitativen Daten, in Anlehnung an Porter (2020) Rechnung zu tragen. Die historische

Quantifizierung in den Wissenschaften als Stärkung einer idealen mechanischen Objektivierung, des nach exakten Methoden entwickelten Wissens, ist verbunden mit dem Wunsch Expert:innenurteile zu ersetzen. Künstlich intelligente Modelle scheinen hier in klarer Linie einer idealen Objektivität zu stehen. Doch wie wir zeigen konnten, ist ganz im Sinne Porters, das »Problem des Vertrauens« (Porter 2020: 214) nicht einfach aufzulösen.

Bereits Helmholtz hatte darauf verwiesen, dass Wissenschaften, die sich der künstlerischen Induktion bedienen, gut beraten sind, zunächst »die Glaubwürdigkeit der Berichterstatter, die ihnen die Thatsachen überliefern« (Helmholtz 1896: 172), also die Entscheidungsgrundlagen liefern, zu prüfen. In gewisser Weise ist die Arbeit an der Wirklichkeit, der Arbeit an und mit künstlich intelligenten Modellen kunstfertig konstruiert (vgl. Porter 2020: 5).

Beispielhaft kann man hier ChatGPT und dessen Entwicklung seit seiner Veröffentlichung im November 2022 heranziehen, bei der sich sowohl Intransparenzen in den letzten Updates der Modelle (ChatGPT 3.5 zu ChatGPT4) zu offenbaren scheinen als auch qualitative Einbußen bemerkt wurden (vgl. Chen et al. 2023: 14f.), welche möglicherweise auch auf den (intentionalen) Eingriff der Entwickler:innen zurückgeführt werden könnten. Diese Veränderungen führten bereits zu einem stärkeren Misstrauen in die Entwicklungsarbeit der Technologie und führten dazu, dass Fachleute zu Skepsis gegenüber den Ergebnissen von ChatGPT aufrufen (vgl. DLF Nova, 24.07.2023). Die Glaub- und Vertrauenswürdigkeit der Entwickler:innen und folglich auch der Ergebnisse des ChatBots selbst scheint dadurch zu leiden und lenkt zudem den Fokus auf die Rolle der Entwickler:innen KI-basierter Technologie.

Die Kunstfertigkeit der Lehre liegt damit nicht nur in der Reflexion der Kunstfertigkeit selbst, sondern im Design eines soziotechnisch konstruierten Lernprozesses, der Lernformen mit- und zueinander, der Daten und Datendokumentation nur in der Kopplung von Mensch und Maschine kennzeichnet. Dies zeigt sich nicht nur darin, dass das Berechnen des nicht Berechenbaren ohne das nicht Berechenbare wirklich ausrechnen zu wollen, von einer soziotechnisch konstruierten künstlichen Intelligenz ausgeht, die sich von der menschlichen unterscheidet, aber nicht ohne diese auskommt. Sondern sie führt gleichzeitig die Bedeutung vor Augen, wie soziotechnisch konstruiert die Eingaben, die Daten, in den Lernern selbst sind und, dass diese einen Effekt haben, sowie der Umgang mit den visualisierten Daten als Ausgaben der Maschine weitere soziotechnische (Ent)Kopplungen prägt. Dabei geht es nicht nur um die Relevanz der Auswahl und Annotierung von Daten, sondern auch um die Verkopplung von Lernformen je nach Kontext und Einbettung.

Die Kunst des Lernens im Machine Learning und auch Deep Learning setzt dabei nicht nur an der soziotechnischen Konstruktion der Daten an, sondern führt mithilfe der Terminologie der White'schen Theorie die artifiziiellen Bereiche des An-Lernens vor Augen. Damit gehört zur Kunst des künstlich intelligenten Lernens auch die Kunst des Lehrens von künstlich intelligenten Lernern.

5. Literatur

- Burrell, Jenna (2016): »How the machine ›thinks‹: Understanding opacity in machine learning algorithms«, in: *Big Data & Society* 3, S. 1–12.
- Cardoso Llach, Daniel (2018): »Daten als Schnittstelle. Die Poetik des maschinellen Lernens im Design«, in: Christoph Engemann/Andreas Sudmann (Hg.), *Machine Learning. Medien, Infrastrukturen und Technologien der Künstlichen Intelligenz*, Bielefeld: transcript, S. 195–218.
- Chen Lingjiao/Zaharia, Matei/Zou, James (2023): »How Is ChatGPT's Behavior Changing over Time?«, in: arXiv preprint arXiv:2307.09009.pdf (arxiv.org) vom 17.08.2023.
- DLF Nova, 24.07.2023: »Updates: ChatGPT wird wohl immer dümmer«, in: DLF Nova (deutschlandfunknova.de).
- Domingos, Pedro (2015): *The Master Algorithm. How the Quest for the Ultimate Learning Machine Will Remake Our World*, New York: Basic Book.
- Häußling, Roger (2012): »Design als soziotechnische Relation. Neue Herausforderungen der Gestaltung inter- und transaktiver Technik am Fallbeispiel humanoider Robotik«, in: Stephan Moebius/Sophia Prinz (Hg.), *Das Design der Gesellschaft. Zur Kultursoziologie des Designs*, Bielefeld: transcript, S. 273–298.
- Häußling, Roger (2016): »Zur Rolle von Entwürfen, Zeichnungen und Modellen im Konstruktionsprozess von Ingenieuren. Eine theoretische Skizze«, in: Thomas H. Schmitz/Roger Häußling/Claudia Mareis/Hannah Groninger (Hg.), *Manifestationen im Entwurf. Design – Architektur – Ingenieurwesen*, Bielefeld: transcript, S. 27–64.
- Häußling, Roger (2020): »Daten als Schnittstellen zwischen algorithmischen und sozialen Prozessen. Konzeptuelle Überlegungen zu einer Relationalen Techniksoziologie der Datafizierung in der digitalen Sphäre«, in: *Soziale Welt (Sonderband 23)*, S. 134–150.

- Helmholtz, Hermann (1896): »Ueber das Verhältnis der Naturwissenschaften zur Gesamtheit der Wissenschaft. Akademische Festrede gehalten zu Heidelberg beim Antritt des Prorektorats 1862«, in: Vorträge und Reden Bd. 1, Braunschweig: Vieweg, S. 157–186.
- Kaminski, Andreas/Glass, Colin W. (2019): »Das Lernen der Maschinen«, in: Kevin Liggeri/Oliver Müller (Hg.), Mensch-Maschine-Interaktion. Handbuch zu Geschichte – Kultur – Ethik, Weimar: J. B. Metzler, S. 128–133.
- Karafilidis, Athanasios (2018) »Die Komplexität von Interfaces. Touchscreens, nationale Identitäten und eine Analytik der Grenzziehung«, in: Berliner Debatte Initial 29, 130–146.
- Koren, István/Klamma, Ralf (2018): »Enabling visual community learning analytics with Internet of Things devices«, in: Computers in Human Behavior 89, S. 385–394.
- La Mettrie (2001): *L'Homme machine. Der Mensch eine Maschine*, Stuttgart: Reclam.
- Lenzen, Manuela (2018): *Künstliche Intelligenz. Was sie kann & was uns erwartet*, München: C. H. Beck.
- Porter, Theodore M. (2020): *Trust in Numbers. The Pursuit of Objectivity in Science and Public Life*, Princeton: Princeton University Press.
- Reigeluth, Tyler/Castelle, Michael (2021): »What Kind of Learning Is Machine Learning?«, in: Jonathan Roberge/Michael Castelle (Hg.), *The Cultural Life of Machine Learning. An Incursion into Critical AI Studies*, Cham: Palgrave Macmillan, S. 79–116.
- Riskin, Jessica (Hg.) (2007): *Genesis Redux. Essays in the History and Philosophy of Artificial Life*, Chicago/London: University of Chicago Press.
- Russel, Stuart J./Norvig, Peter (2016): *Artificial Intelligence. A Modern Approach*, Boston et al.: Pearson.
- Schmitt, Marco/Heckwolf, Christoph (2023): »KI zwischen Blackbox und Transparenz«, in: diesem Band.
- Schütz, Alfred, 1972: »Der gut informierte Bürger. Ein Versuch über die soziale Verteilung des Wissens«, in: Alfred Schütz (Hg.), *Gesammelte Aufsätze*, The Hague: Nijhoff, S. 85–101.
- Shalev-Shwartz, Shai/Ben-David, Shai (2014): *Understanding Machine Learning. From Theory to Algorithms*, Cambridge: Cambridge University Press.
- Simon, Herbert (1984): »Why Should Machines Learn?«, in: Ryszard S. Michalski/Jamie G. Carbonell/Tom M. Mitchell (Hg.), *Machine Learning. An Artificial Intelligence Approach*, Berlin: Springer, S. 25–37.

- Simon, Herbert (1994): Die Wissenschaften vom Künstlichen. Zweite Auflage, Wien: Springer.
- Sudmann, Andreas (2018): »Zur Einführung. Medien, Infrastrukturen und Technologien des maschinellen Lernens«, in: Christoph Engemann/ Andreas Sudmann (Hg.), *Machine Learning. Medien, Infrastrukturen und Technologien der künstlichen Intelligenz*, Bielefeld: transcript, S. 9–23.
- Taffel, Sy (2019): »Automating Creativity – Artificial Intelligence and Distributed Cognition«, in: *spheres 5*, S. 1–9.
- Turing, Alan (2004): »Intelligent Machinery (1948)«, in: B. Jack Copeland (Hg.), *The Essential Turing. Seminal Writings in Computing, Logic, Philosophy, Artificial Intelligence, and Artificial Life plus The Secrets of Enigma*, Oxford: Oxford University Press, S. 410–432.
- White, Harrison C. (1992): *Identity and control: A structural theory of social action*, Princeton/NJ: Princeton University Press.
- White, Harrison C. (2008): *Identity and control: how social formations emerge*, Princeton/NJ: Princeton University Press.
- Ziegler, Siegfried (2007): *Lernen bei Gregory Bateson. Die Veränderung sozialer Systeme durch organisationales Lernen*, Saarbrücken: VDM.

Deep Learning Techniken als Boundary Objects zwischen Entwicklungs- und Anwendungsfeld

Philip Roth, Matthias Dorgeist, Astrid Schulz

1. Einleitung

Personen aus ganz unterschiedlichen Abteilungen, Disziplinen oder Hierarchieebenen beteiligen sich oftmals an Projekten zur Entwicklung neuer Technologien. Eine Herausforderung in solchen Projekten ist es daher, die heterogenen Wissensbestände aus unterschiedlichen Feldern produktiv zu integrieren (Hahn 2013; Heidenreich/Mattes 2017). Praxistheoretische Perspektiven legen dabei nahe, dass sich die einzelnen Wissensbestände nicht auf einfache Weise integrieren lassen, da es schwierig ist, diese zu reflektieren und zu kommunizieren. Das spezifische Wissen der Akteur:innen (zum Beispiel von Ingenieur:innen, Informatiker:innen oder Ökonom:innen) lagert sich in ihren alltäglichen Praktiken und spezifischen Praxiszusammenhängen ab, ist damit Teil ihres Habitus und des jeweiligen Feldzusammenhangs und ist zunächst mehr oder weniger unbewusst. Um die Verständigung und Koordination zwischen den heterogenen Partner:innen in interdisziplinären Projekten zu schaffen, ist es dementsprechend erforderlich, dass die Partner:innen aus ihrem habituell gewohnten Praxiszusammenhang heraustreten und einen Kontextwechsel vollziehen, um einen gemeinsamen Rahmen der Verständigung zu finden (Ibert/Müller/Stein 2014; Heidenreich/Mattes 2017; Carlile 2002). In der Forschung zur Verständigung heterogener Partner:innen in Entwicklungsprojekten wurde gezeigt, dass Entwicklungsgegenstände – also die jeweils zu entwickelnden Geräte, Programme etc. – selbst als *Boundary Objects* Verständigungs- und Koordinationsprozesse strukturieren können (Leonardi/Bailey/Pierce 2019; Barrett/Oborn 2010; Hahn 2013; Carlile 2002). In diesem Beitrag fragen wir danach, wie Deep Learning Techniken (DLT) als

Entwicklungsgegenstand das Verhältnis von DLT-Entwickelnden und DLT-Anwendenden strukturieren.

Bei der Beantwortung dieser Frage stützen wir uns auf feld- und praxistheoretische Konzepte und nicht, wie die übrigen Beiträge des Sammelbandes, auf das Theorieangebot von Harrison White. Diesen Zugang haben wir gewählt, da uns die Exploration des empirischen Materials zu der oben genannten Forschungsfrage und einer Fokussierung von Akteursperspektiven und -praktiken geführt hat. Ausgehend davon haben sich für uns vor allem praxistheoretische Vorarbeiten als unmittelbar hilfreich erwiesen. Dementsprechend war es uns mit geringerem Übersetzungsaufwand möglich, den empirischen Gegenstand praxistheoretisch zu fassen und an den Stand der Forschung anzuschließen. Im abschließenden Abschnitt des Beitrags werden wir vor dem Hintergrund unsere Befunde diskutieren, welche weiterführenden Erkenntnisse eine Integration von Whites Theorie in diesem Forschungsfeld verspricht.

In Entwicklungsprojekten kommt es regelmäßig zu Differenzen zwischen den Perspektiven der technischen Entwicklung und den Perspektiven der Anwendung (Ibert/Müller/Stein 2014; Heidenreich/Mattes 2017; Degele 1996). Als Phasen eines Entwicklungsprozesses folgen sie unterschiedlichen Logiken, was im Prozess zu Schwierigkeiten führen kann, indem beispielsweise in der technischen Entwicklung Anwendungsaspekte vernachlässigt werden, die zu ›zum Scheitern verurteilten‹ Prototypen werden, es zu keiner ausreichenden Zusammenarbeit kommt oder Anwendende nicht realisierbare technische Anforderungen formulieren (Degele 1996). Dem begegnend werden seit den 1980er Jahren mögliche Anwendende gezielt in Produktentwicklungsprojekte eingebunden, um ihre Kenntnisse frühzeitig fruchtbar zu machen (Hippel 1986; Bogers, et al. 2017).¹ Mit dieser Verschiebung werden Anwendende selbst zu Beteiligten im Entwicklungsprozess, indem sie ihr Handlungswissen in den Entwicklungsprozess einbringen. Doch mit der organisationalen Einbindung der Anwendenden in Entwicklungsprojekte ist noch nicht zwingend eine Überbrückung der Differenzen gelungen. So spiegelt sich das unterschiedliche Interesse weiterhin in einer unterschiedlichen Problematisierung der Bedeutung eines spezifischen Entwicklungsobjektes und der Befolgung verschiedener Handlungslogiken wider. Deshalb werden unter Rückgriff auf

1 Siehe auch Degele (1996: 65) zur Entwicklung der Partizipativen Systementwicklung, die beginnt, das Gelingen von Entwicklungsprojekten im Softwarebereich in nicht-technischen Faktoren zwischen Entwickler:innen und Anwender:innen zu verorten.

Praxistheorien Anwendende und technische Entwickler:innen in der diesbezüglichen Forschung als Akteur:innen konzipiert, die in unterschiedliche *Soziale Felder* eingebettet sind (Levina/Vaast 2005, 2008; Bjerregaard 2010). Soziale Felder entstehen, wenn Akteur:innen praktisches Wissen miteinander teilen und diesem folgend nach den gleichen Regeln kontextspezifisch kompetent handeln (Bourdieu/Wacquant 2006; Levina/Vaast 2005). Grenzen zwischen Feldern bestehen dementsprechend dort, wo Diskontinuitäten und Brüche zwischen diesen praktischen Eigenlogiken vorzufinden sind (Levina/Vaast 2005, 2008). Die Brüche müssen dabei nicht offensichtlich im Alltag zu Tage kommen, sondern können sich zum Beispiel auch in Formen des aneinander Vorbeiredens andeuten. Entscheidend für eine gelungene Integration beider Felder ist daher, das eigene praktische Handlungswissen für Akteur:innen des anderen Feldes verständlich machen zu können und so zu einer Überwindung der Feldgrenzen zu gelangen (Leonardi/Bailey/Pierce 2019; Heidenreich/Mattes 2017; Levina/Vaast 2005).

In der Theorie des Boundary Objects spielen bei der Koordination und Verständigung zwischen den heterogenen Partner:innen innerhalb solcher Projekte neben menschlichen Akteur:innen auch die technischen Entwicklungsobjekte² eine wichtige Rolle. Diese haben sich als wichtiger Faktor bei der Vermittlung zwischen den Feldern erwiesen (Barrett/Oborn 2010; Levina/Vaast 2005; Leonardi/Bailey/Pierce 2019). Boundary Objects sind abstrakte oder konkrete Gegenstände, auf die sich in verschiedenen Sozialen Feldern

2 Der Entwicklungsgegenstand erfährt in verschiedenen Phasen des Entwicklungsprozesses unterschiedliche Repräsentationen (z.B. Skizzen, Modelle, Prototypen). Beispielsweise ermöglichen Prototypen, die zwischen technisch Entwickelnden und Anwendenden zur Diskussion gestellt werden können, eine Verständigung zwischen den beteiligten Akteur:innen. Außerdem sind die Repräsentationen der Entwicklungsgegenstände in verschiedenen Entwicklungsphasen mit spezifischen, darauf bezogenen Praktiken assoziiert, die sich sowohl aus den technischen Anforderungen des Entwicklungsgegenstandes in der jeweiligen Phase als auch aus den darauf bezogenen sozialen Konventionen ergeben. Als Boundary Objects schließen Entwicklungsgegenstände dementsprechend die spezifischen Entwicklungsarchitekturen ein, die die auf das jeweilige Endprodukt hin ausgerichteten Zwischenergebnisse und Praktiken beschreiben. Was konkret entwickelt wird, kann sich stark voneinander unterscheiden. Damit gehen unterschiedliche Typen von Boundary Objects einher. Infolgedessen unterscheidet sich auch systematisch, wie Anwendende und technisch Entwickelnde sich in die Entwicklungsprozesse einbringen können und wie die Entwicklungsprozesse verlaufen.

verortete Akteur:innen beziehen. Sie ermöglichen wechselseitige Verständigung und Koordination unterschiedlicher Partner:innen (Star/Griesemer 1989). Dabei zeigt die bisherige Forschung aber auch, dass diese nicht nur Verständnis füreinander ermöglichen, sondern den Entwicklungsprozess diesem entgegen strukturieren können und die Überwindung der Feldgrenzen blockieren. Dies geschieht zum Beispiel, indem sie Akteur:innen aus unterschiedlichen Feldern ungleichen Einfluss ermöglichen und spezifische Verständigungsergebnisse nahelegen (Barrett/Oborn 2010; Leonardi/Bailey/Pierce 2019). Diese Perspektive unterstützt dabei, die Interaktion der Felder über den Entwicklungsgegenstand zu fokussieren (Leonardi/Bailey/Pierce 2019; Hahn 2013: 245f.). Als gelungene Form der Verständigung gilt dabei ein Prozess, der es den jeweiligen Partner:innen ermöglicht, ihr Wissen in die Gestaltung für alle verständlich einbringen zu können. Die so gewonnene Diversität der Zugänge in der Entwicklung reduziert im Ideal die Kontingenz späterer Anwendungsfälle.

In diesem Beitrag argumentieren wir, dass sich Deep Learning Techniken (DLT) als Boundary Objects systematisch von anderen Entwicklungsgegenständen unterscheiden und die Verständigung und Koordination zwischen DLT-Anwendenden und DLT-Entwickler:innen – im Folgenden Programmierende genannt – auf spezifische Weise strukturieren. Wir werden zunächst DLT, Soziale Felder und Boundary Objects einführen, um unser Argument anhand des Forschungsstandes zu entwickeln. Anschließend werden wir das grundsätzliche Argument an empirischem Material konkretisieren. Hierfür ziehen wir zwei Fallbeispiele heran, deren Grundlage Interviews mit Beteiligten solcher Projekte sind, die inhaltsanalytisch ausgewertet wurden. Im Zentrum stehen dabei die DLT-Entwicklungsprozesse aus zwei Projekten, in denen *supervised* DLT zur Bildanalyse in den Lebenswissenschaften entwickelt werden. Anhand dieser Daten zeigen wir, dass DLT als Boundary Objects zu einer paradoxen Einbindung von Lebenswissenschaftler:innen (als DLT-Anwendende) führen: Einerseits erfordert die Entwicklung von DLT eine starke Einbindung der Akteur:innen aus dem Anwendungsfeld, da die Programmierenden (technische Entwickler:innen) von Anfang an darauf angewiesen sind, dass Akteur:innen aus dem Anwendungsfeld ihr Wissen in Form von Lerndaten zur Verfügung stellen und diese von ihnen so aufbereitet werden, dass der Algorithmus von ihnen lernen kann. Ohne adäquate Lerndaten der Lebenswissenschaftler:innen ist es kaum möglich, den DLT-Entwicklungsprozess

erfolgreich abzuschließen.³ DLT als Boundary Objects fördern dementsprechend eine konsequente und frühe Einbindung der Akteur:innen aus dem Anwendungsfeld in den Entwicklungsprozess. Zugleich legt diese Einbindung andererseits aber eine selektive Form der Integration nahe. Innerhalb der Entwicklungsprojekte zeigen sich Tendenzen, die Angewiesenheit auf das Wissen aus dem Anwendungsfeld zu reduzieren und so den Verständigungsprozess zu umgehen. Die DLT ersetzen weitestgehend das Verstehen der Perspektive der Anwendenden (vermittelt durch die von diesen bereitgestellten Lerndaten). Dies reduziert die Relevanz für Programmierende, sich intensiv mit den Anwendenden und den alltäglichen Praktiken des Anwendungsfeldes zu befassen. Diese stark fokussierte und selektive Einbindung der Anwendenden in DLT-Entwicklungsprojekte folgt der Logik einer Generalisierung und verdeutlicht eine Relevanzminderung der alltäglichen Anforderungen der Anwendungsfelder.

Der Artikel leistet damit insbesondere zweierlei. Zunächst trägt er zu einem besseren Verständnis der Spezifika von DLT-Entwicklungsprozessen bei, auf deren Grundlage Szenarien zur Entwicklung künstlicher Intelligenz, die auf der Analyse der technischen Möglichkeiten beruhen, sinnvoll ergänzt werden können. Weiterhin schließt der Artikel an die kritische Forschung zu Boundary Objects an. Entgegen der Annahme, Boundary Objects wirkten grundsätzlich verbindend und egalisierend, weist dieser Beitrag darauf hin, dass Boundary Objects auch Ungleichheiten schaffen oder verfestigen können (Barrett/Oborn 2010; Hawkins/Pye/Correia 2017; Levina/Vaast 2005). Daran anschließend zeigen wir, dass Boundary Objects nicht nur dabei helfen können, Grenzen zu überwinden, sondern es ebenso möglich ist, dass sie das Aufrechterhalten dieser Grenzen fördern.

2. Deep Learning Techniken als Boundary Objects zwischen Anwendungs- und Entwicklungsfeld

Im Folgenden werden wir unser Argument entwickeln, dass DLT als Boundary Objects das Verhältnis von Anwendenden und Programmierenden in DLT-Entwicklungsprojekten systematisch strukturiert und damit die empirische Un-

3 Dabei ist anzumerken, dass wir uns auf supervised Learner konzentrieren, denen die Muster, die sie erkennen sollen, beigebracht werden und diese somit beim Lernen überwacht werden.

tersuchung vorbereiten, die genauer herausstellen soll, in welcher Weise dies geschieht. Dazu werden wir in diesem Abschnitt die drei für unseren Beitrag zentralen Konzepte *Deep Learning Techniken*, *Soziale Felder* und *Boundary Objects* einführen und auf unseren Anwendungsfall beziehen.

Deep Learning Techniken

In den Lebenswissenschaften kommen immer häufiger Deep Learning Techniken (DLT) zum Einsatz. Sie entwickeln diese in gemeinsamen Projekten mit Programmierenden, um aus großen Datenmengen Muster abzuleiten und diese Mustererkennung für weitere Datensätze zu automatisieren. Fortschritte wurden in den letzten Jahren besonders im Bereich der Bildverarbeitung erlangt. Inzwischen stehen riesige Mengen an Bilddaten für das Training bereit und moderne und leistungsstarke Computer sind in der Lage, diese zu verarbeiten, sodass DLT in spezifischen Aufgabenfeldern mittlerweile bessere oder ähnlich verlässliche Ergebnisse wie Menschen hervorbringen (Hosny et al. 2018). Machine Learning produzierte in den vergangenen zehn Jahren besonders dann Schlagzeilen, wenn die besten menschlichen Spieler:innen zum Beispiel im Spiel »Go« von Programmen geschlagen wurden (Serre 2019). Doch neben diesen – die Aufmerksamkeit bündelnden – Ereignissen verbreitet sich die Technik auch zunehmend im Alltag kleinerer Forschungs- und Entwicklungsprojekte. Dies liegt daran, dass DLT heutzutage nach einem Baukastenprinzip und auf der Grundlage vorgefertigter Programmzeilen programmiert werden kann und die technische Hardware für jede:n verfügbar ist. Die Medizin, bzw. weiter gefasst die Lebenswissenschaften, sind der Anwendungsbereich, in dem die Einbindung der DLT in verschiedene Projekte am weitesten fortgeschritten ist (Buchkremer/Witte/Matusiewicz 2020: 388). Dies spiegelt sich auch in dem großen Investitionsaufkommen wider (Razzak/Naz/Zaib 2018: 335). In den Lebenswissenschaften können DLT zum Beispiel wie in den von uns ausgewählten Beispielen unterstützend zur Detektion von Aneurysmen in der Radiologie oder zur Beobachtung von Zellwachstumsprozessen in der Biologie eingesetzt werden.

Trotz der zurückliegenden Entwicklungssprünge ist die alltägliche und vor allem medial gestützte Wahrnehmung von Künstlicher Intelligenz im Sinne eines Hypes zu verstehen (Aust 2021). Überbordende Erwartungen und Ängste gegenüber der Technologie, vermittelt über das Narrativ einer Ersetzung des Menschen durch die Technik wie am Beispiel »Go«, geraten allzu schnell in die Gefahr, dem tatsächlichen Entwicklungsstand nicht gerecht zu werden. Denn

Probleme, wie die begrenzte Fähigkeit, abstrakte Repräsentationen zu bilden und über die Trainingsdaten hinaus zu generalisieren, sind längst in Fachkreisen bekannt (Serre 2019). Ein Blick in aktuelle Entwicklungsprozesse mag so als ein Beitrag zur Korrektur oder Einordnung dieser Erwartung dienen. In Anbetracht des Entwicklungsstandes von DLT und im Kontrast zur bisherigen DLT-Forschung erscheint es vielversprechend, einen Schritt zurückzutreten und statt der vielfach angestellten Überlegungen zu den Folgen und Risiken bei der möglichen Implementierung von DLT in verschiedene Anwendungsbereiche zunächst DLT-Entwicklungsprozesse genauer in den Blick zu nehmen. Hier stellt sich insbesondere die Frage, wie DLT aufgrund ihrer spezifischen soziotechnischen Qualitäten die Entwicklungsprozesse und das Verhältnis der beteiligten Akteur:innen auf charakteristische Art und Weise strukturieren. Welche Rolle spielen DLT in solchen interdisziplinären Forschungsprojekten? Welche Probleme zeigen sich in den Entwicklungsprozessen? Erfordern DLT aufgrund ihres spezifischen Charakters – scheinbar leicht in der Anwendung, aber höchst komplex in ihrer Grundstruktur – besondere Verständigungsformen innerhalb der Entwicklungsteams? Was bedeutet das für die weitere Diffusion dieser Innovation?

Damit entwickeln wir an dieser Stelle eine Perspektive, die DLT in ihrer Gesamtheit als Technologie zu begreifen. Die Aufbereitung, Verarbeitung und Ausgabe der Daten prägen im Entwicklungsprozess ein ganzes Set unterschiedlicher Programme und informationstechnische Architekturen. Gleichzeitig ist unser Fokus weniger darauf gerichtet, wie die Technologie arbeitet, sondern wie mit ihr gearbeitet, geforscht und wie sie entwickelt wird.

Soziale Felder

DLT werden wie andere avancierte Techniken heute typischerweise partizipativ entwickelt und Anwendende von Anfang an eingebunden. Die Forschung zur partizipativen Technikentwicklung hat gezeigt, dass die Konstellationen der verschiedenen Akteur:innen, die an diesem Prozess beteiligt sind, herausfordernde soziotechnische Strukturen sind, die spezifische Dynamiken, Interaktionen und Diskontinuitäten bei der Vermittlung heterogener Wissensbestände produzieren. In unserem Beitrag geht es uns vor diesem Hintergrund darum, zu verstehen, wie diese Konstellationen im Fall von DLT-Entwicklungsprozessen konstituiert werden. Insbesondere werden wir dabei danach fragen, welche Rolle die spezifischen Qualitäten von DLT bei der

Formation der interdisziplinären Akteurskonstellationen spielen. So zeigen sich beispielsweise konkret beim ›supervised learning‹ Herausforderungen in der Phase der Annotation der Lerndaten, bei der Expert:innen des Anwendungsfeldes ihr Wissen integrieren müssen, während bei der Optimierung und Anpassung des Learners Programmierende die Führung übernehmen.

In der Forschung zu Entwicklungsprojekten, in denen Anwendende temporär mit Programmierenden zusammenarbeiten, hat sich das Konzept des Sozialen Feldes als besonders fruchtbar erwiesen (Levina/Vaast 2005; Leonardi/Bailey/Pierce 2019). Für das Konzept sozialer Felder sind vor allem zwei Aspekte charakteristisch (Martin 2003; Fligstein/McAdam 2011; Roth 2019): erstens die inhärente Logik der Praxis in Feldern und zweitens die relative Position der Akteur:innen. Beide Aspekte sind eng miteinander verbunden. Die inhärente Logik von Feldern legt im Wesentlichen die Regeln fest, nach denen Akteur:innen in einem Feld soziale Anerkennung (symbolisches Kapital) und andere Formen von Kapital akkumulieren können, die innerhalb (und außerhalb) des Feldes als erstrebenswert definiert werden und Gestaltungsmöglichkeiten konstituieren. Die hierarchischen Positionen sind dementsprechend eine wichtige Triebkraft für die alltägliche Reproduktion der feldspezifischen Logiken.

Felder und ihre inhärenten praktischen Logiken werden produziert, reproduziert und verändert, indem Akteur:innen immer wieder in interaktive Praktiken miteinander eintreten und dabei Praxiswissen generieren und teilen (Berends/Sydow 2019; Barley/Tolbert 1997; Bourdieu 2009: 174; Giddens 1997: 52). Das geteilte Wissen konstituiert ähnliche Habitus und Institutionen (Barley/Tolbert 1997; Hallett 2003). Habitus werden als geteilte Wahrnehmungs-, Bewertungs-, Denk- und Handlungsschemata verstanden (Bourdieu 1993: 93ff; Kraus 2004; Lizardo 2004). Akteur:innen, die solche habituellen Schemata teilen, richten ihre Handlungen routinemäßig an denselben impliziten und expliziten Regeln und Zielen aus. Darüber hinaus ermöglichen ihnen gemeinsame Schemata, zuverlässige Annahmen darüber zu treffen, wie andere ihre Handlungen und Äußerungen interpretieren und bewerten werden. Diese gemeinsamen Erwartungen definieren Institutionen (Battilana 2006; Berger/Luckmann 1980; Friedland/Alford 1991). In dem Maße, in dem Akteur:innen nach erfolgreichem Verständnis, Anerkennung, sozialem Status und damit verbundenen Ressourcen im Feld streben und umgekehrt soziale Sanktionen vermeiden, orientieren sie ihr Handeln nicht nur routinemäßig, sondern auch strategisch an den impliziten und expliziten Regeln des Feldes (Friedland/Alford 1991; Bourdieu/Wacquant 2006: 129ff; Fligstein/McAdam

2011). Grenzen zwischen Feldern entstehen dort, wo sich das auf vergleichbare Praktiken bezogene Wissen – und die darin begründeten Institutionen und habituellen Dispositionen – voneinander unterscheiden (Levina/Vaast 2005, 2008; Bourdieu/Wacquant 2006; Roth 2019).

Das Feldkonzept ist besonders prominent auf gesellschaftliche Teilbereiche wie Wissenschaft und Wirtschaft angewandt worden. Weiterhin wurde das Konzept aber auch für die Analyse subtilerer Unterschiede zwischen kleineren Kollektiven fruchtbar gemacht (Roth 2019; Levina/Vaast 2005). Insbesondere zur Analyse von Entwicklungsprojekten, in denen Akteur:innen mit heterogenem praktischem Wissen temporär kollaborieren sollen, hat sich das Konzept als fruchtbar erwiesen (Carlile 2002; Majchrzak/Jarvenpaa/Bagherzadeh 2015; Berends/Sydow 2019). In interdisziplinären Kollaborationsprojekten, Partnerschaften zwischen Wissenschaft und Wirtschaft oder bei der Einbindung von Anwendenden in Entwicklungsprozesse, geht es in der Regel darum, die heterogenen Perspektiven und Kenntnisse der unterschiedlichen Akteur:innen erfolgreich zu kombinieren, um Innovationen zu schaffen. Das Feldkonzept weist jedoch darauf hin, dass die Differenzen im Wissen der Akteur:innen oftmals tiefgreifender sind als das auf den Innovationsgegenstand bezogene Faktenwissen. Diesbezügliche Unterschiede können für Kollaborationen grundlegende Praktiken wie die Identifikation der im Entwicklungsprozess auftretenden Probleme, der erstrebenswerten Ziele, die Organisation von Zusammenarbeit oder die Kommunikation betreffen. Erschwerend kommt weiterhin hinzu, dass dieses feldspezifische Wissen von den Akteur:innen in hohem Maße implizit angewandt und als selbstverständlich vorausgesetzt wird. Zwischen Kooperationspartner:innen aus unterschiedlichen sozialen Feldern ergeben sich daher oftmals Interessens-, Koordinations- und Verständigungsschwierigkeiten (Carlile 2002; Gassol 2007; Leonardi/Bailey/Pierce 2019). Die unterschiedlichen Wissensbestände, die oft als Chance für Innovationen und wechselseitiges Lernen gesehen, propagiert und befördert werden, sind dementsprechend zugleich eine Barriere für erfolgreiche Zusammenarbeit.

Voraussetzung dafür, dass diese Barrieren überwunden und Kooperationen über Feldgrenzen hinweg erfolgreich umgesetzt werden können, ist, dass es gelingt, die Unterschiede zwischen den Feldern temporär zu überwinden, indem relevantes Praxiswissen geteilt, generiert, habitualisiert und institutionalisiert wird und es zu gemeinsamen Praktiken zwischen Akteur:innen aus den beiden beteiligten Feldern kommt. Erfolg heißt hierbei aber nicht nur, dass ein spezifisches Produkt entwickelt wird, sondern er misst sich

auch an dem Grad der wechselseitigen Verständigung. Nach dem Konzept der Sozialen Felder gelingt die Verständigung nur innerhalb eines Feldes, sodass Entwicklungsprozesse an der Konstitution eines gemeinsamen Verständnisses erst arbeiten müssen. Die spezifischen kulturellen Logiken ergeben sich im Detail aus den näheren Praxiszusammenhängen der Akteur:innen, womit die kulturelle Differenzierung ausgehend von den makroskopischen Feldern wie Politik, Wirtschaft oder Wissenschaft auf kleiner geartete Interaktionskonstellationen bzw. Arbeitsgemeinschaften herunterzukalieren ist.⁴ In den untersuchten Projekten sind es die individuellen Forschungskontexte, mit der spezifischen Fachcommunity, den Konferenzen und Fachzeitschriften, die dem jeweiligen wissenschaftlichen Streben die strategische Logik geben.

Bezogen auf DLT-Entwicklungsprojekte sind die Felder das *Entwicklungsfeld*, indem DLT-Programmierende DLT entwickeln, sich mit anderen Programmierenden in diesem Feld austauschen und sich ihnen gegenüber profilieren und das *Anwendungsfeld*, indem etwa Ärzt:innen alltäglich Patient:innen Diagnosen stellen, sie untersuchen und behandeln oder in denen Biolog:innen Bildmaterial analysieren und sich darüber mit ihren Kolleg:innen austauschen. Die Bereiche verfügen über spezifische disziplinäre Kenntnisse, die für die Entwicklung von DLT-Anwendungen für bildgebende Verfahren in der Medizin und der Biologie entscheidend sind.

An den hier fokussierten Entwicklungsprojekten von DLT für die Lebenswissenschaften sind typischerweise Akteur:innen aus der Informatik und dem jeweiligen lebenswissenschaftlichen potentiellen Anwendungsbereich beteiligt. In der bisherigen Forschung zu Softwareentwicklung hat sich immer wieder gezeigt, dass sich die sozial geteilten Annahmen über Funktion, Anwendung und Nutzer:innen der DLT systematisch voneinander unterscheiden und deren Ausgestaltung auf unterschiedliche Weise strukturieren können (Degele 1996; Amershi et al. 2019). Typischerweise fokussieren Programmierende die effiziente sowie in sich schlüssige Gestaltung der Software und blenden die spezifische Anwendungssituation und die sich daraus ergebenden Bedarfe weitestgehend aus. Sie werden nur am Rande und insofern berücksichtigt, als Programmierende von ihrer Nutzungssituation auf die

4 Disziplinen sind hierbei, wie Schmitt (2019: 65) schreibt, ungenau, wenn es darum geht die kulturellen Unterschiede in der jeweiligen Forschungspraxis von Wissenschaftler:innen zu beschreiben. Möglich wäre es auch sich hier über die Begriffe der epistemischen Kulturen (Knorr Cetina 2002) oder den wissenschaftlichen Produktionsgemeinschaften (Gläser 2006) zu nähern.

Situation der zukünftig Nutzenden schließen. Solche Schlüsse zeichnen sich in der Regel durch eine rationalistische, stark an Stereotypen ausgerichtete und im Hinblick auf die zentrale Funktion der Anwendung reduktionistische Konzeption der Anwendungsszenarien und Interessen der Anwendenden aus. Dominieren Programmierende also die DLT-Entwicklungsprozesse, erhöht dies das Risiko, dass DLT entwickelt werden, die in der Anwendung nicht praxistauglich sind und somit die Entwicklungsprojekte gegebenenfalls scheitern. Gelingt es doch, die stark durch die Programmierenden geprägten DLT in der Praxis zu etablieren, sind dagegen stärkere Diskontinuitäten in den Anwendungsfeldern zu erwarten. Dabei werden die in die DLT eingeschriebenen Annahmen der Programmierenden tendenziell praktisch reproduziert, weil Anwendende bei der Nutzung dazu veranlasst werden, diesen Erwartungen zu entsprechen. Ziele und Praktiken der Anwendenden geraten dabei gegenüber dem Bestreben der Programmierenden, die DLT in sich zu optimieren, in den Hintergrund.

Für die Ausgestaltung der DLT-Entwicklungsprojekte und ihrer nachhaltigen Folgen ist es entscheidend, inwiefern das Entwicklungs- und Anwendungsfeld durch die Projekte punktuell und temporär integriert werden (Levina/Vaast 2008; Leonardi/Bailey/Pierce 2019). Ausschlaggebend dafür ist das konkrete Verhältnis, das Anwendende und Programmierende in DLT-Entwicklungsprojekten zueinander einnehmen und ob eine Verständigung angestrebt wird und gelingt. Definiert wird dies besonders dadurch, inwiefern Akteur:innen gemeinsam verständigungsorientiert aufeinander Bezug nehmen oder ob versucht wird, feldspezifische Eigenlogiken gegeneinander durchzusetzen. Wir fragen dementsprechend im Folgenden danach, wie Anwendende und Programmierende sich und die Eigenlogiken ihrer Felder im Rahmen von DLT-Entwicklungsprojekten in ein Verhältnis zueinander stellen.

Boundary Objects

In der Forschung zu interdisziplinären Entwicklungsprojekten hat sich gezeigt, dass Boundary Objects das Lernen und die Relationen zwischen den Feldern entscheidend strukturieren können. Wir werden im folgenden Deep Learning Techniken selbst als Boundary Objects konzipieren und danach fragen, wie diese spezifische Technologie die Relationen zwischen Anwendenden und Programmierenden strukturiert. Als Boundary Objects definieren Star und Griesemer abstrakte oder konkrete Gegenstände, »which are both plastic

enough to adapt to local needs and the constraints of the several parties employing them, yet robust enough to maintain a common identity across sites.« (1989: 393)

Die dem Ansatz zugrunde liegende Annahme ist, dass, indem sich Akteur:innen mit verschiedenen Kenntnissen und Perspektiven gemeinsam auf dieselben Boundary Objects beziehen, diese durch geteilte Bezüge eine bessere Verständigung ermöglichen können. Zugleich können Boundary Objects wirksam werden, indem sie den Einfluss verschiedener Akteur:innen auf den Entwicklungsprozess verstärken oder reduzieren.

In Entwicklungsprojekten kommen Prototypen der zu entwickelnden Artefakte eine zentrale Rolle dabei zu, die Explikation impliziten Anwenderwissens zu unterstützen und somit eine Verständigung zwischen Programmierenden und möglichen Anwendenden zu fördern. Beispielsweise explizieren Programmierende selbstverständliche Annahmen, indem sie ihre Programme an Prototypen ausrichten. Mögliche Anwendende können sich dann, wenn sie das Programm in die Anwendung übertragen sehen, an den entsprechenden Ausgestaltungen stoßen. Anhand der entsprechenden Reibungspunkte lassen sich Differenzen zwischen den impliziten Annahmen identifizieren, die deren Artikulation befördern. So können die gemeinsamen Bezüge auf Boundary Objects ein besseres Verständnis für die Perspektiven der anderen schaffen (Leonardi/Bailey/Pierce 2019). Neben physischen Prototypen (Bechky 2003; Carlile 2002) werden in der Literatur beispielsweise auch Skizzen (Bødker 1998), Anwendungsszenarien (Bødker 2000) oder Formulare (Star/Griesemer 1989) als Boundary Objects identifiziert. Entscheidend ist jeweils weniger die Art der Objekte, sondern vielmehr, dass sie zur interaktiven Verständigung zwischen unterschiedlichen Partner:innen beitragen (Levina/Vaast 2005; Barrett/Oborn 2010).

Wie oben bereits erwähnt, werden Anwendende heute typischerweise in (DLT-)Entwicklungsprozesse partizipativ eingebunden. Eine erhebliche Schwierigkeit besteht in interdisziplinären Projekten insofern, als die Programmierenden und Anwendenden ihre praktischen Bedarfe nicht ohne weiteres artikulieren können. Das Wissen über ihre Praxis ist in aller Regel in hohem Maße implizit und schließt zahlreiche unhinterfragte Annahmen ein. Um das Wissen zu erschließen und in den Entwicklungsprozess einzubeziehen, bedarf es daher einer systematischen Explikation des feldspezifischen Wissens, die sich kaum alleine sprachlich realisieren lässt (Roth/Diefenbach 2021). Entscheidend dafür, ob das implizite Anwender:innenwissen erschlos-

sen werden kann und in welchem Maße es Niederschlag in der Gestaltung der DLT findet, kann durch das Wirken von Boundary Objects beeinflusst werden.

Untersuchungen, die Boundary Objects in Entwicklungsprozessen betrachten, zeigen weiterhin, dass diese weder an sich eine spezifische Wirkung auf Entwicklungsprozesse und Akteurskonstellationen entfalten, noch, dass Boundary Objects in dem Sinne neutral sind, dass Akteur:innen sie nach Belieben und entsprechend ihrer reflexiven Ziele einsetzen können. Vielmehr entscheidet die spezifische praktische Einbindung von Objekten, ob sie zu Boundary Objects werden und welche spezifische Wirkung sie als solche entfalten (Leonardi/Bailey/Pierce 2019; Levina/Vaast 2005). Wie Boundary Objects wirken, ergibt sich dementsprechend aus dem praktischen Zusammenspiel der bestehenden Akteurskonstellationen und den technischen Artefakten. Beide strukturieren diesen Prozess in gewissem Maße durch ihre Charakteristika.

Die kritische Forschung zeigt daran anschließend auch, dass die Anwendung von Boundary Objects nicht notwendigerweise zu einer wechselseitigen Verständigung und zu einer egalitären Beteiligung von Programmierenden und Anwendenden führen muss. Vielmehr zeigt sich, dass Boundary Objects von Macht durchdrungen sind (Hawkins/Pye/Correia 2017; Barrett/Oborn 2010). Die Gestaltung und das Verstehen von Boundary Objects setzt mitunter spezifische Kenntnisse und Ressourcen voraus. Sind diese Kenntnisse und Ressourcen asymmetrisch zwischen den Akteur:innen verteilt, sind diejenigen, die über die entsprechenden Ressourcen und Kenntnisse verfügen, in höherem Maße dazu in der Lage, ihre Interessen, aber auch ihre unhinterfragten Annahmen im Entwicklungsprojekt durchzusetzen (Barrett/Oborn 2010; Levina/Vaast 2005). Wir verstehen Boundary Objects dementsprechend als Objekte, auf die sich Akteur:innen aus verschiedenen Feldern interaktiv beziehen und gehen davon aus, dass diese Bezüge das Verhältnis der Akteur:innen strukturieren. Boundary Objects können Verständigung und Koordination über Feldgrenzen hinweg dementsprechend symmetrisch befördern. Sie können aber genauso gut asymmetrische Konstellationen hervorbringen, indem sie Verständigung und Koordination selektiv verhindern.

Der Forschung zu Boundary Objects folgend gehen wir davon aus, dass DLT als Boundary Objects im Entwicklungsprozess eine wichtige Rolle bei der Formation von Akteurskonstellationen spielen. Weiterhin nehmen wir an, dass sich die besonderen Qualitäten von DLT in besondere Akteurskonstellationen übersetzen, die über diese Perspektive sichtbar gemacht werden können. Im Folgenden geht es uns darum, diese Zusammenhänge anhand

zweier Fallbeispiele genauer zu beleuchten. Um die spezifischen Verschiebungen innerhalb von Entwicklungsprozessen differenziert nachvollziehen und identifizieren zu können, untersuchen wir die Konstellationen innerhalb und entlang der einzelnen Phasen und Momente der DLT-Entwicklungsprojekte.

3. Methode

Ausgehend von dem Vorhaben eines Arbeitskreises von Mitarbeiter:innen des Lehrstuhls für Technik- und Organisationssoziologie der RWTH Aachen wurden von 2019 bis 2021 Interviews mit Akteur:innen geführt, die sich in ihren Forschungsaktivitäten und Entwicklungsprojekten mit Machine Learning und Deep Learning auseinandersetzen bzw. maßgeblich in entsprechenden Projekten involviert sind oder diese initiieren.

Aus dem Korpus der geführten Interviews ziehen wir drei von den Autor:innen dieses Beitrags geführte Interviews heran. Diese Interviews beziehen zwei Projekte ein, in denen DLT für eine gezielte Anwendung im Bereich der Bildanalyse entwickelt wurden. Diese beiden Projekte werden als exemplarische Fälle herangezogen und im Folgenden näher beleuchtet.

Dieser Beitrag widmet sich somit zwei anwendungsorientierten Entwicklungs- bzw. Forschungsprojekten, in denen DLT programmiert wurden, um Bilddaten aus dem Bereich der Lebenswissenschaften zu analysieren. Die Betrachtung der beiden Projekte setzt sich aus einer Kombination der jeweiligen Einzelfallanalysen und einer vergleichenden Fallanalyse zusammen. Die Auswahl der beiden Fälle sowie das explorative analytische Vorgehen zielen sowohl darauf ab, etwaige spezifische Gemeinsamkeiten in den Konstellationen und Positionierungen beleuchten zu können, als auch darauf, herausstechende Unterschiede innerhalb solcher anwendungsnahen interdisziplinären DLT-Projekte identifizieren zu können (Yin 2003; Borchardt/Göthlich 2007).

Grundlage der Fallstudien sind drei persönlich geführte problemzentrierte, teilleitfadengestützte Interviews mit Akteur:innen von DLT-Entwicklungsprojekten, die 2019 und 2020 geführt wurden. Die drei Interviewpartner:innen zeichnen sich dadurch aus, dass sie sowohl über Expertise des jeweiligen Anwendungsbereiches als auch über Expertise in der Entwicklung von DLT verfügen. Gleichwohl sie sich klar in einem der Felder verorten lassen, erwarten wir, dass sie in den betrachteten und vorangehenden Projekten als *Boundary*

*Spanner*⁵ fungieren und dadurch in besonderem Maße dazu in der Lage sind, die unterschiedlichen Perspektiven, Interessen und Eigenlogiken der involvierten Anwendungs- und Entwicklungsfelder sowie deren Zusammenspiel in den Projekten zu beschreiben. Befragt wurde ein Universitätsprofessor für Radiologie, der ebenso die Leitung einer Klinik innehat (A2). Er vermittelte im Anschluss an das Interview einen in seinen Projekten tätigen Medizingenieur, mit dem ein weiteres Interview stattfand (A1). Des Weiteren wurde ein Interview mit einem Juniorprofessor für biomedizinische Bildverarbeitung geführt, der in dem Interview seine Projektstätigkeiten als Bioinformatiker schildert (B1). Die Interviews waren teilstandardisiert gestaltet und dem explorativen Charakter des Vorhabens entsprechend narrativ ausgerichtet (Borchardt/Göthlich 2007; Lamnek 2008). Die Interviews bieten somit einen besonderen Zugang zu den Erfahrungen und zu den Perspektiven auf die eigene Tätigkeit. Auch die Positionierungen der beteiligten Akteur:innen in Relation auf die anderen Beteiligten – inklusive der DLT – treten hervor.

Die gewählte explorative Gestaltung ermöglicht ausgehend von den dichten Beschreibungen ein tiefergehendes Verständnis der Prozesse, der Entwicklungen und den Verschiebungen innerhalb dieser Forschungs- und Entwicklungsprojekte. Im Besonderen erscheint dies erkenntnisversprechend, um eventuell konstitutive Momente in solchen DLT-Entwicklungsprozessen und charakteristische Akteurskonstellationen, zu erkennen. Trotz der Limitationen dieses Vorgehens versprechen wir uns erste Beschreibungen und spezifische Einblicke zu generieren, die Grundlage für weiterführende Fragestellungen zu den interdependenten soziotechnischen Konstellationen der Akteur:innen in solchen Projekten sein können (Borchardt/Göthlich 2007).

Die Interviews wurden transkribiert und mithilfe von MAXQDA für die inhaltlich strukturierende Inhaltsanalyse nach Kuckartz aufbereitet (Kuckartz 2018). Das Kategoriensystem wurde mittels eines deduktiv-induktiven Vorgehens entwickelt, ergänzt und fortlaufend ausdifferenziert.

5 Boundary Spanner dient als Bezeichnung für Akteur:innen, die den Austausch des fachlichen Wissens unterschiedlicher Felder erleichtern, ermöglichen und die somit bestehende (disziplinäre) Grenzen überbrücken (Levina & Vaast 2005).

4. Fallbeispiele: DLT Entwicklungsprojekte in den Lebenswissenschaften

Im Folgenden werden zunächst die beiden Fallbeispiele beschrieben, um ein Verständnis ihres Entstehungs- und Entwicklungszusammenhangs herzustellen.

Fallbeispiel A: DLT zur Detektion von Aneurysmen in MRT-Bildern

Dieses Entwicklungsprojekt ist ausgehend von der Qualifizierungsarbeit eines Mediziningenieurs (A1) entstanden, der eine DLT zur Detektion von Aneurysmen entwickelt hat. Als Lerndaten konnte er auf vorhandene Patient:innendaten aus einer Klinik für Neuroradiologie zurückgreifen. Hierbei handelt es sich um MRT-Bilder, die in der Klinik in den letzten 10 Jahren angefertigt wurden, bei denen ein oder mehrere Aneurysmen diagnostiziert wurden (Datensatz aus 85 MRT-Aufnahmen mit 115 diagnostizierten Aneurysmen). Die entwickelte DLT wurde so programmiert, dass er den Bereich, in dem ein Aneurysma identifiziert wurde, im Bild markiert. Die Sensitivität der entwickelten DLT lag ähnlich der von Radiolog:innen bei 90 %. A1 arbeitet an diesem DLT-Entwicklungsprojekt zusammen mit Ärzt:innen, die die bereits bestehenden MRT-Diagnosen überprüfen und die Lerndaten annotieren. Die in den geschriebenen Arztbriefen festgehaltenen zum Teil sprachlich unpräzisen Diagnosen werden von praktizierenden Ärzt:innen bewertet, interpretiert, in »diagnostische Score« (A2) transformiert und so aufbereitet, dass sie als Lerndaten für die DLT fungieren können. In den MRT-Bildern werden die Aneurysmen in jedem Bild exakt markiert, indem die betreffenden Bildbereiche umrandet werden. A1 programmiert dann mit diesen annotierten Lerndaten die DLT, anschließend wird die Güte der DLT an einem anderen Datensatz geprüft und die Ergebnisse wieder von Ärzt:innen bewertet. Die Architektur der DLT wird dabei optimiert und angepasst.

Fallbeispiel B: DLT zur Analyse von Mikroskopbildern biologischer Embryonalforschung

In dem Projekt, das als Fallbeispiel B dient, wird eine DLT entwickelt, die 3D-Bilder und -Videos analysiert, die aus den Aufnahmen von hochauflösenden Mikroskopen stammen. Diese werden zur Analyse von embryonalen Entwicklungsprozessen (hauptsächlich von Zebrafischen) eingesetzt und produzieren

umfangreiche Datenmengen. In diesem Projekt wird eine Kombination aus Detektions-, Segmentierungs- und Trackingnetzwerken entwickelt, die die Zellformänderungen in ihrer zeitlichen Entwicklung erkennen und nachvollziehbar machen können. In diesem Projekt treten die Biolog:innen an den befragten Bioinformatiker (B1) heran, kommunizieren ihm die sie interessierenden Strukturen in den Bildern und bitten ihn, eine DLT für die Bildauswertung zu entwickeln. B1 entwickelt in Folge dieser Abstimmung ein Tool, das die Biolog:innen ähnlich wie das Programm ›Paint‹ nutzen können, um die Bilddaten zu annotieren. Hiermit können sie die sie interessierenden Bildbestandteile präzise kennzeichnen, damit diese als Lerndaten für die Programmierung des Algorithmus eingesetzt werden können. Der nächste Schritt im Prozess nach der Programmierung ist die Optimierung der DLT und die Testung der Detektionsgüte und die daraufhin stattfindende Anpassung des Algorithmus. Im nächsten Schritt wird die entwickelte DLT den Biolog:innen überlassen, die sich gegebenenfalls wieder an den Programmierer wenden, wenn die Qualität der Detektion nicht genügt. Daraufhin würde eine nächste Iterationsschleife und Optimierung durchgeführt werden.

5. Soziotechnische Konstellationen in den DLT-Entwicklungsprojekten

Im Folgenden stehen die spezifischen Charakteristiken von DLT-Entwicklungsprojekten im Fokus. Im Zentrum steht dabei die Frage, wie DLT – verstanden als Boundary Objects – die Akteurskonstellationen in den Entwicklungsprojekten strukturieren.

Erste Beobachtung: Programmierer:innen als Auftragnehmer:innen

In beiden Fallbeispielen werden supervised DLT für eine konkrete Anwendung entwickelt, mittels derer Bilddaten aus den Lebenswissenschaften analysiert werden. Betrachtet man die Entstehungszusammenhänge beider Projekte im Hinblick darauf, zeigen sich Gemeinsamkeiten, aber auch Unterschiede.

In Fallbeispiel A initiiert der Mediziningenieur A1 das Projekt ausgehend von seiner eigenen Qualifizierungsarbeit. Dabei profitiert er von dem für ihn leicht zugänglichen bereits bestehenden Datenpool. Er spricht den Leiter der Klinik (A2) an, um seine Projektidee zu verfolgen. Er begründet die Wahl dieses Themas zum einen mit dem allgegenwärtigen ›Hype‹ um den Einsatz

künstlicher Intelligenz und zum anderen mit den Versprechungen des zukünftigen Einsatzes von DLT als Diagnosetool zur Optimierung des Klinikalltags von Radiolog:innen. Radiolog:innen, die einer hohen Arbeitsbelastung und folgenschweren Diagnose- und somit Entscheidungssituationen ausgesetzt sind, könnten mit diesem zusätzlichen Tool entlastet und ihre Diagnosegüte verbessert, bzw. das Übersehen von Aneurysmen verhindert werden. Hierfür eignet sich A1 die erforderlichen Kenntnisse zur Programmierung an, da er zuvor noch nicht mit DLT gearbeitet hat. Der Leiter der Klinik (A2) wertet den Nutzen des Projektes für die tatsächliche Praxis in der Klinik im Interview und damit nach Abschluss des Projektes als gering ein. Aufgrund der mangelnden zugrundeliegenden Datenmenge und der daraus hervorgehenden mangelnden Generalisierbarkeit besitze das Projekt keine reelle Chance, als Tool im Klinikalltag implementiert zu werden. Es fügt sich somit nicht in das Set der verschiedenen technischen Hilfsmittel ein, die bereits im ärztlichen Alltag zum Einsatz kommen. Letztlich seien solche kleinen Projekte (leider nur) als »Spaßforschung« zu sehen und Neugierde eine wesentliche Motivation.

Da A1 bereits Mitarbeiter des interdisziplinären Teams der Klinik war, konnte das Projekt auf personeller Ebene niedrigschwellig ermöglicht werden. In Hinblick auf die Technik wird erläutert, dass die Daten aus der Klinik für die Forschung genutzt werden konnten, sodass auch hier die Ressourcen intern verfügbar waren. Daneben wurde die notwendige Hard- und Software angeschafft und die DLT erscheint den Mediziner:innen als einfache technische Grundarchitektur, die in Form von Bausteinen zusammengesetzt werden kann. Die Mediziner:innen betreiben das Entwicklungsprojekt in diesem Fall als »Spaßforschung«, auf die sie nicht angewiesen sind. Letztlich stellt sich das Tool als »nice to have« dar und könnte in ganz bestimmten Fällen das Portfolio weiterer Technologien der Klinik erweitern sowie die Entscheidungen in der Radiologie unterstützen. Über den Zustand als Prototyp kommt die DLT-Anwendung nicht hinaus. Es wird auch klar, dass die Forschung nicht aus medizinischer Notwendigkeit heraus erfolgt. In diesem Fallbeispiel zeigt sich eine enge Verknüpfung von Anwendungsfeld und Programmierenden. Verschiedene Hürden verhindern, das Projekt größer zu skalieren, doch durch die leichte Verfügbarkeit der notwendigen Ressourcen können die Mediziner:innen mit der DLT Versuche durchführen und sich ausprobieren.

Im zweiten Fallbeispiel B gestaltet sich die Anbahnung des Projekts anders. Die Entwicklung der DLT entsteht auf die Initiative der Biolog:innen hin, die dem Problem begegnen, dass moderne Mikroskope immens große Datensätze produzieren, die sie selbst nicht mehr mit »klassischen« Methoden auswerten

können. Die Entwicklung der DLT geschieht gewissermaßen als »Auftragsarbeit«, aus der Situation heraus, dass die Datenmenge für sie nicht mehr handhabbar ist und sie darauf angewiesen sind, die Analyse der Bilddaten zu automatisieren und Kooperationen mit Programmierenden anzubahnen.

Die Lebenswissenschaftler:innen sind somit aufgrund der Entwicklung der technologischen Modalitäten dazu aufgefordert, diese durch Automatisierung bearbeit- und handhabbar zu machen. Zugleich mangelt es den Lebenswissenschaftler:innen typischerweise an den dazu erforderlichen Kenntnissen und Kompetenzen. Die entsprechenden Programmierpraktiken entziehen sich daher sowohl ihrer Einsicht als auch ihren Gestaltungsmöglichkeiten. Infolgedessen sind sie auf die technische Expertise von Programmierenden angewiesen. Hierbei bedarf es vor allem informatischer Expertise, da die zu lösenden Probleme vor allem als technische Probleme erscheinen. Die Biolog:innen sind insofern in hohem Maße in ihrer eigenen Disziplin und ihren Möglichkeiten der Problemlösung von der Informatik abhängig. Die Anbahnung des Projekts erscheint dabei jedoch in Form einer Dienstleistung, die erbracht und in Auftrag gegeben wird. Die Entwicklung von DLT scheint sich hier zwischen den Polen von Abhängigkeit und Auftragsarbeit zu bewegen, wodurch die disziplinären Grenzen eher bestehen bleiben, als wenn die Projekte als gemeinsame Forschungs- und Entwicklungsprozesse verstanden werden würden.

Beiden DLT-Entwicklungsprojekten liegt die Erwartung zugrunde, die eigenen Bilddaten durch Automatisierung auszuwerten. Im ersten Fall geht es darum, die Analyse und ihre Güte zu verbessern sowie praktizierende Ärzt:innen im Klinikalltag zu entlasten und im zweiten Fall darum, die Analyse der Datenauswertung für die Erforschung überhaupt erst zu ermöglichen.

In beiden Fällen besitzen letztlich die Anwendenden die Daten z.B. in Form von Bildmaterial und sind in ihren Vorhaben darauf angewiesen, Programmierende einzubeziehen, um das jeweilige Ziel zu erreichen. Im Fallbeispiel A erscheint die DLT als Attraktionsobjekt, dessen allgemeine Versprechungen konkretisiert werden sollen, wobei es bei der letztlichen Umsetzung zu Schwierigkeiten kommt, sodass eine große Disbalance zwischen erster Aneignung und erfolgreicher Umsetzung sichtbar wird. DLT schafft hier also zunächst den Anschein leichter Umsetzbarkeit, der dann durch das Erfordernis für eine größere Datenbasis und disziplinäre Expertise konterkariert wird. Im Fallbeispiel B hingegen zeichnet sich sogleich der hohe Komplexitätsgrad des Problems ab, das bei den Biolog:innen zur Reaktion führt, disziplinäre Expertise zu beauf-

tragen, die dann die DLT zur Lösung einbringt und sie von den Ergebnissen profitieren lässt.

In beiden Fallbeispielen dieser anwendungsnahen Entwicklungsprojekte nehmen die Programmierenden die Funktion von Boundary Spannern ein. Ihre berufliche Ausbildung zeichnet sich dadurch aus, dass sie eine große Expertise in der Schnittmenge der Disziplinen aufbringen. Als Mediziningenieur und als Bioinformatiker besitzen beide schon qua Ausbildung Wissen über die Handlungs- und Wertlogiken von zwei Disziplinen. Sie bilden systematisch parallele Primärreferenzen – und somit eine doppelte Zugehörigkeit – aus, vermögen Felder zu überbrücken und sind so angeregt, Verständnis für die disziplinäre Logik und in diesem Fall das Erkenntnis- und Problemlösungsinteresse des Anwendungsfeldes zu entwickeln. Dennoch wird deutlich, dass die Entwicklung der Technologie und ihre Programmierung das Zentrum ihrer Expertise darstellen. Für B1 gilt dies noch mehr als für A1, da letzterer bereits durch seine Anstellung an einer universitären Klinik auch in praktischer Nähe zum Anwendungsfeld steht. Die Programmierenden sind in der Akteurskonstellation in der Position, die Grenze zwischen den Feldern überbrücken zu müssen. Aber an sie wird die Anforderung gerichtet, Probleme des Anwendungsfeldes zu lösen.

Zweite Beobachtung: Annotieren statt Verstehen

Generell gilt, dass qualitativ hochwertige und zahlenmäßig umfangreiche Lerndaten als »Goldstandard« (A2) betrachtet werden, um eine gute DLT zu entwickeln. Davon ausgehend zeigt sich, dass in den Projekten, in denen supervised DLT für automatisierte Bildanalyse entwickelt werden, die Phase der Datenannotierung⁶ ein konstitutives Element darstellt. Eine charakteristische Konstellation in den betrachteten Fallbeispielen tritt deutlich hervor: Bei der Annotierung – also der Aufbereitung des Datenmaterials als Lerndaten für den Algorithmus – ist die Einbindung der Anwendenden bzw. in unseren Fällen die Einbindung der Lebenswissenschaftler:innen existenziell.

In beiden betrachteten Fallbeispielen übernehmen die Anwendenden die Annotierung der Daten. Dies ist nötig, da die Annotierung der Daten nur mit

6 Im Folgenden verwenden wir bei den Ausführungen zu den Fallbeispielen statt des etablierten Begriffs »Annotation« den im Feld selbst genutzten Begriff der »Annotierung« und verbleiben somit in unmittelbarer Nähe zum Feld und (auch) seiner sprachlichen Praxis.

der spezifischen disziplinären Expertise vorgenommen werden kann. Es geht darum, fachlich fundierte Selektionsentscheidungen vorzunehmen und das Material so aufzubereiten, dass die DLT handhabbare Informationen daraus erhält.

In Fallbeispiel A steht dabei im Vordergrund, die radiologischen Aufnahmen mit den bestehenden schriftlichen Diagnosen aus Arztbriefen in skalierbare Daten umzuwandeln. So geht aus den schriftlichen Diagnosen nicht immer klar hervor, ob ein Aneurysma besteht oder nicht. Diese klare Unterscheidung wird aber seitens der DLT gefordert. Ausgehend von ihrer Expertise treffen praktizierende Radiolog:innen die Entscheidungen, wie diese ›Übersetzung‹ stattfinden kann und inwiefern schriftlich unklare und uneindeutige Diagnosen (z.B. »Blutung nicht sicher erkennbar« (A2)) in einen Score übertragen werden können. Die Aufbereitung der Lerndaten erfolgt somit in zwei Phasen. Zunächst gilt es, in einer Scanabbildung die entsprechenden Strukturen zu markieren, die als Aneurysma in Frage kommen. Bei der Detektion der Strukturen unterscheidet sich die Perspektive der DLT kaum von derjenigen der Ärzt:innen. Beide betrachten ein Bild und suchen nach ungewöhnlichen Mustern in den Daten. Deshalb können bestehende Diagnosen der vergangenen Jahre im Fall A einfach nochmal aufbereitet und gleichermaßen für den Learner übersetzt werden. Die darauffolgende Phase ist die der Klassifikation der detektierten Muster. Hierzu müssen die schriftlichen Diagnosen zu den entsprechenden detektierten Strukturen in einen Score umgewandelt werden.

Im Fallbeispiel B muss zunächst ein Tool programmiert werden, das die Daten zur Annotierung verfügbar macht. Mittels dieses von dem Programmierer entwickelten Tools können die Biolog:innen die relevanten Zellen in den einzelnen Bildern markieren und selektieren. »Genau, das ist so der erste Schritt, dass man einfach mal weiß, was will der Biologe überhaupt haben und dann dafür eben dann diese Label erstellt.« (B1) Dieses Tool ermöglicht zum einen, ein Verständnis des Programmierers für das Erkenntnisinteresse der Biolog:innen zu entwickeln, aber auch den Biolog:innen verständlich zu machen, was die Anforderungen für das Lernen und Programmieren der DLT sind. Die Expertise der Biolog:innen schlägt sich in diesem Prozess in der Annotierung der Lerndaten nieder und wird dann so aufbereitet, dass sie zugänglich für die DLT sind und in die Entwicklung eingespeist werden können. Dennoch besteht seitens des Programmierers keine genaue Kenntnis dessen, was die Label spezifisch bedeuten. So gilt es in dieser Projektphase eine Verständigung darüber zu erlangen, was das Erkenntnisinteresse der Anwendenden ist. Theore-

tisch übersetzt heißt dies, dass die Programmierenden hier den Schritt in das fremde Feld wagen müssen. Das Erkenntnisinteresse muss in die Technologie übersetzt werden, die jedoch selbst Anforderungen bereithält. Diese Anforderungen müssen dabei die Programmierenden vermitteln. In dieser Phase lässt sich die intensivste Zusammenarbeit erkennen.

Die Vorbereitung der Annotierung der Lerndaten findet in Kooperation statt und ist stark geprägt von einem jeweiligen ›aufeinander Angewiesensein‹ auf die Expertise. Um die Technik zur Lösung des Problems nutzen zu können, bedarfes einer Verständigung darüber, wie die Daten aufbereitet werden müssen, sodass ein hochwertiger Lerndatenkorpus zustande kommt. Die Aufforderung zur Mitarbeit stellt aus Sicht des Programmierers B1 in dem Dienstleistungsverhältnis kein Problem dar: »Das ist quasi kein Problem, ihnen zu kommunizieren, dass wir jetzt irgendwie Annotationen von ihnen brauchen, um eben solche Modelle zu trainieren.« (B1)

In Fallbeispiel B fordert B1 als Programmierer die nötige Annotierung der Daten und ihrer Erfordernisse ein und schildert, was er braucht. Es findet aber in beiden Projekten eine Verständigung darüber statt, was die wechselseitigen Bedarfe sind. Die Kommunikation geht jedoch nicht über den Austausch der Bedarfe hinaus, sondern sie erfolgt vielmehr in Form einer ›Übergabe‹ der Verantwortung für den darauffolgenden Teilschritt, der Durchführung der Annotierung. Die Verantwortlichkeiten sind nach wie vor verteilt und klar entlang der Projektphasen getrennt. Kommuniziert wird vor allem zur ›Übergabe‹ in den nächsten Zuständigkeitsbereich, ohne dabei ganz genau zu verstehen, was das Gegenüber macht. Die beiden Disziplinen bleiben weitestgehend über die unterschiedlichen Zuständigkeiten und unterschiedlichen Expertisen, die eingebracht werden, voneinander getrennt und bringen durch die Verschiebungen wechselseitige Abhängigkeiten in den einzelnen Phasen hervor. »Im Prinzip bleibt einem nicht viel anderes übrig, genau. Also ich als Informatiker kann das teilweise nicht so richtig bewerten, was da jetzt genau die Annotierung vom Biologen ist.« (B1)

Die Programmierenden problematisieren diese Abhängigkeit vor allem vor dem Hintergrund, dass die Güte ihrer Arbeit (die DLT) von der Güte der Annotierung der Lerndaten abhängt, auf die sie nur bedingt Einfluss nehmen können, da es der spezifischen disziplinären Expertise bei der Aufbereitung bedarf. B1 problematisiert aus dieser Perspektive heraus, dass die Expert:innen bei der Annotierung der Daten auch müde werden und fehleranfällig sein können. Ob dies aber auftritt, kann er kaum kontrollieren. Aus seiner disziplinären Logik heraus entwickelt er deshalb schon Problemlösungen, die den

Einfluss der annotierenden Expert:innen verringern. Hiermit kann er diese existenzielle Projektphase selbst gestalten und die Kontrolle der Daten wieder übernehmen. B1 schildert, dass es verschiedene Möglichkeiten seitens der Programmierenden gibt, die manuelle, sehr aufwändige und zum Teil auch fehleranfällige Annotierung durch die Simulation und Modellierung künstlicher Trainingsdaten zu begegnen. Auch kann mit einer solchen Augmentierung die Lerndatenmenge künstlich erhöht werden. Hierbei werden die bestehenden Lerndaten simuliert und künstlich variiert. Hier kommt Künstliche Intelligenz zum Einsatz, um die händische Annotierung der Lerndaten zu ersetzen. B1 agiert hier nach der für das Feld der Programmierenden spezifischen Logik und macht das Problem zu seinem eigenen Forschungsfeld, um damit symbolisches Kapital in seinem Feld zu gewinnen.

Im Vergleich zu der vorherigen betrachteten Projektphase des Entstehungszusammenhangs zeigt sich in dieser Phase, dass es zu einem Wechsel in Abhängigkeit und Verantwortung kommt. Die Qualität der Datenannotierung ist ein ausschlaggebendes Erfolgskriterium für die DLT. In beiden Projekten wird der Relevanz der Qualität der Annotierung der Lerndaten begegnet und es werden verschiedene Methoden eingesetzt, damit die disziplinären Expert:innen aus den Anwendungsfeldern ihr Wissen in die Aufbereitung der Daten einbinden können. Dabei sind die Programmierenden auf die Güte der Annotierung angewiesen. Es zeigt sich in beiden Projekten, dass die Anwendenden nicht nur in den Entwicklungsprozess eingebunden werden, sondern sie als disziplinäre Expert:innen für diese Projekte existenziell sind. Doch darüber hinaus wird deutlich, welche große Probleme und Herausforderungen in dieser Phase bestehen. Die Logik der DLT fordert Daten ein, die Entscheidungen auf der Grundlage von Korrelationen ermöglichen. Umso mehr Daten vorhanden sind, mit denen die DLT trainiert werden kann, desto besser. Einzelne Fehler, die die Programmierenden in der Qualität nicht beurteilen können, können so durch die Masse kontrolliert werden. Doch beide Fälle zeigen, dass das Produzieren dieser Daten sehr viele Ressourcen in Anspruch nimmt und auch einen hohen Verständigungsaufwand erfordert. Während es im ersten Fall nicht über den Prototypen hinausgeht, entwickelt sich im zweiten Fall die Tendenz, die Expertise des Anwendungsfeldes durch weitere technische Applikationen zu ersetzen.

Dritte Beobachtung: DLT als Black-Box/Übergaben statt Kollaboration

DLT-Entwicklungsprojekte beinhalten weiterhin eine Phase der Programmierung bzw. eine Phase, in der der Algorithmus optimiert und kontrolliert wird. Häufig werden DLT als Blackbox beschrieben (siehe hierzu auch Beitrag Heckwolf/Schmitt in diesem Band). In dieser idealtypischen Projektphase geht es darum, dass die Programmierenden die Güte der DLT testen und diese kontinuierlich verbessern. B1 beschreibt, dass hierbei wichtig sei, »ein bisschen zu tricksen« und dieser Optimierungsprozess nach dem »Trial and Error«-Prinzip verlaufe. Dabei gäbe es durchaus »Patentrezepte«, die zur Verbesserung der Ergebnisse führen können. Dennoch kommt es hierbei »auf jeden Fall ein bisschen [auf] Erfahrung« an und darauf, »das Gespür zu haben, an welchen Stellschrauben« gedreht werden muss. An der hier beschriebenen Praxis der Programmierenden lassen sich die Übersetzungsschwierigkeiten gut erkennen. Sie entwickeln ein hohes Maß an implizitem praktischem Wissen, das sie bei der Anpassung des Learners einbringen. Nicht nur ist die DLT an sich schwer zu durchschauen – wie es technisch durch die *hidden layers* erwartet wird – sondern es ergibt sich auch aus dem praktischen Umgang mit ihm ein Problem der Verständigung über dessen Funktionsweise innerhalb der Projekte. Auch A1 beschreibt die Charakteristiken dieses Optimierungsvorgangs ähnlich:

»Bei mir war es tatsächlich etwas Glück. Ich habe an Parametern rumgespielt, die ich für sinnvoll erachtet habe, wo ich so das Gefühl hatte, ok, hier könnte man was dran schrauben und dann wird es besser.«

Von den Programmierenden selbst wird der Algorithmus als eine Blackbox und infolgedessen als nicht in Gänze kontrollierbar und verstehbar problematisiert. So bedarf es Erfahrung, Wissen und Glück, um die DLT zu optimieren. Dabei steht in den Beschreibungen der Befragten auch nicht das Verständnis der DLT im Fokus, sondern vor allem die Verbesserung der Güte der Ergebnisse.

»Das ist ja sehr, sehr kompliziert, posthoc rausfinden zu wollen, warum der Algorithmus so oder so gelernt hat. Da gibt es gewisse Verfahren, aber die sind so wahnsinnig aufwendig, die interessieren uns momentan auch nicht.«
(A2)

Der Fokus liegt auf dem Funktionieren der DLT und weniger auf dessen Verständnis. Die DLT gilt als funktionsfähig, wenn gewisse Fehlertoleranzen erfüllt werden. Die Programmierenden betonen, dass ihre Tätigkeiten vor allem anwendungsnah orientiert sind. Sie positionieren sich damit gegenüber theoretisch orientierten Informatiker:innen, deren Hauptaugenmerk dann auf einem durchdringenden Verständnis von DLT liege. Dies macht sichtbar, dass es für sie Schwierigkeiten bereitet, die Logiken der Technologie zu verstehen und entsprechend für Akteur:innen aus dem Anwendungsfeld zu übersetzen.

Die Akteur:innen aus dem Anwendungsfeld übergeben die Verantwortung in dieser Phase des Projektes in Gänze den Programmierenden. Die Gründe hierfür können unterschiedlich sein. So könnte es sein, dass sie nur an dem Outcome interessiert sind und den Programmierenden vertrauen oder sie glauben, weiterhin die Kontrolle darüber zu haben, wie die Technologie angewendet werden wird. Dies drückt insbesondere A2 aus, wenn er davon spricht, dass es letztlich die Ärzt:innen seien, die die Diagnose trafen und somit weiterhin die DLT kontrollieren. Im Folgenden »wechselt die Verantwortlichkeit wieder zurück zu den Ärzten in unserer Abteilung, die dann die gefundenen Ergebnisse auf Plausibilität überprüfen« (A2).

Doch die Abgabe der Verantwortung könnte auch auf ein tiefgreifendes Verständigungsproblem hindeuten. Vonseiten der Programmierenden wird die Verantwortungsabgabe seitens der Akteur:innen aus dem Anwendungsfeld kritisch hinterfragt. B1 gibt an, er nehme es so wahr, dass die Biolog:innen den Ergebnissen der Technik viel zu sehr vertrauen. Für B1 ist durch seine Erfahrungen mit der Programmierung von DLT klar, dass die Ergebnisse variieren können und stets auch Fehler auftauchen können. Scheinbar kann er dieses Verständnis den Biolog:innen nicht ausreichend vermitteln.

DLT als Blackbox wahrzunehmen führt in der Folge dazu, dass es zu einer stärkeren Trennung der beiden Handlungsfelder kommt. Die Optimierung der DLT läuft zwischen den Polen der Expertise der Programmierer:innen und deren Glück und Intuition entlang. Durch den Blackbox-Charakter von DLT entstehen spezifische Verständigungs- und Verstehenshürden und damit einhergehend unterschiedliche Deutungen der Akteurskonstellation. Dadurch, dass die DLT für die Beteiligten nicht voll durchsichtig und vermittelbar sind, kommt es zu Diskrepanzen in den Kontroll- und Expert:innenerwartung, die sich zwischen denen der Anwender:innen und der Programmierer:innen unterscheiden.

Die Wechsel in Zuständigkeit und Verantwortung werden von den Befragten dezidiert als solche beschrieben und im Prozess markiert. Dem stehen

Aussagen zur Zusammenarbeit gegenüber, die aufkommen, wenn es um die gesamte Projektkonstellation geht:

»Also es war auch schon so ein interdisziplinäres Projekt, wo man wirklich zusammen arbeiten musste aus zwei verschiedenen Disziplinen. Aus der Medizin, die nicht viel Ahnung von Technik, Programmieren und Deep-Learning an sich haben und ich, der halt mehr von der technischen Seite kommt, aber weniger Ahnung von Pathologien und solchen Sachen hat.«
(A1)

Das wechselseitige Aufeinander-angewiesen-sein und die Wechsel der Zuständigkeit samt der damit einhergehenden disziplinären Expertise erscheinen als charakteristische Elemente solcher anwendungsnahen DLT-Entwicklungsprojekte. Die Prozesse sind stetig von Übergaben der Verantwortung, Expertise und Kontrolle geprägt.

In der Analyse wird deutlich, dass es durch die DLT nicht zu einem weitreichenden und tiefgehenden Austausch über die disziplinären Verständnisse kommt. Es wird zwar kommuniziert, was das Verständnis der Projektpartner:innen ist, aber nur in Hinblick auf die Anforderungen des nächsten Prozessschritts. Es kommt dagegen vor allem zu Übergaben der Verantwortlichkeiten und weniger zu einer Verständigung über die jeweilige disziplinäre Logik. Die beschriebene existenzielle Einbindung der Anwendenden in den Prozess geschieht zwar und ist relevant, sie ist jedoch sehr reduktionistisch.

6. Diskussion

Die Forschung zu feldübergreifenden Entwicklungsprojekten hat gezeigt, dass die Entwicklungsgegenstände als Boundary Objects strukturieren können, welche Verhältnisse sich zwischen den beteiligten Akteur:innen entwickeln. Das betrifft auch das Verhältnis von Entwickler:innen und Anwender:innen. In unserem Beitrag haben wir argumentiert, dass sich DLT als Boundary Objects in Entwicklungsprojekten auf charakteristische Weise von anderen Technologien unterscheiden. Die explorative Analyse der beiden DLT-Entwicklungsprojekte hat drei Wirkweisen der DLT als Boundary Object offengelegt, die die Verhältnisse von DLT-Entwickler:innen und DLT-Anwender:innen auf spezifische Weise strukturieren.

Erstens zeigt unsere Analyse, dass DLT schon für die initiale Konstellation von DLT-Programmierenden und DLT-Anwendenden in einem Projektzusammenhang strukturierend sind. Als Technologie, die die Lösung von Problemen im möglichen Anwendungsfeld verspricht, gibt sie, wie viele andere Technologien auch, Anlass zu Feldgrenzen übergreifenden Entwicklungsprojekten. In der vorangehenden Forschung hat sich bereits gezeigt, dass die Tiefe der auf einen Entwicklungsgegenstand bezogenen Kenntnisse das Verhältnis grundlegend strukturiert (Leonardi/Bailey/Pierce 2019; Levina/Vaast 2005). In den beiden von uns betrachteten Fällen hat sich gezeigt, dass die DLT-Anwendenden sich selbst, trotz der teilweise beschriebenen leichten Zugänglichkeit der Technologie, besonders geringe auf die DLT bezogene Kenntnisse zuschreiben. Infolgedessen befassen sie sich im Entwicklungsprozess selbst nur rudimentär mit der Konstruktion der DLT und verstehen sich eher als Auftraggeber:innen denn als gleichberechtigte Entwicklungspartner:innen, während die Programmierenden als Dienstleister:innen wahrgenommen werden. Derartige Asymmetrien tauchen unabhängig vom Entwicklungsgegenstand auch in anderen Entwicklungszusammenhängen auf. Wir argumentieren aber, dass die DLT als Entwicklungsgegenstand eine Relationierung der Entwicklungspartner:innen als Auftraggeber:innen und Dienstleister:innen (gegenwärtig) aus zwei Gründen nahelegt. Erstens sind DLT im Gegensatz zu anderen Entwicklungstechnologien in vergleichsweise geringem Ausmaß in den (Anwendungs-)Feldern verbreitet und die potenziellen Anwendenden verfügen über vergleichsweise wenig technologiebezogenes Wissen. Zweitens sind DLT dadurch, dass sie selbst für die Programmierenden nur eingeschränkt zu durchschauen sind und in höherem Maße als Blackbox wahrgenommen werden, an sich besonders schwer verständlich. Die DLT-Anwendenden sind also aufgrund ihrer mangelnden DLT-Erfahrung und weil DLT an sich besonders schwer zu verstehen sind, kaum dazu befähigt, sich die Gestaltung der DLT gedanklich zu erschließen. Dass sich die DLT-Anwendenden im Entwicklungsprozess in eine eher passive Auftraggeber:innen-Rolle begeben, erscheint uns daher als typisch für (gegenwärtige) DLT-Entwicklungsprozesse.

Zweitens zeigt unsere Analyse, dass die DLT-Entwicklung spezifische DLT-Anwendungs-Einbindungspraktiken nahelegt und damit das Programmierungs-Anwendungs-Verhältnis grundlegend strukturiert. Um die DLT entwickeln zu können, sind die Entwickler:innen darauf angewiesen, dass die Anwender:innen ihnen adäquate Lerndaten zur Verfügung stellen. In den beiden betrachteten Fällen ist es dazu notwendig, dass die Anwender:innen auf

Grundlage ihrer Expertise Bilder annotieren. Anhand der Annotation identifizieren die DLT anschließend Muster, die es ihnen schließlich erlauben, Bilder nach eben diesen Mustern zu analysieren. Die Aufgabe der DLT-Entwickler:innen besteht im Wesentlichen darin, den Algorithmus für die spezifische Lernaufgabe zu sensibilisieren. Das Wissen der DLT-Anwendenden findet somit vergleichsweise unmittelbar Eingang in den Entwicklungsgegenstand. Damit ist die Art der Anwender:inneneinbindung im DLT-Entwicklungsprozess auf spezifische Weise vorstrukturiert und unterscheidet sich in der Art der Nutzer:inneneinbindung systematisch von anderen Technikentwicklungsprozessen. Einerseits fordert die DLT die Einbindung von Nutzenden: Während die Entwickler:innen von Maschinen oder klassischer Software in vielen Fällen auch ohne die Einbindung von Anwendenden, alleine ausgehend von ihren eigenen Vorstellungen von Anwendungsfällen, operieren, ist die frühe Einbindung von Anwendenden typisch für die DLT-Entwicklung. Andererseits bleibt die Einbindung aber typischerweise oberflächlich und maschinell. In vielen Technikentwicklungsprozessen wird die Einbindung von Anwendenden durch aufwändige Wissenstransferpraktiken realisiert. Entwickler:innen folgen den Anwender:innen dazu mitunter tagelang durch ihren Alltag, um auch ihr implizites Wissen und dessen kontextuelle Einbettung zu erfassen (Nonaka/Takeuchi 1995; Berends et al. 2006). Entwickler:innen arbeiten hier dementsprechend intensiv daran, Anwender:innen zu verstehen und sich ihr Anwender:innenwissen anzueignen. Die Feldgrenzen, die durch die Differenzen im praktischen Wissen der Akteur:innen definiert sind, verschwimmen dadurch im Erfolgsfall temporär (Leonardi/Bailey/Pierce 2019; Carlile 2002). Dagegen sind die DLT-Entwickler:innen lediglich darauf angewiesen, zu verstehen, welche Mittel sie zur Annotation benötigen. Ein Verstehen der Anwender:innen im engeren und tieferen Sinne ist dabei nicht erforderlich, wenn auch der alltägliche Anwendungskontext weitestgehend ausgeblendet bleibt. Zwar steht es auch den DLT-Entwickler:innen prinzipiell offen, sich intensiver mit dem Wissen der Anwender:innen zu befassen. Die DLT legt das aber zumindest nicht nahe und es erscheint uns daher sehr wahrscheinlich, dass es in DLT-Entwicklungsprojekten typischerweise bei der oberflächlichen und maschinenvermittelten Anwender:inneneinbindung bleibt.

Drittens führen die Undurchsichtigkeit der DLT (Blackbox) und die Möglichkeit, das erforderliche Nutzer:innenwissen maschinell zu integrieren, gemeinsam dazu, dass DLT-Entwickler:innen und DLT-Anwender:innen in den betrachteten DLT-Entwicklungsprojekten eher nebeneinander und unabhän-

gig voneinander als gemeinsam arbeiten (Chettiparamb 2007: 16). Auch in Folge der Arbeitsorganisationen bleiben die beiden Felder daher unabhängig und eigenständig. Eine partielle oder temporäre Integration der beteiligten Felder, wie sie durch andere feldübergreifende Entwicklungsprojekte durchaus vorangetrieben wird, wird in den DLT-Entwicklungsprojekten daher kaum gefördert. Die DLT als zentrales Boundary Object trägt in den betrachteten Entwicklungsprojekten also eher dazu bei, dass die Grenzen zwischen den Feldern trotz Kollaboration aufrechterhalten werden können, als dass sie deren Überwindung befördert.

7. Konklusion

Mit der Analyse der Wirkung von DLT als Boundary Object in Entwicklungsprojekten leistet diese Arbeit insbesondere zwei Beiträge. Mit unserer kritischen Analyse der Wirkung von Boundary Objects tragen wir zu einem differenzierten Verständnis des Konzepts bei. Neben der bereits benannten Machtasymmetrie stiftenden Wirkung weist unser Beitrag darauf hin, dass Boundary Objects nicht nur dabei helfen können, Grenzen zu überwinden, sondern es ebenso möglich ist, dass sie das Aufrechterhalten dieser Grenzen trotz Kollaboration fördern. Am Beispiel von DLT haben wir bereits einige Charakteristika von Boundary Objects identifiziert, die eine derartige Wirkung begünstigen. Es scheint aber äußerst fruchtbar in weiterführender Forschung andere Typen von Boundary Objects miteinander zu vergleichen und detaillierter herauszuarbeiten, welche Boundary Objects-Qualitäten ausschlaggebend dafür sind, ob Boundary Objects eher zur Überwindung oder zur Stabilisierung von Grenzen in Kollaborationsprojekten beitragen.

Zweitens adressiert unser Beitrag die Forschung zu KI und DLT. Viele diesbezügliche Beiträge beschäftigen sich mit den möglichen Auswirkungen der Anwendung von KI in unterschiedlichen Bereichen. Die Techniksoziologie hat jedoch immer wieder gezeigt, dass Bilder von Technik und technische Möglichkeiten bis zu ihrer Anwendung im Alltag einen vielschichtigen sozialen Gestaltungsprozess durchlaufen, der maßgeblich beeinflusst, wie eine vorhandene Technologie tatsächlich wirksam wird. Es erscheint daher sinnvoll, sich diesem Gestaltungsprozess selbst stärker zuzuwenden, um die rezenten Entwicklungen erschließen zu können. Wir haben dieses Ziel in diesem Beitrag mit Blick auf DLT-Entwicklungsprojekte verfolgt. Unsere Befunde weisen darauf hin, dass DLT-Entwicklungsprojekte Anwender:innen und Entwick-

ler:innen in ein spezifisches Verhältnis zueinander bringen. Unsere Daten belegen, dass die beiden Felder im Vergleich zu anderen Entwicklungs- und Kollaborationsprojekten in hohem Maße unabhängig voneinander bleiben und ein wechselseitiges Verstehen nur an der Oberfläche stattfindet. Mit Blick auf die bisherige Forschung zu Akteurskonstellationen und den Ergebnissen von Entwicklungsprojekten erscheinen daran anschließend zwei Szenarien plausibel. Erstens könnte die mangelhafte Anwender:innen-Integration dazu führen, dass die ›inselbegabten‹ Anwendungen, die aus den Projekten hervorgehen, nicht in ausreichendem Maße mit den alltäglichen Anforderungen der Anwender:innen kompatibel sind und daher oftmals nicht dort etabliert werden können. Zweitens führt die Unabhängigkeit der Partner:innen im Entwicklungsprozess dazu, dass die Entwickler:innen die Gestaltung der Anwendungen dominieren. Die Forschung zur Entwicklung neuer Produkte zeigt, dass Entwickler:innen dazu neigen, unhinterfragte Annahmen aus ihrer eigenen sozialen Lebenswelt in die Produkte einzuschreiben. Indem Anwender:innen diesen Regeln entsprechen müssen, wenn sie die Produkte nutzen wollen, werden die Logiken des Entwicklungsfeldes durch die Technik im Anwendungsfeld gewissermaßen reproduziert. Die Eigenständigkeit der DLT-Entwickler:innen führt dazu, dass sie die Logik ihres Feldes weitestgehend ungefiltert in die DLT-Anwendungen einschreiben können. Wenn sich die entsprechenden DLT im Anwendungsfeld etablieren, ist dementsprechend eine Angleichung an das Feld der Programmierer:innen zu erwarten. Weiterführende Forschung, die die konkreten Konsequenzen des Verhältnisses von Anwender:innen und Entwickler:innen für die Ausgestaltung der entwickelten Anwendungen empirisch untersucht, erscheint daher äußerst relevant.

An dieser Stelle nun lässt sich die Perspektive Whites sinnvoll einbinden, indem sie eine andere Perspektive auf das Phänomen ermöglicht. Die hier angewendete feldtheoretische Perspektive folgt dem Impuls, zu klären, warum verinnerlichte Prägungen, die in sozialen Feldern von Akteur:innen gewonnen wurden, ihre Persistenz in divergierenden Kontexten erhalten. Der oder die Akteur:in erhält damit eine relativ stabile sozialräumliche Verortung und wird als zugehörig zu einem bestimmten Feld klassifiziert (Hirschauer 2017: 30ff.). Die Beobachtung von Diskontinuitäten zwischen Feldern ist damit in gewisser Hinsicht vorprogrammiert.⁷

7 Diese eher klassische Lesart der bourdieuschen Feldtheorie wurde zuletzt durch Schmitz et al. (2017) in Frage gestellt. Sie verweisen, um die Grenzen von Feldern zu bestimmen, darauf, sich die Interdependenzen zwischen Feldern genauer anzusehen.

Im Anschluss an White kann allerdings grundsätzlich davon ausgegangen werden, dass Akteur:innen stets in unterschiedlichen Netzwerken eingebettet sind. Dementsprechend stellt sich hier eher die Frage, ob und wie Akteur:innen Kompetenzen darin erlangen, in unterschiedlichen Beziehungen mit heterogenen Handlungsregeln erfolgreich handeln zu können. Der Wechsel zwischen den Kontexten wird bei Whites Einführung des *netdom* Begriffes deutlich: »Identities switch from netdom to netdom, finding footings in different networks in differing domain contexts« (White 2008: 7)⁸. Resultat und Ziel dieser Theorie ist keine saubere und geordnete Verortung des Sozialraums, wie sie die Theorie der sozialen Felder in metaphorischer Anlehnung an die physikalischen Kraftlinien relativ dazu leistet, denn sie geht von einer grundsätzlichen Unordnung der Welt aus (vgl. White 2008: 18). Dies bedeutet in Hinblick auf den zugrunde liegenden Forschungsgegenstand der Zusammenarbeit in Entwicklungsprojekten mit DLT, dass es mittels der Perspektive Whites interessant sein könnte, die tatsächlichen Kommunikationsprozesse zwischen den Partner:innen beobachtend zu folgen und nicht nur die Darstellung in den Interviews zu interpretieren, um genauer herauszuarbeiten, ob sich in der Theoriesprache Whites durch die Events und spezifischen Netzwerksituationen der Entwicklungsprojekte Identitäten und Netzwerkdomänen ausbilden, die Kontrolle ausüben. Es müsste darum gehen, die Domänen herauszuarbeiten, die Mische/White (1998: 702) wie folgt definieren:

So erst werde die relative Autonomie oder Heteronomie eines Feldes zwischen anderen Feldern sichtbar, sodass sich eine stabile sozialräumliche Verortung auch erst in Feldern mit hoher Autonomie ergebe.

- 8 Die Art des Wechsels bzw. Überlagerung und was White dabei unter Identitäten versteht, definieren Mohr/White (2008: 489) folgendermaßen: »White (1992, 2008) provides extensive theoretical specification that begins not with individual embodied persons, but rather with identities that can be generalized to include any source of action. A firm, a community, a crowd, oneself on the tennis court, encounters of strangers on a sidewalk, each may be identities. Identities are triggered out of events—which is to say, relational situations that emerge from the flow of social events, switching from one network situation into another, that produce a need for control over uncertainty and thus a control over situated actions and fellow identities. Lives are organized around this flowing movement from one network situation into another and these situations are themselves phenomenological sites that include certain other identities (in particular networks) organized around specific domains of activity (hence they are network/domains or simply netdoms).«

»We define domain as the perceived array of such signals – including story sets, symbols, idioms, registers, grammatical paternings, and accompanying soporeal markers – that characterize a particular specialized field of interaction.«

Innerhalb der Kommunikationsprozesse wäre dann nachzuvollziehen, ob es zu ›Situationen‹ kommt also »those problematic, high-stakes episodes that cast our prescribed roles and trajectories into question [...] each involves a previously predictable, stylized interaction that suddenly becomes fraught with uncertainty, danger, and/or opportunity« (ebd.: 698). Es ginge darum, die tatsächlichen Vernetzungen nachzuvollziehen, den Wechseln der Identitäten zwischen den jeweiligen Kontexten zu folgen und die Geschichten innerhalb des Netzwerks herauszuarbeiten:

»Active participation (control efforts) in these situations depends upon both the network logic and the shared conceptual understanding of what is going on in the netdom. The shared understanding (localized meaning) is produced through the interactions (and especially the switching back and forth across these interactional sites) and embodied in a shared repertoire of netdom stories that provide the basis for a common understanding of what is going on in the moment.« (Mohr/White 2008: 489)

Dies ermöglicht es weiter zu hinterfragen, inwiefern es in den spezifischen Entwicklungsprojekten gelingt, ein gemeinsames Verständnis aufzubauen oder ob letztlich weiterhin Wege gesucht werden, den Verständigungsprozess zu umgehen und den Herausforderungen, die ein solcher bereitstellt, auszuweichen.

8. Literatur

Amershi, Saleema/Begel, Andrew/Bird, Christian/DeLine, Robert/Gall, Harald/Kamar, Ece/Nagappan, Nachiappan/Nushi, Besmira/Zimmermann, Thomas (2019): »Software Engineering for Machine Learning: A Case Study«, in: 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP), IEEE, S. 291–300.

- Aust, Holger (2021): »KI: Hype oder Technologie der Zukunft?«, in: Holger Aust (Hg.), *Das Zeitalter der Daten*, Berlin/Heidelberg: Springer Berlin Heidelberg, S. 21–32.
- Barley, Stephen R./Tolbert, Pamela S. (1997): »Institutionalization and Structuration: Studying the Links between Action and Institution«, in: *Organization Studies* 18, S. 93–117.
- Barrett, Michael/Oborn, Eivor (2010): »Boundary object use in cross-cultural software development teams«, in: *Human Relations* 63, S. 1199–1221.
- Battilana, Julie (2006): »Agency and Institutions. The Enabling Role of Individuals' Social Position«, in: *Organization* 13, S. 653–676.
- Bechky, Beth A. (2003): »Sharing Meaning Across Occupational Communities. The Transformation of Understanding on a Production Floor«, in: *Organization Science* 14, S. 312–330.
- Berends, Hans/Bij, Hans/Debackere, Koenraad/Weggeman, Mathieu (2006): »Knowledge sharing mechanisms in industrial research«, in: *R and D Management* 36, S. 85–95.
- Berends, Hans/Sydow, Jörg (2019): »Introduction: Process Views on Inter-organizational Collaborations«, in: Hans Berends/Jörg Sydow (Hg.), *Managing Inter-organizational Collaborations: Process Views*, Bingley: Emerald Publishing Limited, S. 1–10.
- Berger, Peter L./Luckmann, Thomas (1980): *Die gesellschaftliche Konstruktion der Wirklichkeit. Eine Theorie der Wissenssoziologie*, Frankfurt a.M.: Fischer Taschenbuch.
- Bjerregaard, Toke (2010): »Industry and academia in convergence: Micro-institutional dimensions of R&D collaboration«, in: *Technovation* 30, S. 100–108.
- Bødker, Susanne (1998): »Understanding Representation in Design«, in: *Human-Computer Interaction* 13, S. 107–125.
- Bødker, Susanne (2000): »Scenarios in user-centred design-setting the stage for reflection and action«, in: *Interacting with Computers* 13, S. 61–75.
- Bogers, Marcel/Zobel, Ann-Kristin/Afuah, Allan/Almirall, Esteve/Brunswick, Sabine/Dahlander, Linus/Frederiksen, Lars/Gawer, Annabelle/Gruber, Marc/Haefliger, Stefan/Hagedoorn, John/Hilgers, Dennis/Laursen, Keld/Magnusson, Mats G./Majchrzak, Ann/McCarthy, Ian P./Moeslein, Kathrin M./Nambisan, Satish/Piller, Frank T./Radziwon, Agnieszka/Rossi-Lamastra, Cristina/Sims, Jonathan/Ter Wal, Anne L. J. (2017): »The open innovation research landscape: established perspectives

- and emerging themes across different levels of analysis«, in: *Industry and Innovation* 24, S. 8–40.
- Borchardt, Andreas/Göthlich, Stephan E. (2007): »Erkenntnisgewinnung durch Fallstudien«, in: Sönke Albers/Daniel Klapper/Udo Konradt/Achim Walter/Joachim Wolf (Hg.), *Methodik der empirischen Forschung*, Wiesbaden: Gabler, S. 33–48.
- Bourdieu, Pierre (1993): *Sozialer Sinn. Kritik der theoretischen Vernunft* (= Suhrkamp-Taschenbuch Wissenschaft, Band 1066), Frankfurt a.M.: Suhrkamp.
- Bourdieu, Pierre (2009): *Entwurf einer Theorie der Praxis auf der ethnologischen Grundlage der kabyllischen Gesellschaft* (= Suhrkamp-Taschenbuch Wissenschaft, Band 291), Frankfurt a.M.: Suhrkamp.
- Bourdieu, Pierre/Wacquant, Loïc J. D. (2006): *Reflexive Anthropologie* (= Suhrkamp Taschenbuch Wissenschaft, Band 1793), Frankfurt a.M.: Suhrkamp.
- Buchkremer, Rüdiger/Witte, Bart de/Matusiewicz, David (2020): »KI in Gesundheit und Medizin«, in: Rüdiger Buchkremer/Thomas Heupel/Oliver Koch (Hg.), *Künstliche Intelligenz in Wirtschaft & Gesellschaft*, Wiesbaden: Springer Fachmedien Wiesbaden, S. 385–395.
- Carlile, Paul R. (2002): »A Pragmatic View of Knowledge and Boundaries: Boundary Objects in New Product Development«, in: *Organization Science*, S. 442–455.
- Chettiparamb, Angélique (2007): *Interdisciplinarity: a literature review*, Southampton: University of Southampton.
- Degele, Nina (1996): »Die Entwicklung und Nutzung von Software. Zur Genese informationstechnischen Handelns«, in: *Zeitschrift für Soziologie* 25, S. 58–70.
- Fligstein, Neil/McAdam, Doug (2011): »Toward a General Theory of Strategic Action Fields«, in: *Sociological Theory* 29, S. 1–26.
- Friedland, Roger/Alford, Robert R. (1991): »Bringing Society Back In: Symbols, Practices, and Institutional Contradictions«, in: Walter W. Powell/Paul DiMaggio (Hg.), *The New institutionalism in organizational analysis*, Chicago: University of Chicago Press, S. 232–266.
- Gassol, Jeanine H. (2007): »The effect of university culture and stakeholders' perceptions on university–business linking activities«, in: *The Journal of Technology Transfer* 32, S. 489–507.
- Giddens, Anthony (1997): *Die Konstitution der Gesellschaft. Grundzüge einer Theorie der Strukturierung*, Frankfurt/New York: Campus.

- Gläser, Jochen (2006): *Wissenschaftliche Produktionsgemeinschaften. Die soziale Ordnung der Forschung* (= Campus Forschung, Bd. 906), Frankfurt a.M./New York: Campus-Verlag.
- Hahn, Katrin (2013): *Heterogene Akteure als Innovationspartner. Zur Strukturierung von Handeln in industriellen Innovationsprojekten*, Wiesbaden: Springer VS.
- Hallett, Tim (2003): »Symbolic Power and Organizational Culture«, in: *Sociological Theory* 21, S. 128–149.
- Hawkins, Beverley/Pye, Annie/Correia, Fernando (2017): »Boundary objects, power, and learning: The matter of developing sustainable practice in organizations«, in: *Management Learning* 48, S. 292–310.
- Heidenreich, Martin/Mattes, Jannika (2017): »Kollaborative Innovationen. Die innerbetriebliche Nutzung externer Wissensbestände in vernetzten Entwicklungsprozessen«, in: Martin Heidenreich/Jürgen Kädtler/Jannika Mattes (Hg.), *Kollaborative Innovationen. Die innerbetriebliche Nutzung externer Wissensbestände in vernetzten Entwicklungsprozessen*, Göttingen: Göttingen University Press, S. 13–43.
- Hippel, Eric von (1986): »Lead Users: A Source of Novel Product Concepts«, in: *Management Science* 32, S. 791–805.
- Hirschauer, Stefan (Hg.) (2017): *Un/doing Differences. Praktiken der Human-differenzierung*, Weilerswist: Velbrück Wissenschaft.
- Hosny, Ahmed/Parmar, Chintan/Quackenbush, John/Schwartz, Lawrence H./Aerts, Hugo J. W. L. (2018): »Artificial intelligence in radiology«, in: *Nature reviews. Cancer* 18, S. 500–510.
- Ibert, Oliver/Müller, Felix C./Stein, Axel (2014): *Produktive Differenzen. Eine dynamische Netzwerkanalyse von Innovationsprozessen* (= Science Studies), Bielefeld: transcript.
- Knorr Cetina, Karin (2002): *Wissenskulturen. Ein Vergleich naturwissenschaftlicher Wissensformen* (= Suhrkamp-Taschenbuch Wissenschaft, Band 1594), Frankfurt a.M.: Suhrkamp.
- Krais, Beate (2004): »Habit und soziale Praxis«, in: Magareta Steiner (Hg.), *Pierre Bourdieu, Politisches Forschen, Denken und Eingreifen*, Hamburg: VSA, S. 91–106.
- Kuckartz, Udo (2018): *Qualitative Inhaltsanalyse. Methoden, Praxis, Computerunterstützung* (= Grundlagentexte Methoden), Weinheim: Beltz.
- Lamnek, Siegfried (2008): *Qualitative Sozialforschung. Lehrbuch* (= Beltz PVU Lehrbuch), Weinheim, Basel: Beltz PVU.

- Leonardi, Paul M./Bailey, Diane E./Pierce, Casey S. (2019): »The Coevolution of Objects and Boundaries over Time: Materiality, Affordances, and Boundary Saliency«, in: *Information Systems Research* 30, S. 665–686.
- Levina, Natalia/Vaast, Emmanuelle (2005): »The Emergence of Boundary Spanning Competence in Practice: Implications for Implementation and Use of Information Systems«, in: *MIS Quarterly* 29, S. 335.
- Levina, Natalia/Vaast, Emmanuelle (2008): »Innovating or doing as Told? Status Differences and Overlapping Boundaries in Offshore Collaboration«, in: *MIS Quarterly* 32, S. 307.
- Lizardo, Omar (2004): »The Cognitive Origins of Bourdieu's Habitus«, in: *Journal for the Theory of Social Behaviour* 34, S. 375–401.
- Majchrzak, Ann/Jarvenpaa, Sirkka L./Bagherzadeh, Mehdi (2015): »A Review of Interorganizational Collaboration Dynamics«, in: *Journal of Management* 41, S. 1338–1360.
- Martin, John L. (2003): »What Is Field Theory?«, in: *American Journal of Sociology* 109, S. 1–49.
- Mische, Ann/White, Harrison C. (1998): »Between Conversation and Situation: Public Switching Dynamics across Network Domains«, in: *Social Research* 65, S. 695–724.
- Mohr, John W./White, Harrison C. (2008): »How to model an institution«, in: *Theory and Society* 37, S. 485–512.
- Nonaka, Ikujiro/Takeuchi, Hirotaka (1995): *The knowledge-creating company*, New York/Oxford: Oxford University Press.
- Razzak, Muhammad I./Naz, Saeeda/Zaib, Ahmad (2018): »Deep Learning for Medical Image Processing: Overview, Challenges and the Future«, in: Nilanjan Dey/Amira S. Ashour/Surekha Borra (Hg.), *Classification in BioApps*, Cham: Springer International Publishing, S. 323–350.
- Roth, Philip (2019): »Die feldspezifische Eigenlogik der praktischen Konstitution sozialer Netzwerke. Konzeptionelle Überlegungen und beispielhafte empirische Befunde«, in: Jan Fuhse/Karoline Krenn (Hg.), *Netzwerke in gesellschaftlichen Feldern*, Wiesbaden: Springer Fachmedien Wiesbaden, S. 171–198.
- Roth, Philip/Diefenbach, Nadine (2021): »The Constitution of Boundaries: How the Embeddedness of Organizational Users Structures the Transfer of their Knowledge«, in: *NOvation – Critical Studies of Innovation* 3, S. 134–162.
- Schmitt, Marco (2019): »Felder und Netzwerkdomänen in der Wissenschaft«, in: Jan Fuhse/Karoline Krenn (Hg.), *Netzwerke in gesellschaftlichen Feldern*, Wiesbaden: Springer Fachmedien Wiesbaden, S. 63–79.

- Schmitz, Andreas/Witte, Daniel/Gengnagel, Vincent (2017): »Pluralizing field analysis: Toward a relational understanding of the field of power«, in: *Social Science Information* 56, S. 49–73.
- Serre, Thomas (2019): »Deep Learning: The Good, the Bad, and the Ugly«, in: *Annual review of vision science* 5, S. 399–426.
- Star, Susan L./Griesemer, James R. (1989): »Institutional ecology, translations and boundary objects: Amateurs and professionals in Berkeley’s Museum of Vertebrate Zoology, 1907–39«, in: *Social Studies of Science* 19, S. 387–420.
- White, Harrison C. (2008): *Identity and Control. How Social Formations Emerge*, Princeton: Princeton University Press.
- Yin, Robert K. (2003): *Case study research. Design and methods* (= Applied social research methods series, Band 5), Thousand Oaks, California: Sage.

Über Identitäten und Selbstverständnis medizinischen Personals in Zeiten künstlicher Intelligenz und Algorithmierung

Tabea Bongert, Dhenya Schwarz

1. Einleitung

Bei einem Blick in die Medien sind solche und ähnliche Schlagzeilen zum Thema KI und Medizin keine Seltenheit: »Wenn der Computer den Arzt ersetzt: Darf der ›Todes-Algorithmus‹ über Leben entscheiden?« (Pfeffer/Matthias 2021). Diese und viele andere Beiträge können medial verfolgt und gelesen werden. Die Missachtung des Datenschutzes – der *gläserne Patient* – oder gar ein kompletter Austausch der Ärzt:innen sowie die Behandlung durch Maschinen allein sind letztendlich verbreitete dystopisch anmutende Vorstellungen (vgl. Ekkerkamp 2019; Maté 2020). Werden Behandlungsmethoden und Entscheidungsprozesse der KI unterliegen? Wie wird sich das Verhältnis zwischen Patient:innen und Ärzt:innen verändern, oder wird es sich gar auflösen?

Fokus dieses Beitrags ist die Untersuchung der Rolle der Ärzt:innen – genauer gesagt, deren Identität bzw. Selbstverständnis als Ärzt:in sowie ihr Verhältnis zu Patient:innen – und der Veränderung durch den Einsatz intelligenter Algorithmen im medizinischen/klinischen Arbeitsalltag. Die Untersuchung fokussiert auf die Perspektive derjenigen, die die Technik einsetzen. Für einen vollständigen Blick auf die Veränderungen im Medizinalltag müssten in weiteren Analysen selbstverständlich auch Perspektiven von Pflegepersonal, Patient:innen und Verwaltung ergänzt werden.

Die Medizinsoziologie untersucht seit den 1950er Jahren als Bindestrich-Soziologie das Gesundheitswesen und deren Teilhaber:innen. Im Zentrum der Forschungsfragen standen und stehen das Gesundheits- und Krankheitsverhalten, die Medizin, ihre Organisation und die darin beschäftigten

Berufsgruppen (Mohr 2000: 188). Eng verknüpft arbeiten die Nachbardisziplinen (Sozialmedizin, medizinische Anthropologie, medizinische Psychologie und die multidisziplinären Gesundheitswissenschaften) an ähnlichen Fragestellungen, dennoch mit anderen Betrachtungsweisen (Mohr 2000: 4).

Die Ärzt:in-Patient:in-Beziehung steht seit jeher im Zentrum des Interesses, da sich dort der zentrale Schnittpunkt von Leistungserbringung und Leistungskonsumption befindet. Bereits Talcott Parsons erforschte 1951 die Rolle der Ärzt:innen und die Komplementärrolle der Patient:innen. Das Rollenkonzept von Parsons wurde erst in den 1970er Jahren in Deutschland in Deutsch veröffentlicht (Bauch 2000: 193). Parsons entwickelte so genannte »pattern variables«, die allgemeine Handlungsorientierungen für das Individuum vorgeben (Mohr 2000: 193). Die sozialen Situationen sind allerdings nicht per se bestimmt, sondern ergeben sich durch Interpretations- und Definitionsleistungen der sozialen Akteur:innen. Erst dadurch wird Eindeutigkeit erlangt (Mohr 2000: 195). Im Bezug zur Rolle des/der Ärzt:in trägt sie idealtypisch vier charakteristische Merkmale. Zum einen ist die Ärzt:innenrolle fachlich kompetent und funktional spezifisch. Das bedeutet, dass die Rolle in modernen Gesellschaften durch Erwerb und Nachweis fachlicher Kompetenzen zugeteilt wird. Zum anderen ist sie universalistisch, der/die Ärzt:in muss also jeder Person Hilfe gewähren. Zudem ist sie emotional neutral, sie soll nicht durch persönliche Vorlieben geleitet sein. Empathie und Anteilnahme gehören zwar dazu, dürfen aber nicht zu einem Übermannen der Gefühle führen. Das letzte Merkmal wird von Parsons als »kollektivitätsorientiert« beschrieben. Die Hilflosigkeit der Person soll weder menschlich noch finanziell ausgenutzt werden. Diese Merkmale werden im beruflichen Sozialisationsprozess erworben (Mohr 2000: 193ff.).

Diese Rollenkonzepte, vor allem unter dem Gesichtspunkt der Reputation, wurden in vielfacher Weise weiter fortgeführt. Beispielsweise wurde das Geschlecht in der Wahrnehmung der Leistung untersucht sowie analysiert, was die Weiterempfehlungsbereitschaft von Patient:innen beeinflusst (vgl. Stiftung Gesundheit 2018: 5; Stiftung Gesundheit 2019: 6; Sieverding/Kendel 2012). Auch der Wandel der ärztlichen Rolle wird immer mehr erforscht. Siegrist beschreibt in seinem Artikel, dass es keinen Berufsstand gibt, dessen »Prozess der Professionalisierung während der vergangenen 150 Jahre so erfolgreich verlaufen ist wie derjenige der Ärzteschaft« (Siegrist 2012: 1100). Dies vergegenwärtigt allerdings auch die neuen Herausforderungen an die Ärzt:innenrolle, die vor allem durch die Begrenzung der Autonomie durch ökonomische Zwänge sowie durch Konsumverhalten und interprofessionelle

– sowohl durch unterschiedliche als auch durch dieselben Fachrichtungen – Konkurrenz gesetzt wird. Er bezieht sich vor allem auf das Rollenverständnis Parsons'. Auch der Einfluss der KI – teils mit Fokus auf die Veränderung der Ärzt:innenrolle – spielt eine immer größere Rolle in Artikeln und Fachtagungen, sie wird aber noch unzureichend empirisch erforscht. Auf der Fachtagung der Kassenärztlichen Bundesvereinigung wurde beispielsweise über den Wandel der Medizin sowie des Arztberufs durch die KI referiert und auch das *Ärzteblatt* beschäftigte sich mit diesem Thema und attestiert, dass der Ärzt:innenberuf zwar nicht ersetzt, aber dafür einen Teil seines Wissensmonopols verlieren wird (Gassen 2019; Schmedt 2018).

Dieser Text möchte dazu beitragen, den noch unzureichend erforschten Ansatz des Einsatzes der KI im medizinischen Arbeitskontext und die Meinung der Ärzt:innen über ihre eigene Rolle in den Mittelpunkt der Forschung zu stellen, um weitere Weichen in diesem Bereich zu stellen. Zunächst bot Andrew Abbott mit seinem Konzept der Profession eine geeignete Orientierung, die aber noch aufgrund der Fragestellung um den Einbezug des Technikhaften, der KI, ergänzt werden musste. Deswegen ist die Theorie sowie deren Begrifflichkeiten von dem Netzwerktheoretiker White eine geeignete Wahl, da er die Begrifflichkeiten der Profession von Abbott in seinem Institutionskonstrukt verwendet. Für eine grundständige Einordnung der KI in die Relationale Soziologie findet sich in diesem Sammelband eine theoretisch einleitende Fundierung, (siehe Häußling/Schmitt in diesem Band). White ist deshalb für diese Untersuchung besonders geeignet, da er sich konsequent an der empirischen Netzwerkforschung ausrichtet und, anstatt eine ontologisch-richtige Abbildung sozialer Zusammenhänge vorzulegen, eher an analytisch fruchtbaren Konzepten für eine Außensicht auf das Soziale interessiert ist. Durch seine skalenfrei nutzbaren Begrifflichkeiten, zum Beispiel der der Identität, ist es möglich, diese ebenfalls auf technische Aspekte anzuwenden. Auch wenn White materielle Dinge nicht explizit in seinen Forschungen als Identität einbezieht, so ist ein Einsatz durch eben diese Skalenfreiheit sowie abstrakten Begriffsdefinition möglich (Häußling 2020: 54). Zugleich sind auch die Relationen zwischen Menschlichem und Nicht-Menschlichem elementar angelegt, allerdings nicht weiter konzeptionell ausgearbeitet (Häußling 2020: 52). Die Theorie bietet somit viel Potenzial für weitere Forschung mit Technikbezug. Es ist ein Instrumentarium an nützlichen Konzepten für die Empirie, in der sich seine Begrifflichkeit sowie theoretischen Erwartungen unabhängig voneinander verwenden lassen (Schmitt/Fuhse 2015: 4f.).

Denn Entwicklungen entsprechend der medial verbreiteten Sorgen würden gravierende Veränderungen für diese mit sich bringen. Die *Profession* der Medizin ist ausschlaggebend für das Selbstbewusstsein der Akteur:innen. Zusammen mit dem *Stil* der Profession (nach Abbott), der *Institution* (nach White) und den verschiedenen *Netdoms* (Netzwerke mit den dazugehörigen kulturellen Formen nach White) können geeignete Marker gesetzt werden, die disruptive Veränderungen abtasten können. Warum die gesellschaftliche Debatte sehr empfindlich auf die KI in diesem Bereich reagiert und inwiefern die Ärzt:innenschaft von ihrer Rollenveränderung durch die KI ausgeht, wurde in vier Interviews mit Ärzt:innen in Forschung und Praxis sowie einem Forschenden aus dem biomedizinischen Ingenieurwesen erhoben und analysiert.

In einem ersten Schritt wird in Kapitel 2 das theoretische Instrumentarium vorgestellt, damit das wechselhafte Spannungsverhältnis zwischen Wandel und Stabilität in der *Institution Medizin* deutlich wird. In Kapitel 3 werden anhand dieses Werkzeugkastens die Interviews implementiert und ausgewertet sowie auf ihre Hauptaussagen für die Fragestellung reduziert. In Kapitel 4 werden schlussendlich die Ergebnisse der Analyse problematisiert, um abschließend Schlüsse zu ziehen und einen Ausblick zu schaffen.

2. Die Institution Medizin im Spannungsfeld zwischen Stabilisatoren und Wandlungspotenzial

Die *Institution Medizin* ist seit jeher ein Ort ständiger Innovationen (Interview Staatjes/Regli: Z. 329f.). Die Institution steht für ein Netzwerk, welches durch ein kulturelles Modell – durch die Rhetorik – geprägt wird. Die Institution begleitet soziale Prozesse, indem sie sie kanalisiert, also in eine bestimmte Richtung lenkt. Doch trotz dieses innovativen Ortes und somit des potenziellen Wandels, existieren in einer institutionellen Konstellation verschiedene Arten von meist dauerhaft angelegten Netzwerken und *Stilen* von zuständigen Personen – und eben auch der *Stil* der Ärzt:innen – zu arbeiten (White 2008: 171). Die immer wiederkehrenden Gespräche, Praktiken und Geschichten führen zur Selbsterhaltung der Institution. Der institutionelle Wandel steht den dauerhaften, nebeneinander existierenden *Stilen* in dieser Institution gegenüber (Mohr/White 2008: 1). Damit steht die Institution der Medizin im Spannungsfeld zwischen Selbsterhaltung und Weiterentwicklung. Sie ist wie ein Bindeglied zwischen den sozialen Aggregationsleveln – sowohl der Mikroebene mit sozialen Interaktionen als auch der Mesoebene der Organisation. Weiterhin

werden durch sie Brücken zwischen dem Symbolischen und dem Materiellen sowie zwischen Struktur und Agency geschlagen (Mohr/White 2008: 1). Die Institution Medizin ist somit nicht nur durch das Spannungsfeld charakterisiert, viel eher ist es ein immanenter Bestandteil. Die Stabilisation erfolgt durch vorläufig gefestigte Verhältnisse der *Kontrollkämpfe* – ein Versuch der Reduktion von Unsicherheit der Umwelt. Diese Kontrollkämpfe entspringen jedem Individuum und halten es zudem an ihrem Platz. Wie kann aber die Institution der Medizin so stabil sein, wenn die Kontrollkämpfe durch die erwünschten und auch selbst hervorgebrachten Innovationen stetig im Wandel sind? Abbott hat sich in seinem Buch *The System of Professions* intensiv mit der Frage beschäftigt, wie sich in modernen Gesellschaften Expertise institutionalisiert und diese Fragestellung vor allem vor dem Hintergrund der Medizin beleuchtet. Die Institutionalisierung der Profession der Ärzt:innen steht somit nicht nur im engen Zusammenhang mit der Gründung der Institution Medizin, sondern auch mit der Rückwirkung dieser auf dessen Eigenverständnis. Es entsteht eine Art Wechselwirkung. Er untersuchte vor allem die Bedingungen und Kontexte der Kontrolle der Arbeit, da daraus der Zuständigkeitsanspruch der jeweiligen Profession erwächst. Abbotts Kontrollbegriff, der Fokus auf Wettbewerb (Kontrollkämpfe) sowie die Ausarbeitung der in den einzelnen *Netdoms* liegenden Zuständigkeitsansprüche (Diagnose, Behandlung, Schlussfolgerung, akademische Arbeit), die neben der sozialen Struktur die kulturelle Maschinerie der Zuständigkeit ausmacht, werden bereits in Whites erstem Buch *Identity and Control* von 1992 aufgegriffen, woraus sich eine geeignete Synergie mit seinen Begrifflichkeiten ergibt (Abbott 1988: Preface; Abbott 1988: 66; vgl. White 1992: 223).

Die Institution bringt neben dem Spannungsverhältnis ein Set an Normen und Werten als Orientierungswissen mit und ist interdependent mit dem/den *Stil/en* der Institution verknüpft, in der die Stile sogar erst durch die Institution hervorgebracht werden (Mohr/White 2008: 7). Der *Stil* der Institution entsteht durch gängige Praktiken und dessen Reproduktion im situativen Kontext (White 2008: 112ff.). Diese gängigen Praktiken können im Krankenhausalltag die Blutabnahmen darstellen, die stets zu Beginn durchgeführt werden, um weitere Maßnahmen abzuklären. Nach der weiteren Überweisung in einen Fachbereich wird die weitere Behandlung durch den dort vorherrschenden Stil weiter spezifiziert (Beispiel der sich stets wiederholenden Praktik der Magenspiegelung bei Gastroenterolog:innen). Er wird durch die Institution und ihr Norm-Set beeinflusst und begrenzt, bestärkt die Institution jedoch zugleich durch die Reproduktion (Mohr/White 2008: 8). Die Profession als Stil

mit ihrem Zuständigkeitsanspruch in bestimmten *Netdoms* findet sich ebenfalls in der Institution wieder. Allerdings besitzen auch andere Institutionen Zuständigkeitsansprüche, die einen gewissen Druck ausüben können. Dies ist bei Abbott einmal das Rechtssystem, das eine formale Kontrolle der Arbeit ermöglichen kann und die öffentliche Meinung, in der die Berufung, also die mit Leidenschaft und aus Überzeugung ausgeführte Tätigkeit, der Medizin Bilder aufbaut und verbreitet. Die in der öffentlichen Meinung erstellten Bilder der Berufung können die verschiedenen Bereiche (Bsp. Rechtssystem) unter Druck setzen. Der Bereich des Arbeitsplatzes ist laut Abbott noch zu unzureichend erforscht, da gerade der Arbeitsplatz die offiziellen Grenzen zwischen rechtlich und öffentlich etablierten Zuständigkeiten verzerren kann. Somit wird auch diese teils paradoxe Problematik hier adressiert und eine Versöhnung der öffentlichen Ärzt:innenposition und den Positionen im Arbeitskontext angestrebt (Abbott 1988: 66).

Der Zuständigkeitsanspruch muss sowohl artikuliert als auch anerkannt werden. Demnach liegt hier auch das Wandlungspotenzial der Institution, da nur durch die Aberkennung oder Abgabe der Zuständigkeit der Kern der Institution ins Wanken geraten kann (Mohr/White 2008: 16f.; Schmitt/Fuhse 2015: 122). Dem gegenüberliegend finden sich die *Netdoms* (Domäne und Netzwerk) wieder (zum Beispiel charakterisiert durch Narrative für Sozialbeziehungen, Sprachmuster, Interaktionsregeln oder Bewertungen), welche die kulturellen Formen zwischen und in den verschiedenen sozialen und organisationalen Netzwerken widerspiegeln (Schmitt/Fuhse 2015: 109). Die Domäne eines Netzwerkes folgt aber niemals nur allgemeinen kulturellen Mustern. Dies ist eher ein Merkmal der Institution. Die *Netdoms* greifen zwar auf diese Muster zurück, entwickeln jedoch auf dieser Grundlage immer spezifische Sinnformen für diesen einzelnen sozialen Kontext (Schmitt/Fuhse 2015: 109). Dies bedeutet, dass die *Netdoms* auf die Regeln der Institution Medizin (Prävention, Diagnose, Behandlung von Krankheiten) zurückgreifen, aber die verschiedenen Fachbereiche als *Netdoms* nochmals zusätzliche Sinnformen entwickelt haben. Die Reproduktion der Interaktion zwischen den einzelnen *Netdoms* (Ärzt:innen, Pflegepersonal, Patient:innen) und der den Netzwerken zugehörigen kulturellen Formen kann bei gegenseitiger Anerkennung für einen Pol der Stabilität sorgen (White 2008: 36f.). Dies wird auch als *Switching* zwischen den *Netdoms* bezeichnet, also ein kommunikativ vollzogener Wechsel zwischen Kontexten (vgl. White 2008: 2; Schmitt/Fuhse 2015: 138). Die Ärzt:in wechselt von ihrem Kommunikationsspektrum mit ihren Kolleg:innen, in der eine andere Art der Sprechweise vorherrscht (Fachbegriffe etc.), in

das der Patient:innen. Werden die *Stories* zum Einsatz der KI-Methoden in der Medizin, die innerhalb der Netdom der Ärzteschaft existieren, im Switching-Prozess mit der Patient:in-Netdom kommuniziert? Welche *Stories* werden wo geteilt (vgl. Schmitt/Fuhse 2015: 137f.)?

Wie sich im weiteren Verlauf zeigen wird, sind die *Stories*, also die Rhetorik (Sprechakte) innerhalb dieser Interaktionen, dabei ausschlaggebend (Schmitt/Fuhse 2015: 95f.). Wie später noch eingehend diskutiert wird, fließen inzwischen viele algorithmische Faktoren in Diagnose und Therapie mit ein, finden aber in den Gesprächen zwischen Patient:in und Ärzt:in kaum statt. Hier werden Kontrollkämpfe verschränkt, indem die *Stories* geteilt, anerkannt und berichtet werden. Die Wiederholungen dieser Vorgänge bringen den Stil hervor und symbolisieren ein gemeinsames Verständnis innerhalb der Institution (Schmitt/Fuhse 2015: 120ff.; vgl. White 2008: 112).

Und damit bleibt ein weiterer Fokus der Analyse vorzustellen, der oben bereits kurz angerissen wurde: *Kontrollkämpfe*, welche entstehen, wenn neue Akteur:innen, Praktiken oder veränderte Kontexte eine Überprüfung und gegebenenfalls Veränderung von Gegebenem veranlassen. Sie entstehen immer und überall, wenn es darum geht, die Identität – also den Zuständigkeitsanspruch – zu koppeln und zu entkoppeln (White 2008: 7f.). Dementsprechend müssen die Beteiligten (vor allem die Ärzt:innen) in der Institution die Kontrolle über das Wissen erhalten und sich darin selbstregulativ organisieren. Nicht nur die Anwendung des Wissens, sondern auch dessen Schaffung und Weitergabe müssen in der Hand der Institutionsmitglieder liegen (Abbott 1988: 16; Mieg 2005: 342f.). Die Integration neuer Praktiken im Sinne neuer Technik für Diagnose, Therapie und Prävention, informatische Einflüsse in der Medizin und damit auch neue Akteur:innen, sind als solche Kontrollkämpfe interpretierbar. Sie werden durch *Stories* ausgedrückt und beeinflussen den Stil – stabilisierend als auch destabilisierend.

In Abbildung 1 kann dieses Zusammenwirken von stabilisierenden und innovativen Effekten für Wandel nachvollzogen werden. Während aus der Profession heraus ein ständiger Wandel angestoßen wird, indem neue Diagnose- und Behandlungsansätze eingesetzt werden, wirken die Reproduktionsmechanismen der Netdoms äußerst stabilisierend auf die Institution. Diese kann sich so beständig weiterentwickeln – ohne dabei Gefahr zu laufen, sich selbst abzuschaffen.

Erleben wir heute schon einen Kontrollkampf zwischen Medizin und der informationstechnischen Wissenschaft? Müssen wir in Zukunft damit rechnen, eher Informatiker:innen und intelligenten Programmen im Gesundheits-

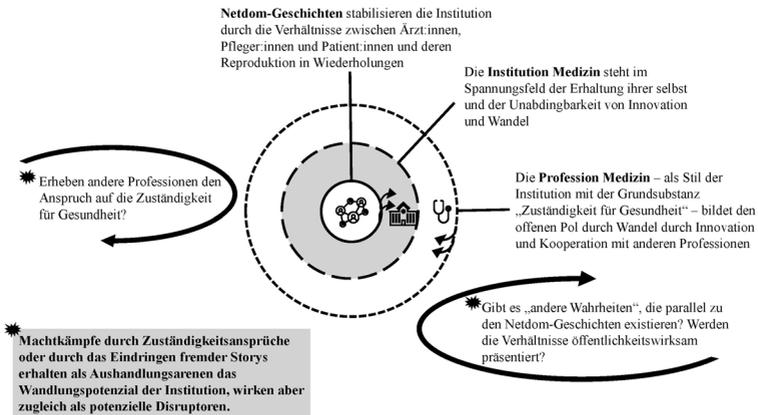
bereich zu begegnen? Oder wären wir damit einer medial ausgestalteten Geschichte auf den Leim gegangen?

Um diesen Fragen nachzugehen und zu beantworten, ob die Institution Medizin Gefahr läuft, den Kontrollkampf gegen informatische Einflüsse zu verlieren, werden nun Auszüge aus den Interviews vorgestellt und mit den Theoriebausteinen verbunden.

3. Selbstbewusste Forscher:innen und Anwender:innen im Innovationsfeld der Medizin

»Medizin ist grundsätzlich ein extrem innovatives Fach und wir leben seit Jahrzehnten in einem innovativen Umfeld. Also Innovation macht in der Medizin keine Angst und sollte keine Angst machen.« (Interview Staartjes/Regli: Z. 325–327)

Abbildung 1: Die Institution Medizin zwischen Stabilitäts- und Wandlungspotenzial



Die Medizin als genuin innovative Institution

Was in den Medien seltener im Zusammenhang mit KI und Medizin behandelt wird, ist, dass die künstliche Intelligenz (KI) schon längst intensiven Einzug in den Klinikalltag erhalten hat. Viele Algorithmen kommen in der klinischen und bildgebenden Forschung bereits automatisiert zur Anwendung, vor allem im Operationssaal. Die Prädiktion des Erfolgs einer Operation kann kalkuliert, beispielsweise ein Tumor genau und ohne störende anatomische Strukturen lokalisiert und mit verschiedenen KI-basierten Instrumenten entfernt werden (Interview Kernbach: Z. 24f.). Vorreiter in der KI war die Radiologie, da in diesem Bereich schon sehr lange bildgebend und digital gearbeitet wird – es läge hier quasi »in der Natur der Dinge« (Interview Echle: Z. 73–77). Diese Kenntnisse wurden in andere Fachrichtungen getragen, sodass Chirurgie, aber auch die Innere Medizin von den Algorithmen profitieren konnten. Teilweise sei das Deep Learning schon »out« bzw. so fest etabliert, dass viele Forscher:innen bereits versuchen, diese in Methodik und Anwendung weiterzuentwickeln. Sie wollen die Methodik so weiterentwickeln, dass Vorhersagen über spezielle und seltene Krebsarten durch histologische Bilder geschlossen werden können und eben nicht nur die gängigen, höchst wahrscheinlichen Krebsarten gut erkannt werden. Auch spezielle Varianzen bzw. Mutationen sollen sofort und genau erkannt werden können, um dem/der Patient:in eine differenzierte und personalisierte Behandlung anbieten zu können (Interview Echle: Z. 39f.). Dabei ginge es vor allem um die Sichtbarmachung von Erkrankungen, die für das bloße ärztliche Auge nicht sichtbar sind (Interview Echle: Z. 34). Es wird deutlich, dass Anwendungen mit KI schon lange im Praxisalltag der Medizin angekommen sind, ohne dass damit ein paradigmatischer Wechsel für die Institution einhergeht.

Verständnis für ein vulnerables Kollektiv und dessen Befürchtungen

Trotzdem findet sich neben der Befürwortung von KI in der Medizin vor allem auch Verständnis für die Ängste der Patient:innen und der Bevölkerung. Laut den interviewten Ärzt:innen und Forschenden gehören die Patient:innen zu einem vulnerablen Kollektiv, da die Gesundheit des Menschen ein hohes und sehr sensibles Gut darstelle. Die Assoziation der Gesundheit mit dem Tod berühre eine Grundangst des Menschen, weswegen der Gesundheitsstatus und auch die Daten dazu einen hohen Wert besäßen und besonderen Schutz bedürfen (»[...] health is always associated with death [...]« (Interview

Hadjilioniadis: Z. 444f.)). Nicht umsonst sind mit der Offenlegung einer Krankheit viele soziale Stellungen (Arbeit, Versicherung etc.) verbunden. Da die Zuständigkeit für die Gesundheit des Menschen in die Institution Medizin verlegt wurde, dieser Schutz aus dem Berufsstand erwachsen ist und viele Themen zudem immer noch tabuisiert sind, liegt allen Beteiligten daran, diese Daten sicher zu wissen. Dem Datenschutz gilt noch immer die ausgeprägteste Sorge der Patient:innen, da die KI aus dem Wirtschaftswesen mit anderen ethischen Anforderungen käme und es Sorge um Sicherheitslücken gäbe (Interview Kernbach: Z. 308–319). Des Weiteren besteht die Befürchtung, dass das Ärzt:innen-Patient:innen-Verhältnis unpersönlicher wird und die Entscheidungen, welche Handlungsoptionen und weitere medizinische Maßnahmen nun eingeleitet werden müssen, am Ende ausschließlich bei der KI liegen. Dass die Institution Medizin ihre *Menschlichkeit* verliert, ist ein großer Bestandteil dieser Befürchtungen. Auch die Interviewees wissen um die Befürchtungen der Patient:innen, die sie in ihrem eigenen Umfeld spüren, wenn es um die Speicherung und Weitergabe von Patient:innendaten geht, die durch den Einsatz der KI weiter beflügelt werden (Interview Echle: Z. 407–414). Die interviewten Ärzt:innen zeigen Verständnis und betonen die Wichtigkeit der transparenten Kommunikation, aber argumentieren im Gegenzug, weswegen sich ihr Verständnis durch diese mangelnde Kommunikation ihrer eigenen Rolle als Ärzt:in nicht verändern würde (vgl. Interview Echle 416f., 421f.; vgl. Interview Kernbach: Z. 123ff.; Interview Staartjes/Regli: Z. 508ff.). Die transparente Kommunikation zwischen Ärzt:innen und Patient:innen und wie sich ihr Fehlen auswirken könnte, wird im weiteren Verlauf noch einmal wichtig sein.

Stil, Kontrollkämpfe und Zuständigkeit als Garanten für Resilienz und Wandel

Zuerst soll nun aber die Resilienzwahrnehmung des interviewten medizinischen Personals weiter erörtert werden. Die Art und Weise, wie Ärzt:innen arbeiten, ist fest im Stil der Profession verankert. Ein Stil entsteht durch die Kontexte und reproduziert sich selbst durch stochastische und bewusste, aus der Gewohnheit herausgebildeten Wiederholungen. Sie stellen vor allem kombinierte Verständnisse von Situationen zusammen mit bestimmten Sets an Praktiken dar. Beispielsweise nimmt ein:e Internist:in stets zu Beginn einer jeden Untersuchung Blut ab, um bestimmte Werte zu analysieren. Die Kontexte selbst ergeben sich aus den Beziehungen in den Netzwerken der Ärzt:innen,

die durch die Kontrollkämpfe der jeweiligen Identitäten entstehen. Besteht ein Stil einmal, reproduziert er sich in seinen eigenen Kontexten (White 2008: 112ff.). Neben den stochastischen Verteilungen braucht es allerdings auch die Beobachtung von Dritten, die diese auch als Stil anerkennen und sensibel für die sich daraus ergebenden Stilmuster sind. Die Praxis der Ärzt:innen muss also von den Patient:innen richtig erkannt und interpretiert werden, ein weiterer wichtiger Faktor, der im nächsten Abschnitt näher erläutert wird.

Die Profession als Stil setzt laut Abbott Kontrolle über Wissen voraus, welches durch die Selbstorganisation beziehungsweise -regulation der Ärzte gegeben ist. Diese wird durch die Kontrollkämpfe untereinander, die auch notwendig für den Erhalt und die Weiterentwicklung des Stils sind, gestärkt beziehungsweise auch destabilisiert. Gerade in der Forschung und Ausbildung, in der die KI weiter ausgebaut sowie verändert wird und die Etablierung in die Praxis erfolgt sowie in der bereits angewandten Praxis, liegt diese Kontrolle des Wissens (Interview Kernbach: Z. 165ff.; Interview Echle Z. 372ff.; Abbott 1988: 3, 89ff; Miege 2005: 342). Denn die Stärke der Profession liegt in der Zuständigkeit, die sich im Laufe der Geschichte entwickelt hat – in diesem Fall der Gesundheit. Da die Zuständigkeit die Grundsubstanz jeder Profession ist, könnte eine Unterhöhlung lediglich durch die Ärzt:innen selbst erfolgen, da hier Wissen (re-)produziert und kontrolliert wird.

In der Ausbildung von Ärzt:innen erfolgt de facto eine Zunahme der möglichen Zusatzkurse für die Anwendung der KI, viele praktizierende und forschende Ärzt:innen arbeiten aber auch mit technischen Angestellten zusammen, die die KI auf Grundlage ihrer Forschung in die Tat umsetzen (Interview Kernbach: Z. 168ff.; Interview Staartjes/Regli Z. 320ff; Interview Echle Z. 387ff.). Es findet somit ein enger Austausch in Form von Kontrollkämpfen und Stories zwischen den verschiedenen Netdoms der Ärzt:innen in Forschung, Bildung und Anwendung statt sowie zwischen den zwei Netdoms der Informatiker:innen und Ärzt:innen. Der Beitrag »Deep Learning Algorithmen als Boundary Objects zwischen Entwicklungs- und Anwendungsfeld« des Sammelbandes befasst sich intensiv mit diesem Kontextwechsel der Anwendenden und Programmierenden und betrachtet, inwiefern Deep Learning Algorithmen als Boundary Objects für die Kommunikation zwischen ihnen dienen können (siehe Roth/Dorgeist/Schulz in diesem Band). Als Kompetenzabgabe sei diese interdisziplinäre Zusammenarbeit jedoch nicht zu verstehen, sondern eher als Kooperation. Die Etablierung des Stils würde gerade durch diese Aushandlungsprozesse weiter vorangetragen (Interview Echle: Z. 379ff.; vgl. Interview Kernbach: Z. 165ff.). Denn eine komplette Abgabe der Kontrolle

ist seitens der Ärzt:innen nicht gewünscht. Alle komplexen technischen Zusammenhänge zu verstehen, welches zum Beispiel das Black-Box-Problem mit einschließt, steht bei den meisten Ärzt:innen jedoch nicht im Fokus. Die Interviewees betonen aber, dass das Hinterfragen dieser neuen Ergebnisse, die durch komplexe Algorithmen produziert werden, immer noch notwendig sei. Immerhin befinden sich Ärzt:innen in einem »geschützten Kollektiv«, welches gerade wegen der Patient:innen auch weiter geschützt werden sollte: Ein Plädoyer für *explainable KI* (Interview Kernbach: Z. 94–96; 123–126)! Schmitt und Heckwolf thematisieren in ihrem Beitrag das Black-Box-Problem in seiner gesamten Komplexität und analysieren die hier beschriebene resultierende Intransparenz der tatsächlichen Funktion für die Expert:innen, die an der Konstruktion beteiligt waren (siehe Schmitt/Heckwolf in diesem Band). Durch die Zusammenarbeit der Nettdoms der Medizin und der Informatik werden Ärzt:innen nicht zu verkappten Informatiker:innen. Die Grundlagen sind laut den Interviewees das Entscheidende, um die medizinische Expertise zu gewährleisten. Zudem sind bereits viele Praktiken der KI etabliert, die laut Ärzteschaft nicht als Bruch in ihrer Art zu arbeiten, sondern als standardisiertes Instrumentarium anzusehen sind. Der KI wird kein Einfluss auf die Rolle der Ärzt:in zugeschrieben (Interview Staartjes/Regli: Z. 83f.).

Wie oben beschrieben, geben erst die Zuständigkeitsansprüche und Konkurrenz dabei den Anstoß für organisatorische Entwicklung und Professionalisierung (Abbott 1988: 16; siehe auch Miege 2003; Abbott 1988: 17). In den Interviews wurde sogar für diese Kontrollkämpfe der verschiedenen Ärzt:innen-Nettdoms geworben, um die KI weiter voranzubringen und einen Fortschritt zu erzielen (Interview Echle: Z. 253f.). Diese Vernetzung sieht eine der Interviewees auch in der digitalen Patient:innenakte, die einerseits die Kommunikation der Ärzt:innen untereinander erleichtern und zugleich zu weiteren – gewollten – Kontrollkämpfen führen würde (Interview Echle: Z. 404f.). Nun bietet der Stil der Profession Orientierung im sozialen Leben. Stile sind jedoch ephemere und können durch diese Kontrollkämpfe Lücken sowie Brüche aufweisen, durch die eine Veränderung möglich ist (Schmitt/Fuhse 2015: 122). Diese Kontrollkämpfe und Klüfte geben gerade nach obigen Argumentationen dem Stil der Ärzt:innen diese Stabilität in ihren Rollenverständnissen, solange – und das ist bislang der Fall – die Grundzuständigkeit der Medizin für die Gesundheit erhalten bleibt (Interview Staartjes/Regli: Z. 501f.) Die theoretischen Annahmen zu innerinstitutionellen Wechselwirkungen sowie den Fremdeinflüssen und ihrer Wirkung decken sich also mit den Aussagen der Interviews.

Die KI hat also längst Einzug in die Praxis der Ärzt:innenschaft erhalten. Die Adaption der KI gelang durch die Zuständigkeit der Ärzt:innen, die diese neuen Tools in ihren Arbeitsalltag integriert haben – auch durch Kontrollkämpfe, die zu einem Gleichgewicht führten. Mit Einzug von KI-basierten Tools in der Medizin hat es Veränderungen der Rahmenbedingungen gegeben (Datenschutz, Speicherung von Daten, Datenaustausch). So ist ebenfalls sichtbar geworden, dass die Profession der Ärzt:in auch abhängig von und im ständigen Austausch mit anderen Institutionen und Netdoms ist (Institution Rechtssystem zum Thema Datenschutz aber auch wirtschaftliche Unternehmen wie Google). Damit hat es allerdings nicht so deutliche Veränderungen in der Kontrolle des Wissens gegeben, dass sie zur Grundlage für einen Zuständigkeitswechsel werden könnten (vgl. Interview Staartjes/Regli: 354ff., 378ff., 398ff.; Interview Kernbach: Z. 225ff.). Sowohl die Erzeugung neuen Wissens wie auch die Weitergabe dessen in Bildungseinrichtungen stehen in der absoluten und unbestrittenen Zuständigkeit der Medizin, so die Auffassung der medizinischen Expertise.

Switching-Prozesse und die Notwendigkeit der Thematisierung neuer Technologien

Doch woher kommt dann das medial artikuliert Unbehagen gegenüber neuem intelligenten Assistenzsystemen in der Medizin? Ein Grund dafür könnte in der Kommunikation liegen. In der Kommunikation zwischen Ärzt:innen und Patient:innen werden diese neuen Tools laut Aussagen der Ärzt:innen nämlich nicht erwähnt. Lediglich die Diagnostik spielt hier eine übergeordnete Rolle für die Patient:innen. In den Switching-Prozessen zwischen den einzelnen Netdoms findet keine hinreichende Rhetorik in Form von Stories statt (in Form von Ärzt:in-Patient:innen Gesprächen, Interaktionsregeln, Sprachmuster, Bewertungen), die eine geeignete Interpretation der Patient:innen über die Praxis der Ärzt:innen mit der KI zulässt. Die Sensibilität und Beobachtung der Patient:innen ist für eine geeignete Interpretation nicht hinreichend – die öffentliche Sorge ist somit eher als Narrativ, denn als begründete Angst zu analysieren. Die verschiedenen Sphären (oder laut White Institutionen) sowie Netdoms koexistieren zwar ausreichend mit Informationsvorbehalten, sie führen aber auch in der öffentlichen Zuständigkeit zu einem Verschwinden der Welt der Tatsachen zu einer Welt der Möglichkeiten (Abbott 1988: 89). Von den Interviewees wird deswegen zu verstehen gegeben, dass bezüglich digitaler Innovationen in Krankenhäusern »besser oder anders« (Interview Echle: Z. 206–207)

als in anderen Fällen kommuniziert werden muss, als es beispielsweise bei Bekanntem, wie zum Beispiel Spritzen der Fall wäre. Im Switching-Prozess von Ärzt:innen und Patient:innen müssen die Netdom-Stories zur KI ausgetauscht werden, um die Akzeptanz und das Rollenverständnis der Ärzt:innen auch in der Patient:innen-Netdom fest zu etablieren und interpretierbar zu machen. Die transparente Kommunikation in Richtung der Patient:innen ist also ein integraler Bestandteil, um ein Gegengewicht zum medialen Narrativ der abdankenden Ärzt:innen zu bilden oder es bestenfalls sogar zu entkräften.

Ein weiterer bedeutender Faktor einer Institution ist ihre Bedeutungsebene, die wichtig für die Aushandlung des täglichen Lebens der Akteur:innen ist. Belangvoll sind vor allem die zusammenhängende bedeutungsvolle interpretative Erfahrung der täglichen Existenz mit der physisch verkörperten Erfahrung der Akteur:innen, die durch soziale Strukturen verbunden sind. Das bedeutet, dass die Interaktionen in den einzelnen Netdoms interpretierbar sind und durch den Switching-Prozess zwischen den Netdoms unterschiedliche Verständnisse – in dem Fall Stories über den Arbeitsstil von Ärzt:innen – generiert und gebildet werden (Mohr/White 2008: 8f.; vgl. White 1995).

Wie an vorherigen Textstellen jedoch deutlich wurde, ist die Problematik die mangelnde Kommunikation der Ärzt:innen, die nicht erwähnen, welche KI-Mechanismen hinter der Diagnostik stecken. Deswegen werden vermeintliche Befürchtungen auch durch eine andere Institution – durch die der Medien – herangetragen. Denn was eine Institution folglich so robust gegenüber Veränderungen macht, ist die Eigenständigkeit der verschiedenen Netdoms innerhalb der Institution, die zur Gesamtkohärenz dieser beitragen (Mohr/White 2008: 15). Die Switching-Prozesse sind allerdings nicht mit Bedeutungen rund um die KI behaftet, sondern beinhalten nur die wichtigen Informationen rund um das Thema Gesundheit und Genesung der Patient:innen. Ein Grund ist auch die tägliche Benutzung von KI, ohne sich deren Anwesenheit und dessen Weiterentwicklung bewusst zu sein. Beispielsweise werden Smart Home oder auch die Smart Watch, die einfache Herzfrequenzen aufnehmen und verarbeiten kann, bereits von einer Vielzahl von Menschen im Alltag genutzt. Die Ärzt:innen können es sich im Alltag nicht mehr vorstellen, dass sich solche Techniken und Praktiken nicht nachhaltig ausbreiten – sie tun es wie gesagt bereits (Interview Staartjes/Regli: Z. 43–47; »Wir benutzen täglich Sachen, vielleicht ohne zu wissen, dass es AI-Algorithmen drin hat. Aber eigentlich wie jeder Mensch heute, wenn er täglich etwas macht.«, (ebd.: Z. 46–47)).

Aber auch der Austausch mit dem Krankenhauspersonal und anderen Ärzt:innen, also in der eigenen Netdom, ist entscheidend (Interview Echle: Z.

204–213; Interview Kernbach: Z. 232–234). Denn Ziel der medizinischen Forschung ist zum einen die Verbesserung der Diagnostik und zum anderen eine individuelle Behandlung, die entgegen des *Selection Bias* des *weißen, männlichen und 50-jährigen Patientenkollektivs* arbeitet, um damit der Diskriminierung in der Medizinpraxis etwas entgegenzusetzen (Interview Echle: Z. 126, 268; Staartjes/Regli: Z. 186, 280ff.). Folglich wären generalisierbare Methoden zu erforschen und zu erproben, die dann im Einsatz individuell an den Kontext der Patient:innen angepasst werden können (Interview Staartjes/Regli: Z. 545ff.).

Arbeit mit KI als neuer Stil der Institution

Der Treiber von Wandel kann laut White und Mohr nur der Stil sein. Er wird durch Chaos und Kämpfe innerhalb des Stils hervorgerufen. Da Medizin jedoch ein grundsätzlich innovatives Fach sei und in einem innovativen Umfeld gearbeitet und geforscht würde, wären Wandlungsprozesse und auch Normalität erwünscht (Interview Staartjes/Regli: Z. 325f.). Zumeist ist ein komplett neu eröffnetes Paradigma – vor allen Dingen eine kulturelle Reformulierung – die Veränderungsvariable, die allerdings eine größere Umstrukturierung erfordert. Es müsste somit beispielsweise ein neues Krankenhaus mit einem neuen Kollegium geschaffen werden, oder eine Abspaltung eines etablierten Krankenhauses erfolgen, die dieses neue Paradigma (beispielsweise die Verbannung der KI) im Arbeitsprozess implementiert und ausführt (Mohr/White 2008: 16f.). Da auch etablierte und erfahrene Ärzt:innen auf die KI vertrauen, passen sich diese lediglich dem neu etablierten Stil mit KI zu arbeiten an, die Institution selbst wird aber nicht tangiert. Auch lang praktizierende Ärzt:innen betonen ihre Unbesorgtheit gegenüber der KI, die nicht kreierte, sondern lediglich – wenn auch gelungen – kopiere und ein perfektes »Plagiat« erstellen könne. Die Ärzt:innen hätten jedoch mehr Aufgaben, sodass die KI Ärzt:innen nicht ersetzen werde. Sie stelle aber ein geeignetes Instrumentarium dar, um diese zu verbessern (Interview Staartjes/Regli: Z. 202f., 206–210, 214–216, 226f.). »Also ich glaube nicht, dass viele Ärzte Angst haben, dass diese Rolle zerstört wird, diese Super Hero Rolle. Ich denke, wir werden nur besser dadurch werden.« (Ebd.: Z. 500–502)). Denn die KI könne keine Entscheidung vorgeben oder treffen, die es noch nicht gesehen oder gelernt hätte, sie beziehe auch nicht die sozialen Situationen der Patient:innen mit in die Analyse ein. Gerade deswegen sei die Ärzt:innenschaft in ihrer Entscheidungshoheit

so wichtig, um kreativ über andere Lösungsmöglichkeiten zu entscheiden (Interview Staartjes/Regli: Z. 238ff., 276):

»Eigentlich, was KI macht, ist, was wir alle universitär hassen, es ist ein Plagiat. KI wiederholt, was er gelernt hat und reproduziert, was man ihm gelernt hat und er macht das extrem gut. [...] Der Arzt macht mehrere Aufgaben und das denke ich, hat die KI noch nicht auf den Punkt gebracht. Und deswegen denke ich nicht, dass KI den Arzt ersetzen wird, er wird den Arzt aber besser machen.« (Interview Staartjes/Regli: Z. 206–216)

Nicht umsonst argumentieren die Interviewees, dass der Einsatz der KI auch schon im Bereich der Nachbehandlung und Prävention eingesetzt wird. Es findet also bereits eine Erweiterung des Wirkungskreises in den Privatbereich statt (Handy, Smart Watch etc.). Laut den Interviewten sei dies »nicht komplett Fiktion« (Interview Echle: Z. 356), da jetzt schon simple EKG-Streifen von Apple Watch und Co. aufgezeichnet werden können (Interview Kernbach: Z. 270–276). Es wird auf Forschungsprojekte hingewiesen, in denen die Nachbehandlung und somit der Geltungsbereich der Medizin auf den Haushalt ausgeweitet und mithilfe der dort generierten Daten sowohl die Behandlung der Patient:innen verbessert werden kann, als auch eine Objektivität durch die Daten gewährleistet wird (Interview Staartjes/Regli: Z. 35–38; vgl. Abbott 1988: 86). Zu nennen sei hier das Forschungsprojekt iPrognosis, bei dem Handydaten zur Früherkennung von Parkinsonerkrankten benutzt werden, um auch dort die wertvollen Daten für die weitere Forschung nutzen zu können (Interview Hadjilioniadis: Z. 177–192). Diese für die Forschung nicht zu verwenden, wäre laut Ärzt:innen und Forschenden eine Verschwendung von Daten, die sowieso zur Verfügung stehen könnten. Dementsprechend wäre in Zukunft eher eine Aktivierung der Patient:innen und damit auch deren Ermächtigung zu erwarten, mit einer progressiven Passivierung von Ärzt:innen jedoch nicht:

»Eigentlich kann man fast sagen, die KI macht vielleicht eine physische Distanz, aber [...] wir verstehen den Patienten noch besser Dank der ganzen Informationen. Also eigentlich kommen wir ihm fast mehr unter die Haut, obwohl wir physisch viel weiter weg sein können, um das zu regeln.« (Interview Staartjes/Regli: Z. 487–490)

Die Rolle der Institution der Medien

Wodurch kommen dann die medial formulierten Annahmen und Ängste, wenn die Institution Medizin und die Profession der Ärzt:innen zwar Veränderungsprozesse durch die KI erfährt, sich aber das Selbstbild und die grundlegende Arbeitspraxis nicht verändert? Es werden weitere Verständnisse durch verschiedene Netdom-Switchings an die Patient:innen und die breite Bevölkerung getragen. Diese werden durch die Institution der Medien in verschiedenen Stories durch die Netdoms verbreitet, die ein anderes Verständnis der KI implizieren. Da diese Bedeutungen besser beobachtbar, interpretierbar, rhetorisch aufgeladener und vor allem leichter zugänglich für die Patient:innen und die Bevölkerung sind, wird dieses neue, drastische Verständnis auf die Profession der Ärzt:innen seitens der Patient:innen übertragen. Zudem liegt laut Abbott dem öffentlichen Diskurs über die Zuständigkeit die Annahme zugrunde, dass die Aufgaben der Profession objektiv definiert sind (Abbott 1988: 69). Durch den Mangel an geteilten Stories – oder auch Geschichten – zu Patient:innen und der Öffentlichkeit wird diese Definition verzerrt und eben nicht mehr objektiv definierbar. Auch in den Interviews wurde diese Annahme bestätigt: Die Abschaffung der Ärzt:insei in erster Linie eine »medieninteressante Frage« (Interview Kernbach: Z. 188f.), also eine, die von der klinischen Alltagsrealität vollkommen losgelöst sei. Denn dass die KI schon längst angewendet wird, bestreitet niemand. Es entstehen jedoch immer kompliziertere Anwendungen (Black-Box-Problem: siehe Schmitt/Heckwolf in diesem Band), die die Aufmerksamkeit der Bevölkerung erregen, da sie nicht leicht verständlich sind und von der Institution der Medien leicht durch Narrative der Sorge und Angst aufgegriffen werden können. Die Interviewees sind sich allerdings sicher, dass diese Verfahren trotz Undurchsichtigkeit zu Vertrauen bei den Patient:innen führen können, da sie vorerst und so lange notwendig durch historisch sehr etablierte Verfahren abgesichert werden (Beispiel Biopsie¹). Hier lässt sich also ein Übergangsprozess erkennen – der Versuch, die Akzeptanz und das Vertrauen der Patient:innen zu gewinnen (Interview Staartjes/Regli: Z. 135f., 141–144). Auch hier wird das Problem von den Kolleg:innen (Roth/Dorgeist/Schulz in diesem Band) des Sammelbandes aufgegriffen. Den befürchteten Risiken und Folgen sollten genauere Analysen

1 Bei einer Biopsie entnehmen Ärzt:innen mit unterschiedlichen Techniken Zellen oder Gewebeproben (Biopsat) aus dem Körper (Kohn 2013).

und Beobachtungen der Entwicklung von DeepLearning-Algorithmen mit Anwendenden und Programmierenden entgegengestellt werden.

4. Fazit

Problematisierung der Selbstbilder

Wie sich gezeigt hat, äußern sich die interviewten Mitglieder der medizinischen Institution sehr selbstbewusst. So gelesen gäbe es keinen Grund zur Sorge. Dennoch fallen bei genauem Hinsehen einige Widersprüche auf. Bestehen die Selbstbilder den Realitätstest?

Insbesondere fallen Widersprüche bei Betrachtung der Selbstbilder auf. Die interviewten Personen zeigten deutliches Verständnis für die Sorgen des vulnerablen Patient:innenkollektivs – sie sprachen sogar konkrete Arbeitsfelder an: die Sensibilität der Patient:innendaten, die Rolle des Datenschutzes sowie das Lösen des Black-Box-Problems. Gleichzeitig jedoch wird deutlich gemacht, dass die Technologie im Dialog mit diesem Kollektiv quasi nicht stattfindet. Die Patient:innen wissen also nicht, wann und welche Technologie zur Diagnose und Therapie eingesetzt wird, da es für die ärztliche Arbeit schlicht keinen Unterschied macht. Daraus kann man zwei Schlüsse ziehen: Erstens könnte der Ärzt:innenschaft somit unterstellt werden, die technologischen Leistungen als eigene zu verbuchen, da Patient:innen ohne dessen Thematisierung davon ausgehen müssen, dass es sich um eine vollkommen menschliche Leistung handle (Schluss 1). Zwar liegt der Zuständigkeitsanspruch und somit die Möglichkeit des Wandels der Institution bei den Ärzt:innen, jedoch muss diese Weiterentwicklung, eben die Nutzung von KI, an die Netdoms der Patient:innen weitergegeben werden, um die Möglichkeiten und Behandlungsverbesserungen durch die KI weiter zu unterstreichen. Zweitens – und dieser Schluss wiegt deutlich schwerer – wird hier eventuell eine Chance verspielt, aktiv für Aufklärung zu sorgen (Schluss 2). Indem Mediziner:innen die Technologien in der Rhetorik zwischen Patient:innen und Ärzt:innen einbauen, würden sie Schluss 1 entkräften und gleichzeitig Vertrauen in die Technologien schaffen, denen sie als Expert:innen längst vertrauen. Somit würden sie – mit den White'schen Begriffen gesprochen – dafür sorgen, dass die Technologie nicht mehr nur in der Netdom der Ärzt:innenschaft stattfindet, sondern durch die Switching-Prozesse auch in der der Patient:innen ankommt. Davon wären nach diesem Modell sowohl stabili-

sierende Effekte für die Institution als auch die Entschärfung der medialen Sorgen zu erwarten, da die mediale Institution bei besserer Aufklärung der Patient:innen schlicht kein Publikum mehr für das Narrativ der abdankenden Ärzt:innen hätte. Denn der Einfluss der Stories, die von der Netdom der Medien in die Netdom der Patient:innen und folglich auch in die Netdom der Bevölkerung haben, können negative Rückwirkungen auf die Institution der Ärzt:innen haben. Diese Intransparenz für das Patient:innenkollektiv zu beseitigen sollte Ziel der Institution sein, da das Selbst- sowie Fremdbild der Ärzt:innen deutlich ins Straucheln geraten könnte. Sonst überließe sie diese Arbeit einfach anderen Institutionen und gäbe folglich eine gewisse Menge an Kontrolle ab.

Ausblick

Um diesen Ansatzpunkt weiter zu verfolgen und die Argumentation zu bekräftigen, sollten folglich auch Interviews mit Patient:innen und jeglichen Personen aus der Institution Medizin (Pflege, Verwaltung, Leitung etc.) geführt werden, um einen Abgleich mit den dargestellten Interviews zu schaffen. Zudem ist es wichtig, die Medien stärker miteinzubeziehen und die Narrative zu finden, die in der breiten Öffentlichkeit kursieren, da diese einen erheblichen Einfluss auf die Netdoms der Patient:innen haben. Ein weiterer interessanter Untersuchungsgegenstand ist, wie sich die Profession der/des Ärzt:in, näher der Zuständigkeitsanspruch, im Laufe der heutigen Zeit verändert haben könnte. Liegt der Zuständigkeitsanspruch immer noch in der Ärzt:innen-schaft, oder sind die Kontrollkämpfe der Institution Medizin so stark geworden, dass sie diesen Zuständigkeitsanspruch untergraben könnten? Denn die Medien haben durch Fake News, ChatGPT und durch ihren uneingeschränkten Zugang deutlich an Stärke und Einfluss gewonnen.

Die Argumentation, welche mithilfe der Theorie von White sowie auch Abbott durchgeführt wurde, zeigt bereits deutlich, dass sich das Rollenverständnis der Ärzt:innen durch den Einsatz der KI im Kern nicht verändern wird. Lediglich die Praxis – der Stil zu arbeiten – wird gerade durch die disruptiven Kontrollkämpfe eine (positive) Veränderung erfahren.

Ärzt:innen büßen nicht an Selbstbewusstsein bezüglich Ihrer Kompetenzen ein, sondern sehen neue digitale Assistenzsysteme als das was sie sind: Assistenzsysteme. Selbst bei einer möglichen Verselbstständigung der KI wird optimistisch in die Zukunft geblickt und betont, dass zwar sehr viel geforscht wird, aber die Etablierung der neuen Tools im medizinischen Alltag bei wei-

tem nicht derart schnell voranschreitet, wie die mediale Berichterstattung oder Forschung dies nahelegen. Der finale Einsatz in der Praxis unterliegt weiterhin der größten Sorgfaltspflicht. Es bleibt allerdings eine Frage der Zeit, bis mehr und mehr Anwendungen gemäß diesen Regeln in den Alltag integriert werden (vgl. Interview Staartjes/Regli: Z. 575f., 580–584).

6. Geführte Interviews

Prof. Lontios Hadjilontiadis (Chair of Biomedical Engineering-Khalifa University (UAE); Prof. of Electrical & Computer Engineering Aristotle University of Thessaloniki; Projekt: iPrognosis)

Amelie Echle (Universitätsklinikum Aachen; Forschungsteam: Entwicklung von DeepLearning-Methoden für die digitale Pathologie)

Dr. med. J. M. Kernbach (Universitätsklinikum Aachen; Sectionleader Machine Learning & Artificial Intelligence; Neurosurgical Artificial Intelligence Laboratory Aachen (NAILA))

Prof. Dr. med. Luca Regli (University Hospital Zurich, University of Zurich; Chefarzt des Department of Neurosurgery, Clinical Neuroscience Center) und **Dr. Victor Staartjes** (University Hospital Zurich, University of Zurich; Machine Intelligence in Clinical Neuroscience (MICN) Lab)

7. Literatur

Abbott, Andrew (1988): *The system of professions*, Chicago: University of Chicago Press.

Bauch, Jost (2000): *Medizinsoziologie*, München: Oldenbourg.

Ekkerkamp, Axel (2019): »Künstliche Intelligenz in der Medizin – Am Krankenbett mit Doktor KI«, in: *Tagesspiegel* vom 24.07.2019, <https://www.tagesspiegel.de/wissen/kuenstliche-intelligenz-in-der-medizin-am-krankenbett-mit-doktor-ki/24691876.html>.

Freidson, Eliot (1970): *Profession of medicine: A study of the sociology of applied knowledge*, New York: Dodd, Mead and Co.

Freidson, Eliot (1983): »The theory of professions: State of the art«, in: Robert Dingwall/Philip Lewis (Hg.), *The sociology of the professions*, London: Macmillan, S. 19–37.

- Freidson, Eliot (1986): *Professional powers*, Chicago: University of Chicago Press.
- Freidson, Eliot (2001): *Professionalism: The third logic*, Cambridge/UK: Polity.
- Freidson, Eliot 1994: *Professionalism reborn: Theory, prophecy and policy*, Chicago: University of Chicago Press.
- Gassen, Andreas (2019): »Forum Digitalisierung: Medizin im Wandel – Wie verändert die Digitalisierung den Praxisalltag und Arztberuf?«, in: kbv vom 29.04.2022. <https://www.kbv.de/html/38790.php>.
- Häußling, Roger (2020): »Zur Materialität von sozialen Netzwerken in der Digitalen Gesellschaft«, in: Werner Friedrichs/Sebastian Hamm (Hg.), *Zurück zu den Dingen!: politische Bildungen im Medium gesellschaftlicher Materialität*. 1. Auflage, Baden-Baden: Nomos, S. 49–76.
- Kohn, Dieter (2013): »Biopsie«, in: Dieter Kohn/Tim Pholemann (Hg.), *Operationsatlas für die orthopädisch-unfallchirurgische Weiterbildung*, Berlin/Heidelberg: Springer, S. 151–156.
- Maté, Christian (2020): *Medizin ohne Ärzte. Ersetzt künstliche Intelligenz die menschliche Heilkunst?*, Salzburg/Wien: Residenzverlag.
- Mieg, Harald A. (2003): »The science of professional knowledge & work: Bridging sociology and cognition theory«, in: Lennart Svensson (Hg.), *Conceptual and comparative studies of Continental and Anglo-American professions (Research Report No 129)*, Göteborg: Department of Sociology-Göteborg University, S. 43–54.
- Mieg, Harald A. (2005): »Professionalisierung«, in: Felix Rauner (Hg.), *Handbuch der Berufsbildungsforschung*, Bielefeld: Bertelsmann, S. 342–349.
- Mohr, Arno (2000): *Medizinsoziologie*, München/Wien: R. Oldenbourg Verlag.
- Mohr, John W./White, Harrison C. (2008): »How to Model an Institution«, in: *Theory and Society* 37(5), S. 485–512.
- Pfeffer, Matthias (2021): »Wenn der Computer den Arzt ersetzt: Darf der ›Todes-Algorithmus‹ über Leben entscheiden?«, in: Online-Fokus vom 18.10.2021, https://www.focus.de/wissen/kuenstliche-intelligenz-in-der-medizin-der-todesalgorithmus_id_24332671.html
- Schmedt, Michael (2018): »Künstliche Intelligenz wird Ärzte nicht ersetzen«, in: *aerzteblatt.de* vom 16.11.2018, <https://www.aerzteblatt.de/nachrichten/99193/Kuenstliche-Intelligenz-wird-Aerzte-nicht-ersetzen>.
- Schmitt, Marco/Fuhse, Jan (2015): *Zur Aktualität von Harrison White. Einführung in sein Werk*, Wiesbaden: Springer.
- White, Harisson C. (1992): *Identity and Control. A Structural Theory of Social Action*, Princeton/New Jersey: Princeton University Press.

- White, Harisson C. (1995): »Network switchings and Bayesian forks: Reconstructing the social and behavioral sciences«, in: *Social Research* 62, S. 1035–1063.
- White, Harisson C. (2008): *Identity and Control. How social formations emerge*. Second edition, New Jersey: Princeton University Press.

Plattform und jetzt?

Gründungsstories digitaler Start-ups zwischen Weltverbesserung und digitaler Transformation

Tim Franke, Niklas Strüver, Sascha Zantis

1. Einleitung

»The rise of the platform as the dominant infrastructural and economic model of the social web« (Helmond 2015: 5), ist mittlerweile ein weitreichend belegtes und untersuchtes Faktum. Dabei beschränken sich Plattformen schon längst nicht mehr auf den Social Media-Bereich. Plattformen durchziehen zunehmend neue Lebensbereiche (vgl. Dolata 2015) und sind im sozialen und gesellschaftlichen Alltag gängige Technologien zur Lösung sozialer Probleme. In der Regel steht hinter modernen Plattformen eine Form Künstlicher Intelligenz (KI), die Prozesse leitet, optimiert und ermöglicht.

KI-Systeme wie Deep Learning- und Recommender-Algorithmen sind auf große Datenmengen angewiesen, um entweder in geleiteten (supervised) oder ergebnisoffenen (unsupervised) Verfahren trainiert zu werden. Der Schlüssel zum Erfolg sind bereits identifizierte oder zu identifizierende Muster, die sich in den erhobenen Datensätzen wiederholen. Je mehr Daten, je mehr Wiederholungen und je eindeutiger das Muster, desto genauer bestenfalls der Algorithmus. Im Idealfall lassen sich auf der Grundlage der erhobenen Daten Muster identifizieren oder sogar zukünftige Muster antizipieren. Welche Muster anhand welcher Daten erkannt werden können, hängt letztlich von der Plattform und den dort anfallenden Nutzer:innendaten ab. Plattformen sind entsprechend so mannigfaltig wie die Probleme, zu deren Lösung sie entworfen und aufgebaut werden.

Die Plattform hat sich somit als Lösung diverser Probleme in allen Ebenen des Internets verwurzelt. Allerdings stellt das Einrichten und Etablieren einer solchen Plattform die Entwickler:innen und Designer:innen, die Ingenieur:in-

nen und Informatiker:innen, die *Enabler* und die *Maker* vor einige Herausforderungen. Es bedarf der technologischen Infrastruktur, der Investor:innen, des Personals, der Unternehmenssitzte und so weiter. Die Gründung eines Unternehmens zur Entwicklung digitaler Plattform ist von chronischer Unsicherheit geprägt, die Gestaltungsspielräume und -zwänge mit sich bringt. Den oft netzwerkförmig aufgebauten digitalen Start-ups muss es gelingen »eine gemeinsame Narration zu etablieren« (vgl. Laux 2014: 163). Sie entwickeln sich so vom »konturlose[n] Urschleim des *Gemenges*« in eine »immer noch flüssige, aber überschaubare *Assoziation*, in der es zu einer experimentellen und konfliktgeladenen Programmierung der Struktur kommt.« (Vgl. ebd.)

Die gewebte Narration, oder auch Story, muss dabei so überzeugend sein, dass die Gründer:innen (a) sich der Unsicherheit der Unternehmensgründung aussetzen und (b) darüber hinaus bereit sind, bei schlechter oder keiner Bezahlung mehr monetäre, temporäre und kreative Ressourcen zu investieren, als sie es in traditionellen Beschäftigungsverhältnissen müssten. Sollte das geteilte Narrativ also nicht überzeugen, zerfällt die Komposition und das Start-up scheitert. Wieso setzen sich Gründer:innen digitaler Unternehmen diesem Risiko aus? Wie begründen sie ihr rastloses Entwickeln digitaler Plattformen sich selbst und ihrem sozialem Umfeld gegenüber und wie werden ihre Begründungen von außen wahrgenommen? Was macht die Start-up-Story aus? Die Frage nach normativen Motivationsgrundlagen adressiert im Sinne Boltanski/Chiapello (2003: 42f.) »den Geist des Kapitalismus [als] eine Ideologie« und fokussiert Strategien der Rechtfertigung des Engagements für den Kapitalismus, die über die »materielle Beteiligungsmotivation« hinausgeht. Sowohl die soziologische als auch die ökonomische Forschungslandschaft weist hier Unschärfen auf: Zwar erheben Umfragen wie der KfW-Gründungsmonitor turnusmäßig Motive für die Unternehmensgründung, allerdings differenzieren sie lediglich zwischen Gründungen, die als Chance verstanden, oder aus der Not heraus geboren werden (vgl. Metzger 2018: 2).

Die Vorlage für erfolgreiche Gründungsstories bilden einige wenige Galionsfiguren der digitalen Plattformökonomie, in der Gründer:innen meist eine zentrale Rolle einnehmen: Google, Facebook, Amazon, Microsoft, Apple sowie Alibaba und Tencent. Diesen Vorbildern folgend versuchen sich die (idealtypischen) digitalen Eliten im Silicon Valley daran, das nächste Facebook oder Amazon zu entwickeln (Nachtwey/Seidl 2017). Sie stehen exemplarisch für eine neue, sich herausbildende »Polis der Solution«, d.h. einer neuen Rechtfertigungsordnung. Rechtfertigungsordnungen sind Boltanski und Thévenot (2007) zufolge historisch variierende Sinn- und Legitimationsquellen, auf die

sich Subjekte berufen, um ihr Engagement im kapitalistischen System, das (weit) über die bloße Existenzsicherung hinaus geht, sich selbst und anderen gegenüber zu legitimieren.

Im folgenden Beitrag sollen die Konzepte der Story von Harrison White (2008) sowie der Polis der Solution von Nachtwey/Seidl (2017) als Werkzeuge genutzt werden, um die formationskonstituierenden Narrative von sich in der Gründung befindlichen Plattformunternehmen zu analysieren (Kap. 2). Dafür werden zunächst die theoretischen Hintergründe der Konzepte besprochen, sprich die Netzwerktheorie Harrison Whites sowie die von Boltanski und Thévenot (2007) identifizierten Rechtfertigungsordnungen. Der theoretische Referenzrahmen dient als Grundlage für die Analyse qualitativer Interviews mit Start-up-Gründer:innen, die im Rahmen des BMBF-geförderten Forschungsprojekts INDIZ erhoben und nun einer Sekundäranalyse (Kap. 3) unterzogen werden. Zur Auswertung werden acht Interviews herangezogen, die zunächst die formationsinterne Perspektive auf die Komposition des digitalen Unternehmens abbilden und aus denen sich die Stories bzw. Legitimationsgrundlagen des Start-ups und der Gründer:innen destillieren lassen (Kap. 4). Abschließend wird die Frage diskutiert, was die Stories digitaler Start-ups ausmacht und, ob sich Gründer:innen tatsächlich als »Weltverbesserungsunternehmer:innen« (Nachtwey/Seidl 2017) darstellen bzw. als solche wahrgenommen werden (Kap. 5, 6).

2. Solutionistische Rechtfertigungen als Geschichten von Identität und Kontrolle

Die von Boltanski/Thévenot (2007: 28) entwickelte sogenannte Theorie der »Rechtfertigung« stützt sich auf die empirische Untersuchung von konfliktträchtigen, tendenziell unsicheren Situationen¹ (1. Ebene), in denen soziotechnische Akteur:innen auf eine überschaubare Zahl übergeordneter

1 » [...] vielmehr versuchen wir zu zeigen, wie Personen mit Unsicherheiten umgehen und sich dabei auf Objekte stützen, um Ordnung herzustellen, und wie sie umgekehrt auch die Objekte konsolidieren, indem sie sie mit schon konstituierten Ordnungen in Verbindung bringen. Unser Unternehmen zielt damit auf einen dynamischen Realismus ab, insofern es die Konstruktionsarbeit sichtbar machen will, ohne sie zugleich auf eine lokal beschränkte Einigung über Bedeutungen reduzieren zu wollen. Aus dieser theoretischen Festlegung, für die das Handeln im Zusammenhang mit Unsicherheiten zu begreifen ist, folgt auf der Ebene der Beobachtungsmethode, die empirische For-

Formen geteilter Allgemeinheit, sogenannte Rechtfertigungsordnungen zurückgreifen (2. Ebene), um in situativen Prüfungen temporär Einigungen zu erzielen und somit soziale Ordnung herzustellen. Setzt sich eine Rechtfertigungsordnung durch, bestimmt sie als geteiltes Ordnungsprinzip den situativen Modus Operandi, stabilisiert kurzfristig soziale Ordnung gegenüber einer kontingenten Umwelt, verlangt aber auch Anpassungsleistungen von den beteiligten Identitäten (vgl. ebd.: 34).

In den Prüfungen beziehen sich die Akteur:innen auf normative Wissensbestände und Begründungszusammenhänge, die Boltanski/Thévenot identifizieren, verdichten und mit politischen Philosophien des Gemeinwohls anreichern, um sie so zu einer überschaubaren Zahl von Rechtfertigungsordnungen zu kondensieren (vgl. ebd.: 97ff., 108ff.). Zusammen mit Chiapello bettet Boltanski die Rechtfertigungsordnungen in eine weitreichendere Diagnose kapitalistischer Gesellschaften: Sie artikulieren in Anlehnung an Max Webers »Geist des Kapitalismus« sinnstiftende Quellen zur Legitimation des individuellen und kollektiven Engagements in kapitalistischen Gesellschaften, die über monetäre Anreize hinaus gehen. Mit dem Wandel dominanter Rechtfertigungsordnungen lässt sich so der Wandel kapitalistischer Gesellschaftsformationen untersuchen (vgl. Boltanski/Chiapello 2003). Nachtwey/Seidl (2017: 30) greifen die Rechtfertigungsordnungen auf und diagnostizieren im Aufkommen des Solutionismus einen erneuten Wandel des kapitalistischen zum digitalkapitalistischen Geist. Für idealtypische Solutionist:innen, die Nachtwey/Seidl etwa in den digitalen Eliten des Silicon Valley identifizieren, bemisst sich Erfolg daran, inwiefern sie mittels technologischer Innovationen zur Verbesserung sozialer Probleme beitragen (vgl. ebd.: 17).²

Der analytische Fokus auf tendenziell unsichere Situationen und das Forschungsinteresse an abstrakten, situationsübergreifend-gültigen Formationen ist auch der Ausgangspunkt von Harrison Whites Theorie von Identität und Kontrolle (TIK):

schung auf solche Momente zu konzentrieren, in denen etwas in Frage gestellt oder kritisiert wird.« (Boltanski/Thévenot 2007: 34)

- 2 Für Nachtwey & Seidl (2017: 17) vollzieht sich der Wandel auf zwei Ebenen: Mit dem Aufkommen des Solutionismus verschiebt sich einerseits das dominante Akkumulationsregime vom Finanzmarktkapitalismus auf den digitalen Kapitalismus. Gleichzeitig stellt der Solutionismus eine Reaktion auf die legitimatorische Krise eines »moralisch und politisch entfesselten Finanzmarktkapitalismus« (ebd.) dar, der seit der Finanz- und Wirtschaftskrise 2008 zunehmend in den Fokus der Sozialkritik geriet.

»Sociology has to account for chaos and normality together, and this book works toward suitable flexible framings. [...] I claim that all scopes of social processes induce themselves in some fashion as the following: Identities trigger out of events – that is to say, out of switches in surroundings – seeking control over uncertainty and thus over fellow identities. Identities build and articulate ties to other identities in network-domains, *netdoms* for short. [...] In their search for control, identities switch from *netdom* to *netdom*, and each switching is at once decoupling from somewhere and embedding into somewhere.« (White 2008: 1f.)

Geschichten (engl. stories) stellen für White (2008) eine elementare Netzwerkformation dar, die in Netzwerken erlebt und über das Netzwerk erzählt werden. Ausgehend von den konvergierenden Forschungsperspektiven von Boltanski/Thévenot (2007) und Harrison White (2008) soll im Folgenden der Solutionismus, sprich der neue digitalkapitalistische Geist, als eine besondere, ordnungsstiftende Story im Sinne Whites (2008) untersucht werden.

Das Vorgehen verspricht in dreifacher Weise fruchtbar zu sein: (1) Die netzwerktheoretische Perspektive Whites (2008) auf kulturelle Formationen findet bisher kaum Anwendung in empirischen Analysen (Karafillidis 2019). (2) Dabei erlauben es gerade die Granularität und Flexibilität der von White eingeführten kulturellen »Formationsbegriffe« (vgl. Reckwitz/Rosa 2021: 157), den idealtypischen Solutionismus zu dechiffrieren. Whites TIK soll daher als analytisches Werkzeug vorgestellt werden, um die Komposition und Kristallisation der Story vom Solutionismus herauszuarbeiten sowie implizierte retrospektive Bezugspunkte und prospektive Erwartungshaltungen offen zu legen. (3) Zwar kann kein ausführlicher Theorievergleich zwischen der Theorie der Rechtfertigungsordnungen (Boltanski/Thévenot 2007) und der TIK (White 2008) erfolgen, dennoch ergeben sich aus der vergleichenden Perspektive auf einen Untersuchungsgegenstand – solutionistische Stories unter Gründer:innen digitaler Start-up-Unternehmen – bereits Hinweise auf wesentliche Schnittmengen und Differenzen. Um diese Konvergenz zu leisten, wird zunächst der Solutionismus in der Auslegung von Morozov dargestellt. Im Anschluss wird das Instrumentarium von White vorgestellt, um die folgende Analyse vorzubereiten.

Problembhebung durch Technologieversprechen: Solutionismus nach Morozov

Für Nachtwey und Seidl (2017) ist Morozov (2013) Ausgangspunkt der Überlegungen zur solutionistischen Polis. Auch wenn Morozov in seiner Konzeption des »Solutionismus« einräumt, dass es wohl kaum ein neues Konzept ist (vgl. Morozov 2013: 14), hat sein Buch »To save everything, click here: the folly of technological solutionism« den Begriff salonfähig gemacht. Morozov sieht solutionistische Lösungsansätze überall dort, wo blind »sexy, monumental, and narrow-minded solutions« (ebd.: 6) auf komplexe Probleme angewendet werden. Er versteht daher Solutionismus als eine Einstellung, die jegliche komplexe soziale Situation als ein technologisch, rechnerisch lös- und optimierbares, wohldefiniertes Problem betrachtet. Mögliche (unerwartete) Konsequenzen werden in den meisten Fällen nicht in diesen Problemlösungsprozess einbezogen und eine Abwägung der Konsequenzen gegen das Problem, das es zu lösen gilt, ebenso wenig (vgl. ebd.: 5). Das ehemalige Motto Facebooks: *move fast and break things* illustriert diese Einstellung. So betont Morozov auch, dass es besonders die intellektuellen Einflüsse eines Internetzentrismus sind, die den modernen Solutionismus so gefährlich machen (vgl. ebd.: 16). Dennoch liegt das Hauptaugenmerk des Solutionismus für Morozov zunächst nicht auf der technischen Lösung, sondern auf der Art der rhetorischen Problematisierung. Entsprechend den Traditionen und Routinen der Professionen, sind bestimmte Arten und Weisen Probleme zu formulieren bereits ritualisiert (vgl. Agre 1997: 150f.). So lassen sich im Dunstkreis der Silicon Valley Solutionist:innen zur Eröffnung einer solchen Problematisierung rhetorische Räume der Optimierung und Objektivität identifizieren, die häufig beschworen werden, um die technologische Lösung eines vermeintlichen Problems zu legitimieren (vgl. Sadowski/Selinger 2014). Der Problematisierungsprozess (Jaton 2020: 82), ist besonders in der Machine Learning-Community des Silicon Valleys ein zentraler Aspekt, der die ange-dachten algorithmischen Lösungen als Generatoren positiver Unterschiede zu identifizieren sucht (vgl. Jaton 2021: 9). Diese Formulierungen der Probleme sind »both descriptive and prescriptive in nature and the initial diagnosis justifies and orients the technical propositions that follow« (Rieder/Skop 2021: 4). Diese Vorgehensweise erlaubt es dem »solutionistischen tool-kit« technische, algorithmische und plattform-förmige Produkte als Lösung aller Probleme zu sehen. Morozov kontrastiert hierzu, dass nicht nur nicht jedes dieser Probleme durch solche Ansätze lösbar wäre, sondern dass auch viele

dieser *Probleme* keine seien. Oftmals seien die vermeintliche Ambiguität, Ineffizienz und Intransparenz keine Probleme, sondern schützenswürdige Werte und soziale Güter (vgl. Morozov 2013: 6). Jedoch wird durch die rhetorische Rahmung als technologisch lösbares Problem jeglicher Raum für philosophische und politische Reflektion verengt und technologische Innovation wird zur notwendigen Ideologie (vgl. Sadowski 2020: 67). Hierbei ist der Aspekt des monetären Profits im Silicon Valley nicht von der Hand zu weisen: »Die Industrie stellt die von ihr herbeigeführten Veränderungen gerne so dar, als gehorchten sie einem Naturgesetz: Wenn ich oder mein Unternehmen das nicht tun, wird es jemand anderes tun.« (Daub 2021: 9)

Es ergibt sich eine Kombination aus der Logik der Notwendigkeit von Problemlösungen durch geschickte Problematisierungen und Marktrhetorik und der schnellen Durchführung von technologischen Lösungsansätzen. Morozov identifiziert in diesem Sinne eine Gefahr im Solutionismus: »[...] the quick fixes it peddles do not exist in a political vacuum. In promising almost immediate and much cheaper results, they can easily undermine support for more ambitious, more intellectually stimulating, but also more demanding reform projects.« (Morozov 2013: 8) Situieret man den Solutionismus folgerichtig im sozialen, ökonomischen und politischen Feld des Silicon Valleys, fällt auf, dass hier die »coding elite« Lösungen für Probleme entwickelt, die letztlich ihre eigene Statusgruppe gesellschaftlich und ökonomisch aufwertet (vgl. Burrell/Fourcade 2021: 217).

Start-ups stehen im Zentrum der solutionistischen Antwort auf jegliche Form der Probleme: »problems are to be dealt with via apps, sensors, and feedback loops – all provided by startups« (Morozov 2014). Um den Kreislauf zu schließen, sei erneut vor diesem Hintergrund auf die Verengung der Problemlage durch den Solutionismus hingewiesen. Probleme wie soziale oder ökonomische Ungleichheit lassen sich nicht durch Tracking ergründen und sind somit nicht Teil des Problems: »The devil doesn't wear data. Social injustices are much harder to track than the everyday lives of the individuals whose lives they affect.« (Morozov 2014) Digitale Plattformen spielen eine zentrale Rolle in der Verbreitung und Kommerzialisierung sozialer Innovationen und beeinflussen somit den gesellschaftlichen Wandel (vgl. Morozov 2022: 6). Paradoxerweise werden zusätzliche Potenziale zur Reflexivität und zu transformativem Handeln, die digitale Tools und Plattformen eröffnen mögen, genau dadurch begrenzt, dass sie sich der Verwertungslogik der Plattformen und den algorithmischen Zwängen anpassen müssen (vgl. ebd.: 7). Folgerichtig fällt es zunehmend schwerer den gesellschaftlichen Wandel nicht »als simplen Zweiklang

von Plattform und App« (vgl. ebd.: 8) zu denken. Vor diesem Hintergrund soll nun Whites TIK angeführt werden.

Stories of Control – Story-Sets und Story-Lines bei Harrison White

In der neuen Wirtschaftssoziologie hat sich laut Mützel (2007) der Ansatz etabliert, ökonomisches Handeln als sozialen Prozess in sozialen Kontexten zu verstehen. Sie beschreibt (Produktions-)Märkte im Anschluss an Harrison White als Beziehungsgefüge, auf die sich Marktteilnehmende zur gegenseitigen Beobachtung, Koordinierung und Positionierung beziehen. Märkte seien »Ergebnisse von Netzwerkstrukturen, die einer relationalen Logik der Ko-Konstruktion folgen« (2010: 90), sodass besonders die Verkettung und soziale Interaktion selbstdefinierter (wirtschaftlicher) Cliques betrachtet werden. Stories bzw. Geschichten werden auf Märkten als eine Form reziproker Beobachtungs- und Kontrollversuche beschrieben, die der temporären Ordnungsbildung dienen. Management und Marketing etwa bedienen sich an Erfolgs- und Know-How-Geschichten, welche für andere Marktteilnehmende sinnstiftend wirken können, ein bestimmtes Bild der Herausgeber:innen hervorrufen (Mützel 2009: 93) und gleichzeitig einen Eindruck von der Struktur des Marktes – mitsamt der eigenen Position – geben (Mützel 2009: 88). Geschichten lassen sich daher als Kontrollprojekte begreifen (White 1992: 9f.), mithilfe derer die (Markt-)Akteur:innen versuchen können, »die Identitäten der beteiligten Akteure im jeweiligen Kontext« (Mützel 2007: 483) mitzugestalten.

In der Netzwerktheorie Whites (2008) erfüllen Stories eine vergleichbare soziale Funktion, wie Rechtfertigungsordnungen: In sozialen Konstellationen wirken sie ordnungsstiftend und prägen die Art des Relationierens. Rechtfertigungsordnungen stellen für White einen empirischen Leitfaden dar, mit dem Stories zu identifizieren seien. Gleichzeitig fokussierten sie einen spezifischen *tie*, d.h. eine bestimmte Art der Beziehung, nämlich die des Streits (vgl. White 2008: 182f.). Einen konkreten Anknüpfungspunkt an einen spezifischen Formationsbegriff gibt White jedoch nicht Preis: Rechtfertigungsordnungen seien »close to rhetorics« (ebd.), historisch, wie »institutions« oder »styles«. Ebenso wie er seien Boltanski, Thévenot und Chiapello auf der Suche nach »story-sets that sustain instances of a type of tie in common as viable« (ebd.) und »[...] also comport with the valuation orders [...]« (ebd.). *Stories* und die damit assoziierten Begriffe der White'schen Terminologie werden von ver-

schiedensten Autor:innen diskutiert und entwickelt.³ Eine Voraussetzung der Verwendung des Story-Begriffs ist etwa, dass sich die sozialtheoretische Bedeutung des Begriffs bereits in den neunziger Jahren verschob, wie Somers (1994: 613) diagnostiziert. Statt Stories und Narrative lediglich als das deskriptive Beschreiben und Ordnen von komplexen Interaktionsketten zu verstehen, schlägt Somers vor, die ontologische Ebene von Stories einzubeziehen. Dadurch, dass Stories menschliche Erfahrungen prägen und identitätsstiftend sein können, leiten sie soziales Handeln:

»[...] stories guide action that people construct identities (however multiple and changing) by locating themselves or being located within a repertoire of emplotted stories; that ›experience‹ is constitute through narratives; that people make sense of what has happened and is happening to them by attempting or assemble or in some way to integrate this happenings within one or more narratives [...]«. (Somers 1994: 613f.)

Als alternative Explanans positioniert sie die kultursensibleren und insbesondere weniger exklusiven Stories gegenüber klassisch-sozialwissenschaftlichen Taxonomien wie Geschlecht, Klasse, oder Herkunft (vgl. ebd.: 634). Hieraus resultiert ein Story- bzw. Narrativverständnis, das – vergleichbar mit Whites – prozessual und relational angelegt ist, sodass Stories eine temporäre Stabilisierung von sozialen Netzwerken und das Selbstverständnis der darin befindlichen Identitäten sowie deren Handlungen erklären (vgl. ebd.: 616). Ontologische Stories definiert Somers wie folgt: »*Ontological* narratives are used to define who we *are*; This in turn can be a precondition for knowing what to *do*. This ›doing‹ will in turn produce new narratives and hence, new actions [...]«. (Somers 1994: 618)

White selbst bleibt in der Definition von Stories vage: »Stories include everything from the simplest line heard on the playground [...] through artful excuses and daily accounts and on through recondite nuggets of professional gossip. Stories are invoked, without hesitation, endlessly.« (White 2008: 31)

3 In dem vorliegenden Beitrag werden insbesondere die Rezeptionen von Schmitt und Fuhse (2015), von Laux (2014) und von Phillip (2017) zur näheren Bestimmung des Story-Begriffs herangezogen. Das Werk Tillys (2008) ist zwar explizit dem Story-Konzept gewidmet. Da es sich bei seiner Begriffsbestimmung um ein eher assoziatives Verständnis des Story-Begriffs handelt, welches nicht in Relation mit den anderen Begriffen der TIK gesetzt wird, wird es hier nicht weiter diskutiert.

Gleichzeitig bilden Stories eine Säule der TIK, weil sie kommunikativer Ausgangspunkt der Genese sozialer Identitäten innerhalb von Netzwerken sind (vgl. White 1995: 1037; Phillip 2017: 77). Stories verleihen der Struktur des Netzwerks Farbe⁴ und bilden das »narrative Gedächtnis einer Sozialbeziehung« (vgl. Laux 2014: 44). Damit mittels Stories sozialer Sinn und soziale Zeit ko-konstruiert werden können, bedarf es sogenannter Switchings⁵. In stets unsicheren Situationen können Identitäten auf eine Reihe optionaler Geschichten zurückgreifen, die abhängig vom Kontext, d.h. dem »[...] Thema und die Sinnebene, unter der Relationierung stattfindet [...]«, sowie den Präferenzen der Identitäten sind (vgl. Phillip 2017: 69f.). Switchings bezeichnen erstens die Festlegung auf eine dominante Story, vor dessen Hintergrund die gegenseitige Bezugnahme prozessual vollzogen wird, und zweitens geht mit der Festlegung auf eine dominante Story eine Festlegung des Netdom einher, die wiederum die beteiligten Identitäten definiert (vgl. ebd.). Netdoms bilden in der White'schen Terminologie eine zentrale Formation, die linguistisch und logisch Netzwerke (»net«) und Domänen (»domains«), also Struktur und Kultur verschmelzen. Netdoms konstituieren Netzwerke aus mehreren Identitäten. Reziproke Kontrollversuche erzeugen eine Adhäsionskraft, die Identitäten innerhalb der Netdoms zusammenhalten. Mit dem Begriff der Stories ließe sich ausgehend von einer kulturellen Perspektive auf das Netzwerk fragen, welche Art von Story dem Netdom zu Grunde liegt.

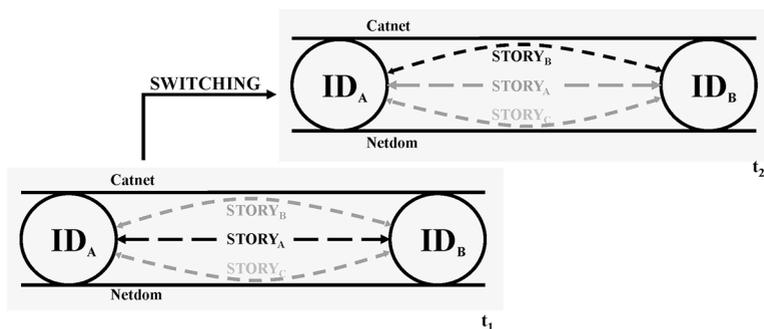
Switchings und die damit wahrgenommenen Kontextwechsel ermöglichen den Identitäten Vergleiche vorzunehmen und sich so über verschiedene Netzwerkkonstellationen und -themen selbstreflexiv zu definieren (vgl. ebd.). »Erst durch den Akt des Switchings gewinnen Akteure Agency, neue Handlungsoptionen und generieren Sinn.« (Ebd.) Dementsprechend sind Stories bzw. in der Mehrzahl Set of Stories bewährte, beobachtete, pragmatische oder antizipierte Relationen, auf die Identitäten zurückgreifen, um in Beziehung zu treten (vgl. ebd.: 73). Stories verweisen nicht ausschließlich auf erprobte,

4 »Without stories, social action would have a monotone quality; there would not be all the ›colors‹ that humans observe and use in social settings.« (Harrison C. White und Frédéric Godart, 2007, p. 4)

5 »Social networks are traces from dynamics across netdom switchings. As two identities come over time to focus control attention upon each other, a stymied struggle can settle down into some story that marks a tie between them. A story is a tie placed in context. Stories structure switchings into accounts with a beginning, middle, and the end (Tilly 2002); so story making frames social time.« (White 2008: 20)

vergangene Erfahrungen im Repertoire geglückter Bezugnahmen der beteiligten Identitäten, sondern integrieren darüber hinaus (ausgehend von den vergangenen Bezugnahmen) narrative Zukunftserwartungen, die das Switching, d.h. die Entscheidung für oder gegen eine spezifische Story oder einen Kontextwechsel, beeinflussen (vgl. ebd.: 80f.). Switchings prägen so via Stories die soziale Wahrnehmung von Zeit und verleihen der Bezugnahme Sinn, da sie unter Rückgriff auf miss- oder geglückte Relationierungen eine Kontrolle der chaotischen Umwelt und damit die Identitätskonstitution ermöglichen (vgl. Mützel 2009: 875). In der Wahl einer optionalen Story entscheiden sich die Identitäten prinzipiell für die Story, die den Identitäten die bestmögliche Identitätskonstitution und größtmögliche Kontrolle über ihre kontingente Umwelt erlaubt (vgl. ebd.: 75).⁶

Abbildung 1: Schematisches Beispiel von Harrison White's Begriffen der »story« und des »switching«. Die Darstellung ist angelehnt an Phillip (2017: 71).



Wenn es zur Operationalisierung von Stories bzw. Switchings kommt, sieht Phillip primär zwei Hürden: (a) Erstens sind Switchings aus der Beobachter:innenperspektive schwer zu erfassen, weil neben situativen Kontexten individuelle Prioris »als individuelles Vorwissen über bestehende Stories« einbezogen werden müssen (Phillip 2017: 77). (b) Aus dem gleichen Grund »kann

6 Die situationsinterne Entscheidung für oder gegen eine Story ist hier verkürzt dargestellt. White fügt hier das Modell der Bayesian Forks ein, welches es erlaubt, die Entscheidungsfindung für oder gegen eine Story bzw. gegenseitige Bezugnahme statistisch und unter der Berücksichtigung des bestehenden Wissens der Identitäten zu modellieren (vgl. Phillip 2017: 74; White 1995).

[es] deshalb keine übergreifenden, immer und überall gültigen Gesetze des Sozialen geben« (vgl. ebd.). Nicht zuletzt, weil es sich bei Stories um höchst dynamische Konzepte handelt, die permanenten Switchings unterliegen. Die Suche nach übergreifenden Metastories⁷, die einen vergleichbar universellen Gültigkeitsanspruch wie die von Boltanski/Thévenot (2007) identifizierten Rechtfertigungsordnungen aufweisen, scheint von vornherein an der Situativität und Subjektivität von Stories zu scheitern. Es stellt sich die Frage, ob und wie Stories überhaupt situationsübergreifend miteinander verglichen werden können.

Im Anschluss an White (2008) und Phillip (2017) lassen sich zunächst zwei Anhaltspunkte formulieren, mittels derer vergleichbare Stories identifiziert werden können: Stories werden bisher als kleinstes narratives Plasma verstanden, in welches Identitäten und ihre Relationen eingebettet sind. Ihre Verortung auf einem eher niedrigen Skalen- bzw. Aggregationsniveau (über-)betont die Varianz verschiedener Stories bzw. die Besonderheit einzelner Stories. Um vergleichbare Stories zu identifizieren, bietet es sich an, zunächst nach höherskaligen kulturellen Formationen Ausschau zu halten – etwa nach Netdoms⁸ als narrative Kontexte. Phillip zufolge ist es denkbar, dass verschiedene Netdoms auf vergleichbare Story-Sets zurückgreifen, sich jedoch darin unterscheiden, wie (oft) wann welche Story zur Relationierung herangezogen wird (vgl. Phillip 2017: 73).

Von zufälligen Konstellationen lassen sich storybasierte Netzwerke außerdem auf der temporalen Ebene unterscheiden. Situative Bezugnahmen zweier oder mehrerer Identitäten verdichten sich über die Zeit, indem sie von den Identitäten aufgegriffen werden, sich in Storys über das Verhältnis der involvierten Identitäten verdichten und somit eine relative Halbwertszeit gewinnen. Sie bilden vergangene Interaktionen ab, die gegenwärtige Interaktionen prägen und eine Erwartungshaltung gegenüber zukünftigen Interaktionen erlauben (vgl. Schmitt/Fuhse 2015: 96). Dass Beziehungen situationsübergreifend anschlussfähig sind, setzt voraus, dass sie generalisiert und in einer abstrakteren Beziehungsbeschreibung verdichtet werden, die dann wiederum für

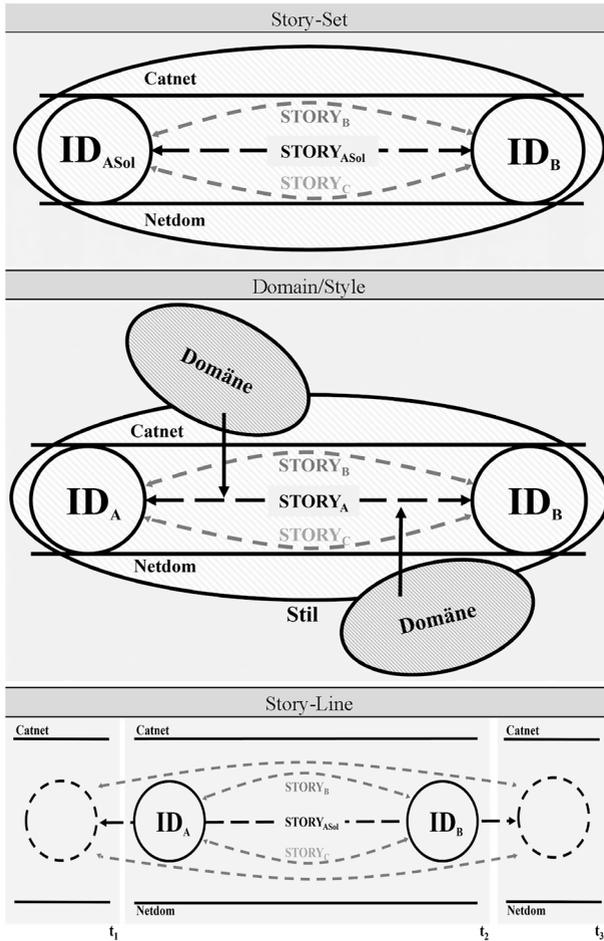
7 Auch Somers beschreibt »Metanarrativity« als »epische Dramen unserer Zeit«, die zwar subtil die Wissensgenerierung prägen, allerdings jenseits der bewussten Wahrnehmung (Somers 1994: 619).

8 »[Netdoms, T.F.] sind eine Kombination aus technischer Vermittlung, Wahrnehmbarkeit der in ihnen erzählten Stories und inhaltlicher Orientierung dort stattfindender Relationierung an spezifischen Themen.« (Phillip 2017: 70)

wissenschaftliche Beobachter:innen zugänglich sind, mitunter sogar von ihnen produziert werden, denn auch Beobachter:innen können Beziehungskonstellationen identifizieren und in Stories über die Art der Beziehung verdichten (vgl. ebd.: 98).

White, der sich nicht nur für die »Entstehungs- und Reproduktionsbedingungen des Sozialen« interessiert, fragt gleichermaßen nach dessen »Verfestigung und Verknöcherung«, »[...] also nach den Ursachen für die besondere Alchemie und Stabilität der Netzwerke.« (Laux 2014: 41). Um die Stabilität der Netzwerke und Geschichten aus der kulturellen Perspektive in den Blick nehmen zu können, bietet die White'sche Terminologie eine ganze Reihe an Formationsbegriffen, Stories anhand (a) ihrer Komposition, (b) ihrem Kontext und (c) ihrer pro- und retrospektiven Referenzpunkte zu differenzieren. (a) Wenn Stories parallel zueinander auftreten, sich demnach mehrere Stories innerhalb eines Netzdoms bewegen, spricht White von »set of stories«. (b) Sich zeitlich aneinanderreihende Stories bezeichnet er als »story-lines«, die meistens im Plural erscheinen, konstitutiv sind für (ein Verständnis von) soziale(r) Zeit und sich wiederum in unterschiedliche Set of Stories differenzieren lassen. Sie verweisen auf Positionen im Netzwerk/in Netzwerken, die Identitäten eingenommen haben und bilden somit einen narrativen Pfad durch das Netzwerk (vgl. White/Godart 2007: 8ff.). (c) Stories, Story-Lines und Story-Sets können aushärten: Faustregeln/Rules of Thumb etwa sind robust gegenüber verschiedenen Netzwerkkontexten und lassen sich daher in diversen Situationen anwenden. Sie können umgekehrt für die jeweilige Situation ein spezifisches Story-Set liefern (vgl. ebd.: 6).

Abbildung 2: Schematisches Beispiel von Harrison White's Begriffen des »story-sets«, »domain/style« und der »story-line«. Die Darstellung ist angelehnt an die von Phillip (2017: 71).



Stories können einerseits eine individuelle, eine spezifische oder einzigartige Beziehungskonstellation explizieren, d.h. so in kulturellen Sinn übersetzen, dass sie nachvollziehbar und interpretierbar ist; sie können darüber hinaus generalisiert werden. Die Generalisierung einer Story verleiht ihr (temporäre) Stabilität, gibt ihr eine abstraktere soziale Bedeutung, die sich auf ver-

gleichbare Beziehungskonstellationen anwenden lässt und somit *sozial bedeutsam* und über ihren spezifischen Kontext hinaus verfügbar werden lässt (vgl. ebd.). Stories verfestigen und verdichten sich in Domänen und Institutionen. *Domänen* werden hier verstanden als »[...] Ensemble an geteilten Sinnformen in einem abgrenzbaren sozialen Kontext.« (Vgl. Schmitt/Fuhse 2015: 108) Stories im Sinne Whites sind somit relational zu verstehen: Ihr ordnungs- und identitätsstiftendes Potenzial entfaltet sich durch den Verweis auf vorherige, zukünftige oder parallele Stories. Die Erzählung über eine Situation oder einen Prozess wird dann plausibel und anschlussfähig, wenn sie in die Form eines Narratives gebracht wird, das mit anderen Erzählungen vergleichbar ist. Der Solutionismus grenzt sich von anderen Narrativen insofern ab, dass er gegenwärtige oder zukünftige *technologische* Lösungsansätze für soziale Problemfelder gegenüber rein *politischen* oder *ökonomischen* Ansätzen priorisiert. So wird z. B. die Frage nach der Kompetenz in der Bedienung der Devices zu einer Frage des technologischen Designs (Apple) oder die Frage nach sozialer Teilhabe zu einer der Konnektivität (Facebook).

3. Sekundäranalytische Identifikation von Stories

Die zu Grunde liegenden Interviews (N=38) wurden im Kontext des BMBF-geförderten Forschungsprojektes »INDIZ – Innovationen digitaler Plattformsysteme für die ethische, rechtliche und soziale Zukunftsgestaltung« [FKZ: 16ITA206] zwischen 2019 und 2020 erhoben. Die Proband:innen wurden in verschiedene Anspruchsgruppen⁹ kategorisiert, um die je eigenen Perspektiven auf digitale Plattformen herauszustellen. Der starke thematische Bezug zu digitalen Plattformen prädestiniert den Datensatz für eine Sekundäranalyse solutionistischer Perspektiven.¹⁰ Durch die thematische und prozessuale

9 Im Forschungsprojekt wurden insgesamt vier Anspruchsgruppen differenziert: Gründer:innen [n=10], Expert:innen [n=13], Mitarbeiter:innen digitaler Plattformunternehmen [n=6] und Bürger:innen [n=9]. Die Zahl der interviewten Personen entspricht nicht der Zahl der geführten Interviews, da z.T. Interviews mit zwei Personen geführt wurden.

10 Da einer der Autoren bei der primären Datenerhebung und -bearbeitung beteiligt war, kann auf Kontextwissen bezüglich Erhebung, Datensatz und Forschungsziel zurückgegriffen werden, sodass die Originaldaten nicht zuwider ihrer Intention bzw. Erhebungskontext verwendet werden.

Nähe können die Daten also im Sinne einer neuen Fragestellung rekontextualisiert werden (vgl. Medjedović 2014: 226ff.), auch wenn die Identifikation der »Polis der Solution« nicht das ursprüngliche Ziel der Datenerhebung darstellte.

Anschließend an die in Kapitel 2 dargestellten Ansätze zur Polis der Solution soll das Datenmaterial mittels der von Harrison White formulierten Formationsbegriffe der Netzwerktheorie untersucht werden. Hierzu wird ein qualitativ-inhaltsanalytischer Ansatz verwendet, um den Argumentationslinien inhärente Stories, Story-Sets, Domänen und Stile sowie Story-Lines zu extrahieren. Die Selbstdarstellungslogiken digitaler Eliten, auf die Seidl durch statistische und quantitative Analysen solutionistischer Einstellungen schließt (vgl. 2021: 192ff.), sollen hier aus den Blickwinkeln der qualitativen Forschung beleuchtet werden. Es wird eine qualitative inhaltliche Strukturierung durchgeführt (vgl. Mayring 2015: 67, 97ff., 103ff.), die die Interviews sequenziell auf solutionistische Argumentationslogiken untersucht. Diese wird genutzt, um das »Material zu bestimmten Themen, zu bestimmten Inhaltsbereichen [zu] extrahieren und zusammen[zul]fassen« (ebd.: 99). Die Analyse wurde hauptsächlich von deduktiv generierten Schlüsselkategorien geleitet, die anhand der Literatur zum Solutionismus entwickelt wurden und den Fokus der Analyse darstellen (vgl. Schreier 2012: 59).

Relevanten Textsegmenten wurden Codes in MaxQDA zugewiesen. Die kleinste zu kodierende Einheit waren einzelne Sätze, wobei darauf geachtet wurde, dass der Kontext der kodierten Segmente auch außerhalb des Volltexts nachvollziehbar blieb (Mayring 2015: 61, 115ff.). Der Fokus der Analyse liegt auf den zehn interviewten Start-up-Gründer:innen, da ihnen – ausgehend von den theoretischen Vorüberlegungen – am ehesten solutionistische Denk- und Artikulationsmuster unterstellt werden können.

Im Forschungsdesign werden implizite Annahmen getroffen, die im Anschluss an das White'sche Vokabular der Erklärung bedürfen. Erstens wird im Anschluss an Morozov (2013) vorausgesetzt, dass der Solutionismus ein zeitgenössisches Story-Set ist, auf welches insbesondere innerhalb eines spezifischen Milieus zurückgegriffen wird. Der Archetyp dieser Geisteshaltung findet sich im Silicon Valley (vgl. Nachtwey/Seidl 2017: 17). Den Vorarbeiten zur »Polis der Solution« (Nachtwey/Seidl 2017; Seidl 2021) folgend, wird angenommen, dass sich digitalaffine Milieus als kulturelles *netdom* und lokales *catnet* auch außerhalb des Silicon Valleys wiederfinden lassen. Als vergleichbare *netdoms* werden Stories-Sets verstanden, auf die Solutionist:innen zurückgreifen. Sie weisen Schnittmengen auf, die sich als solutionistische Narrative identifi-

zieren lassen, gehen jedoch nicht im Solutionismus auf, weil es sich hierbei in der Theorie um ein idealtypisches Konstrukt handelt. Die Wahrscheinlichkeit, dass die digitalen Eliten weltweit auf vergleichbare Story-Sets zurückgreifen steigt dadurch, dass es sich bei digitalen Eliten um relativ homogene Gruppen in Bezug auf die Zugehörigkeit zu bestimmten Alterskohorten oder sozialen Schichten handelt. Die Analyse von Stories begibt sich in ein dynamisches Wechselspiel zwischen situativ gültigen Narrativen, die permanenten Switchings unterworfen sind und reziprok die Prioris der Proband:innen prägen, aber auch durch sie geprägt werden.

Häußling (2006: 126ff.) zufolge ist es möglich, innerhalb von sozialen Netzwerken den analytischen Fokus auf die Motive der beteiligten Identitäten zu legen und so den Wandel von Netzwerken zu beschreiben: Somit ist es möglich »[...] die Sicht der einzelnen Akteure mit derjenigen zu verknüpfen, die sich ergibt, wenn man zur Emergenzebene der Interaktionen mit deren Regeln, Ressourcen und Eigendynamiken wechselt.« (Ebd.) Individuelle Motive, die für Solutionist:innen handlungsleitend sind, wenn sie in gesellschaftliche und soziale Prozesse eingreifen, treten so zu Tage. Der methodologische Zugang beschränkt sich also nicht auf die deskriptive Rekonstruktion der strukturellen Position(en) der betrachteten Identität(en), sondern spitzt diese zu (vgl. ebd.: 128f.).

»Vielmehr wird die Position von den Akteuren individuell angeeignet und motivational angereichert, ja regelrecht zugespitzt. Die Motive können [...] darin bestehen, die eigene Position im Relationsgefüge zu anderen Akteuren zu festigen, Machtmittel zu akkumulieren, aktives Networking zu betreiben oder eine Positionsveränderung anzustreben.« (Vgl. ebd.: 128f.)

Die Begriffe *stories*, *storysets*, *domain/style* und *story-line* werden als »analytisches Werkzeug« (Reckwitz/Rosa 2021: 47f.) verwendet¹¹. Die ordnungsstiftende Kohärenz solutionistischer Couleur wird so dekonstruiert, dass differente und

11 Fuhse/Schmitt legen nahe, dass Whites Theorie nach dem »Baukastenprinzip« (Fuhse/Schmitt 2015: 3ff.) aufgebaut sei, sodass sich Begriffe separat anwenden lassen, besondere Anschlussfähigkeit zur empirischen Netzwerkforschung besteht und, dass sie mit Theorien mittlerer Reichweite kombinierbar sei (vgl. Fuhse/Schmitt 2015: 2ff.). Die Metapher von Theorien als »Werkzeug« ist kein Novum, sondern erscheint bereits 1975 bei Foucault, der seine Theorien als »kleine Werkzeugkisten« bezeichnet (vgl. Foucault 2002: 887f.).

kohäsive Story-Sets und Story-Lines sichtbar werden. Für die analytische Dekonstruktion des Solutionismus werden drei nicht trennscharfe, strukturgebende Dimensionen festgelegt, die sich an den Begriffen (a) »story-set«, (b) »domain & style« sowie (c) »story-line« und »narrative« orientieren und jeweils spezifische Perspektiven auf den Untersuchungsgegenstand eröffnen.

4. KI und Solutionismus: Geschichten digitaler Start-ups

Solutionistische Unternehmer:innen werden charakterisiert durch eine bestimmte Perspektive auf Erfolg als Ausdruck des Beitrags, der zur Verbesserung der Welt geleistet wurde.

»At the heart of this beliefs is the idea that all good things go together, that one can make money while making the world a better place. [...] Much like the early protestants believed, that economic success is a sign of chosenness, the solutionist entrepreneurs are convinced that if they are doing good, they will also do well; and conversly, that if they are doing well, they must also be doing good.« (Seidl 2021: 173)

Die White'schen Stories unterscheiden sich von der solutionistischen Polis, da sie nicht an eine normative Rechtfertigungsordnung gebunden sind, sondern sich in erster Linie als analytisches Tool eignen, mit dessen Hilfe der Sinn beschrieben werden kann, welchen die vernetzten Identitäten ihren Beziehungen zuschreiben. Der zugeschriebene Sinn kumuliert in Geschichten über die Beziehung und legt fest, »[...] welche Art von Kontrollversuchen darin zu finden und vorgesehen sind.« (Fuhse/Schmitt 2015: 97f.)

Die Gründer:innen sind in der Regel aktiv eingebunden in die Entwicklung einer technischen Lösung, dementsprechend ist ihnen – ausgehend von ihrer Position im Netzwerk – eine Argumentationsstruktur solutionistischer Prägung zuzuschreiben. Da es sich bei den von ihnen entwickelten Technologien um digitale Plattformen oder Software handelt, stehen sie indirekt mit den technologischen Identitäten und indirekt auch mit Nutzer:innen in Relation. Die folgende Analyse orientiert sich an den in Kapitel 3 herausgearbeiteten Analysekategorien und identifiziert mit der a) Rolle von KI in der Gesellschaft und b) dem Weltverbesserungsnarrativ, zunächst zwei essenzielle Story-Sets. Anschließend werden Domänen, Stile und Story-Lines spezifiziert.

Zur Rolle von digitalen Plattformen und KI in Gründungsprozessen

Auch wenn sowohl Boltanski/Thévenot (2007: 39) als auch White (2008: 26) technologischen Identitäten eine ordnungsbildende Funktion einräumen, ist bei White nicht abschließend geklärt, welche Rolle sie in Netzwerken einnehmen können. Entgegen der einschlägigen Verwendung der TIK als anthropozentrische Theorie, in der nur menschliche Akteur:innen Stories erzählen können (vgl. White 2008: 346; Philip 2017: 78)¹², wird mit Blick auf digitale Technologien aus zwei Gründen widersprochen: (1) Das Überführen von Geschichten in digitale Relationen ist wesentlicher Teil des Geschäftsmodells digitaler Plattformen, die wiederum die Art und Weise prägen, in der Geschichten erzählt und wahrgenommen werden. Digitale Plattformen sind somit soziotechnische Identitäten, die in den Prozess der Storygenese involviert sind. (2) Insbesondere KI-Systeme sind aktiv in die Produktion von Stories involviert, etwa wenn sie konkret Geschichten schreiben (vgl. Köbis/Mossing 2021) oder aber die Kategorisierung menschlicher Akteur:innen beeinflusst. Digitale Plattformen werden außerdem als »Extraktionsarchitekturen«¹³ (Vogl 2021: 79) verstanden, mittels derer Digitalunternehmen Daten gewinnen, die sie in anschließenden Arbeitsschritten in Informationen aufbereiten und zu neuen Stories verdichten können: »Die Intelligenz einer Maschine richtet sich danach, wie viele Daten sie frisst.« (Zuboff 2018: 118f.)

Die Relevanz von Daten in informationskapitalistischen Geschäftsmodellen ist den Gründer:innen entsprechend bewusst. Ein Proband bildet eine Analogie zwischen Amazons Angebot an Filmen und traditionellen Videotheken:

»[...] so eine Plattform lebt ja in der Regel davon, dass die halt Daten hat, die man in einem klassischen Modell nicht bekommt. Also statt jetzt ein Nutzer, der da jetzt irgendwie durchgeht und am Ende eine DVD rauspickt,

12 Tilly, der das Story-Konzept Whites geprägt hat, relativiert, dass menschliche Akteur:innen in Geschichten meist priorisiert werden, nicht-menschlichen Akteur:innen jedoch ähnliche Attribute zugeschrieben werden (vgl. Tilly 2006: 65).

13 »Dabei geht es zunächst darum, die durch das Online-Verhalten hinterlassenen Informationsspuren so zu quantifizieren, zu aggregieren, zu filtern, zu analysieren und zu transformieren, dass aus Vorlieben, Suchanfragen, Lüsten, Selbstdarstellungen, Querelen, Intimitäten oder Sozialkontakten überhaupt Muster und aus diesen Mustern Produkte oder Waren zum Verkauf an Kunden aus der Werbebranche gewonnen werden, die sich insbesondere für individuell und mikrologisch adressierbare Kampagnen interessieren.« (Vogl 2021: 78)

sieht Amazon ja genau welche DVD ich vorher schon durchgeblickt habe. Das können die wiederum nehmen und andere empfehlen. Und die können das besser machen, je mehr Daten die haben. Das heißt je mehr Nutzer die haben, desto besser sind halt solche, solche Effekte. Und ich glaube, das sorgt dafür, dass in diesem ganzen Plattformgeschäft immer der gewinnt, der am meisten Daten hat oder halt die meisten Nutzer hat.« (I2)

Das Story-Set besteht hier aus der Gegenüberstellung von analogen und digitalen Vorgehensweisen, wobei analoges Vorgehen – aus heutiger Perspektive – eigentümlich und vor allem unpraktisch bzw. ineffizient erscheint. Die Gründer:innen sind sich einig, dass maschinelles Lernen und KI-Methoden (neben der Blockchain-Technologie) eine zentrale Rolle einnehmen werden (vgl. I1; I7). Den Anwendungsmöglichkeiten für KI stehen sie aus mehreren Gründen skeptisch gegenüber. Ein Proband erklärt das KI-basierte Geschäftsmodell Amazons und ergänzt, dass die Algorithmen zur Aggregation von Attributen zu Nutzer:innenprofilen nach wie vor fehleranfällig sind, Learner aber gleichzeitig die Lösung des Problems sein können, sofern sie falsche Attributionen identifizieren.

»[...] das heißt, KI-gesteuerte Auswertungen von Nutzerprofilen, die pseudo-anonymisiert worden sind, wo Personas zusammengefasst werden und dann aufgrund der Persona-Zugehörigkeit Empfehlungen für den nächsten Arbeitsplatz äh vielleicht gemacht werden. Von daher sehe ich schon, dass KI ähm eine technologische Sache ist, die Plattformen beeinflussen wird. Wir sehen es eben heute relativ statisch von Amazon äh, »Kunden, die das gekauft haben, haben auch das gekauft«. Aber Amazon nimmt heute nicht Rücksicht darauf, wenn ich ein Geschenk gekauft habe, sondern dann ordnet Amazon mir das Geschenk als mein Interesse zu und empfiehlt mir dann trotzdem weiterhin Sachen, die eigentlich gar nicht in mein Profil reinpassen. Ähm das ist zurzeit ein Nachteil von Amazon, aber ich glaube schon, dass man da über eine KI dann eben sieht, ok, das passt jetzt überhaupt nicht in den, in das Einkaufsverhalten raus.« (I7)

Als Story-Set wird KI mit den führenden Tech-Unternehmen in Verbindung gebracht, um Nähe zu diesen zu suggerieren. Ebenso werden Schwierigkeiten beim Einsatz von KI, auch und besonders bei diesen Firmen, aufgezeigt, um mögliche eigene Probleme zu relativieren. Die Anwendung von KI-Verfahren gewinnt an Attraktivität, wenn eine Verarbeitung der Datenmengen mit klassischen Methoden nicht mehr gewährleistet ist.

»Wenn ich das beides [Datenmengen und die Dysfunktionalität klassischer Verfahren, T.F.] nicht habe, wenn ich Bereiche habe, wo ich weiß, okay, da sind Menschen immer noch effizienter, weil es zum Beispiel einen großen empathischen Aspekt aufweist oder auch Erfahrungswerte mit sich bringt, die wir brauchen, um effektiv arbeiten zu können, dann ist es relativ sinnlos.« (I1)

Der Einsatz von maschinellem Lernen im Gründungskontext will durchdacht sein, weil sowohl Expert:innen zur Anwendung als auch die Hardware-Komponenten kostspielig sind. Dieser Drahtseilakt lässt sich erneut mit den KI-Strategien der großen Plattformunternehmen vergleichen. Start-ups agieren häufig in technologischen Nischen und entwickeln »new, artisanal AI services« (Srnicek 2022: 243), für konkrete Problemsituationen, die von den großen Firmen nicht bespielt werden. In dieser Situation der simultanen Rechtfertigung und Abgrenzung zeigt sich jedoch ebenfalls die relative technische Nähe (vgl. Pfeiffer 2021: 232) der KI-Verfahren von großen Unternehmen und Start-ups. So lässt sich im Start-up-Bereich eher von einem inkrementellen Wechsel durch die Erprobung von Teillösungen sprechen (vgl. Pongratz 2022: 150; Hepp 2022: 238).

KI lässt sich nicht nur als Technologie, sondern auch als narrativer Hype verstehen, der als Story interpretiert werden kann, in der »[...] die gefühlten Auswirkungen größer sind, als die realen Auswirkungen.« (I1) Die Anziehungskraft von Hypes besteht darin, dass sie nicht in Gänze entschlüsselt sind, jedoch in einigen Bereichen substanziellen Mehrwert bieten. Sie berufen sich auf zukünftige Potenziale, obwohl sie in anderen Bereichen sehr viel Zeit und Geld verschlingen, ohne einen Mehrwert zu generieren. Ein Proband beschreibt, dass die mit KI in Verbindung stehenden Stories weit entfernt sind von dem aktuellen technologischen Entwicklungsstand. KI wird als Label genutzt, damit sich das Start-up vor den Investor:innen rechtfertigen kann.

»Also momentan ist ja/, ich weiß nicht, ob es überhaupt noch irgendein digitales Start-up gibt, das nicht irgendwo erzählt, sie machen auch irgendwas mit KI. Da kriege ich schon jedes Mal einen Schreikrampf, wenn ich das höre, vor allem, wenn man halt weiß, wo wir da aktuell technisch stehen, es ist viel weiter hinten, als es die meisten Leute eigentlich immer glauben, [...] aber trotzdem glaubt jeder heute, jedes kleine Start-up, jede Plattform: Ich mache jetzt noch etwas mit KI [...] und dann generiere ich einen substantiellen Mehrwert für meinen, ja, vielleicht für meinen Investor, da-

mit der das Gefühl hat, die sind voll cool, aber für mehr halt auch nicht.«
(17)

Das investor:innenorientierte Denken zeigt, wie Start-ups KI-Stories strategisch als Versprechen einsetzen können, um Gelder für die geplanten Innovationsprojekte zu sichern (vgl. Van Lente/Rip 1998). »Stories influence the confidence of investors that markets will develop in a certain direction and thereby influence investment decisions. [...] Fictional expectations provide justifications for investment decisions whose success is uncertain.« (Beckert 2013: 228) Die Legitimation einer Idee im Gründungsprozess ist essenziell für die Konstruktion der Story (Pongratz et al. 2014: 417), für Plattform-Start-ups dient KI so als narratives Zugpferd. Dass die Gründer:innen dem Einsatz von maschinellem Lernen in der Praxis skeptisch gegenüberstehen, erscheint daher plausibel. Durch die inflationäre Verwendung des Begriffs zu Marketing- und Werbungszwecken, werden gründliche Überlegungen zur Notwendigkeit und Angemessenheit der KI-Lösungen häufig vernachlässigt. KI als Marketing-Finte wiederum steht im Widerspruch zum fachlichen Selbstverständnis der Gründer:innen und birgt gegebenenfalls das Risiko eines nachhaltigen Reputationsverlustes. Der Frage nach der Sinnhaftigkeit des Einsatzes von Verfahren maschinellen Lernens begegnen sie differenzierend: Ist der Einsatz vor dem Hintergrund des Geschäftsmodell, der vorhandenen Daten, der monetären Ressourcen und fachlichen Kompetenzen zielführend? Welches Problem soll mittels KI bearbeitet werden? Erst das Problemfeld bestimmt, ob der Einsatz maschineller Lerner ergiebig erscheint, wobei die Einschätzung gegenüber den Problemräumen nüchtern ausfällt. Nicht alle Problemräume können mittels maschineller Lerner sinnvoll bearbeitet werden:

»Aus fachtechnischer Sicht heißt es dann/, sind es verschiedene Ansätze Maschine Learning, Supervise and Supervise Learning, verschiedene Ansätze, was dann auch gerne mal als KI zusammengefasst wird. Ich glaube das Potenzial ist auf jeden Fall da, man sollte nicht unbedingt versuchen jedes Problem jetzt damit zu lösen, also es ist nicht der universal Hammer, mit dem alles funktioniert, sondern es gibt halt gewisse Probleme die funktionieren besser, andere funktionieren schlechter. [...] Ich finde immer man sollte da ein bisschen betrachten, was ist der Ausgangszustand und wo kann man potenziell hinkommen.« (I2, Proband 2)

Neben Fragen der Angemessenheit werden auch gesellschaftliche Implikationen reflektiert. Ein Hype um Technologien wie diese ist nichts Neues und auch mit der Gefahr eines plötzlichen Zusammenbruchs besonders vor dem Hintergrund des Platzens der Dotcom-Blase verbunden (vgl. Sadowski 2022: 236). Ein Proband prognostiziert, dass der Hype um die Technologien in einem Boom im Ausbildungsbereich mündet, der – sofern sich der Hype nicht bestätigt – dazu führt, dass im Bereich der KI mehr ausgebildetes Personal zur Verfügung steht, als gesellschaftlich gebraucht werde (vgl. I1). Somit wird auch ein Scheitern von KI-Technologien bzw. eine Entwicklung, die ggf. nicht den Erwartungen des (Arbeits-)Markts entspricht, antizipiert. Die Konsequenzen von Hype-Stories gehen daher weit über das Sichern von Geldern und den Gründungsprozess hinaus.

Vor dem Hintergrund einiger durchaus selbstkritischen Betrachtungen der Start-up-Szene, lassen sich KI-Technologien als facettenreiches Set an Stories betrachten. KI-Stories stehen nicht für sich allein, sondern werden stets mit anderen Kontexten zu komplexen Story-Sets verwoben, die letztlich Kontrollprojekte zu verbergen suchen (vgl. White/Godart 2007: 6). Die Anwendung von KI variiert je nach kontextualisierendem Netdom: Ist die individuelle Expertise bzw. Einschätzung einzelner Gründer:innen gefragt? Steht die Verwendung des Begriffs in der Branche zur Diskussion? Oder wird nach einer generellen Einschätzung von KI gefragt? KI wird daher als Story-Set wahrgenommen, dass abhängig vom Kontext anders interpretiert und erzählt wird.

Story-Set: Das Weltverbesserungsunternehmertum

Um die Komposition der angeführten Stories abzubilden, gilt es individuelle Motive zu dechiffrieren, sodass Differenzen zwischen *Story-Sets* offengelegt werden. Story-sets legen einen Schwerpunkt auf die Kombination und Rekombination solutionistischer Argumente. Die Gründer:innen beziehen sich in unterschiedlichem Maße auf die von ihnen identifizierten gesellschaftlichen Probleme, um die Motivation zur Entwicklung einer digitalen Lösung zu fundieren. Sie greifen soziale Probleme auf und verweben sie in einem solutionistischen Story-Set. Die Varianz reicht von pragmatischen Hürden bis hin zu (gesamt-)gesellschaftlichen Problemen. Ein eher nüchternes Ziel der Plattform wäre, allgemeine Alltagsprobleme der Nutzer:innen zu adressieren und ihnen das Leben zu vereinfachen, während ein kostengünstiger Mehrwert geboten wird (I1):

»Also ich glaube wir haben es relativ leicht einen gesellschaftlichen Mehrwert aufzuzeigen, weil wir Menschen Bildungsformate geben, die vielleicht vorher keinen Zugang zu diesen Bildungsformaten hatten, weil es halt einfach nicht kosteneffizient war [...]. Und jetzt machen wir halt aus dieser Präsenzveranstaltung ein digitales Lernformat und wir bieten es ganz vielen Menschen an. Also selbst unabhängig von Monetarisierung und wie zugänglich das ist, in dem Moment, wenn es existiert, ist es schonmal schon Mal zugänglicher, als wenn es eine Präsenzveranstaltung gewesen wäre. Also den gesellschaftlichen Mehrwert sehe ich vor allem darin, dass wir Bildungsformate skalieren können.« (12, Proband 1)

Die Problematisierung niedrigschwelliger Bildungsangebote und des lebenslangen Lernens als soziales und gesellschaftliches Problem sind hier das narrative Fundament des Netdoms. Wie häufig in Entstehungsgeschichten von Plattformunternehmen zu erkennen, sind die persönlichen Erfahrungen der Gründer:innen (vgl. Bucher 2021: 35–39) essenzielle Ausgangspunkte für die Ausrichtung der Plattformfunktionen: Eine geschickte Problematisierung (vgl. Morozov 2013: 8) der Impressionen des Stuttgarter Stadtverkehrs werden so narrativer Bezugspunkt für die Gründer:innen, ein Produkt mit technologischem Mehrwert, sowie gesellschaftlichen, sozialen und Umweltaspekten zu entwickeln.

»Und das ganze Thema Stau und pendeln und zusammen pendeln und so, das ist mir zum ersten Mal bewusst geworden als ich in Stuttgart gelebt habe. [...] Weil hier in NRW [...] da ist das mir nicht so ins Auge gestoßen. Und da habe ich es schon gemerkt, Stuttgart ist natürlich [...] schon krass. Gerade mit den wenig Zubringerautobahnen. Und das war 2015, das ist schon lange her. Aber das ist mir lange im Kopf geblieben.« (13, Proband 1)

Trotz moderater Darstellung des Transformationspotenzials ihrer Plattform und dem lokalen Ausgangspunkt der Story, verdeutlichen dieselben Gründer:innen eine radikale Erwartungshaltung gegenüber der gesellschaftlichen Transformation:

»Alles wird sich verändern. Ja, alles MUSS sich verändern. Also, meiner Meinung nach und ich glaube, dass das auch ziemlich destruktiv kommen wird. [...] ich glaube, gerade für Städter wird der Privat-Pkw-Besitz unheimlich unwichtig. [...] Und das wird der Moment sein, wo es, glaube ich, alles ändern wird. [...] in dem Zeitpunkt, wo es günstiger ist [autonome Fahrzeuge zu nut-

zen N.St], ich glaube, da wird sich alles ändern. Auch Auswirkungen auf, wie die Leute leben. Ob die Leute vielleicht nicht in der Stadt leben, die Kinder müssen nicht von den Eltern nicht mehr außerhalb der Stadt zur Schule gebracht werden, sondern können sich einfach in ein Fahrzeug reinstellen/setzen, auch wenn das weiter weg ist. Also es wird halt ab dem Zeitpunkt auch wirklich ein soziale [...] Veränderung haben.« (I3, Proband 1)

Noch deutlicher wird ein Start-up-Unternehmen, das im Sinne der »New Work« freie Bürokapazitäten möglichst flexibel vermieten möchte. Auch dieses Start-up stilisiert ihre Plattform als Teil eines gesellschaftlichen Wandels, der unausweichlich ist: »Ich glaube schon, dass es ein Kulturwandel geben wird, Richtung New Work, wo einfach auch nachkommende Generationen kommen und sagen: ›Ich möchte nicht mehr zwischen neun und 17 Uhr arbeiten, sondern dann, wenn ich Zeit und Lust dazu habe!« (I7)

Die Antizipation von großflächiger Veränderung und die Haltung, Teil des Transformationsprozesses zu sein, sind konstitutive Elemente des Geistes des Silicon Valleys, der durch Plattformunternehmer:innen global exportiert wird (Daub 2021). Pfeiffer notiert dazu kritisch: »Der eigentliche Unterschied zwischen Narrativ und wirtschaftlicher Realität einerseits und zwischen den hier betrachteten Unternehmen andererseits ist dabei, wie laut, überzogen oder realistisch die jeweiligen Versprechen zur Marktausdehnung sind.« (2022: 232) Die von Gründer:innen entwickelten Plattformen adressieren nicht primär die Optimierung technologischer Prozesse, binden diese aber notwendigerweise ein, da Plattformtechnologien leicht zugängliche Lösungen repräsentieren (Morozov 2013). Sie können als Teil einer solutionistischen Story bezeichnet werden, da sie in den meisten dargestellten Aussagen eng verwoben werden mit ökonomischen Zielen, wie dem Schaffen von »Mehrwert« (I1) oder »Kosteneffizienz« (I3) und »Profit« (I12).

Auch für die Interviewten scheint das Ziel klar formuliert, dass die entwickelten Lösungen über den lokalen, bestenfalls nationalen Kontext hinaus Anwendung finden, d.h. *skaliert* werden. »Die Welt ist voller ›Bugs‹, und es ist die Mission jeder Solutionistin, diese nach und nach zu ›fixen‹. Die Solutionistin ist eine Sozialingenieurin im wahrsten Sinne des Wortes.« (Nachtwey/Seidl 2017: 22; Herv. T.F.) Dieser Maxime folgend postuliert ein/e Gründer:in: »Also, ein Wort reicht fast schon, um das finale Ziel zu beschreiben: Skalieren.« (I12) Die Möglichkeit zu Skalieren und Reichweite aufzubauen, ist elementarer

Bestandteil des Geschäftsmodells¹⁴ – weniger altruistisch motivierte Problemlösekompetenz. Der monetäre Gewinn steht zwar für manche Gründer:innen im Vordergrund, jedoch nach eigenen Aussagen nicht für alle. Zumindest relativieren sie ihre Ansprüche und ergänzen pragmatischere, facettenreichere Motive, die sich aus flexiblen Arbeitszeiten, autonomer Arbeitsgestaltung und Verantwortung gegenüber ihren Mitarbeiter:innen, aber auch Spaß an der Arbeit, sowie Einfluss ergeben:

»Also [...] was mir wichtig ist halt irgendwie, so die Mitarbeiter, die wir haben, dass es denen gut geht. [...] Dass es mir selber natürlich AUCH gut geht. Ich will jetzt nicht unter der Brücke schlafen. Aber ich habe [...] den Eindruck, dass Du ab einem gewissen Einkommenslevel einfach da auch keinen [...] Mehrnutzen mehr daraus ZIEHST.« (I6)

Andere Gründer:innen heben das Ausüben und Aufbauen von Einfluss als primäre normative Motivationsgrundlage hervor:

»Wir bauen jetzt nicht das Produkt und dann verkaufen wir das. Sondern wir sagen, hey, wenn wir Einfluss haben, dann aber so richtig. Und die Vision, die wir skizzieren und [...] als interne Motivation schon nutzen, was möglich ist. Ich glaube, das ist der größte Unterschied und das erreichst du in einer Nische nicht. Und deswegen ist das Großdenken, die Skalierung schon der Kernaspekt, glaube ich.« (I3)

Vor dem Hintergrund der oft außergewöhnlichen Arbeitsumständen in Start-ups, d.h. häufig initial geringer Lohn bzw. Bezahlung in Firmenanteilen, bei hohen Arbeitszeiten und gleichzeitiger Unsicherheit ob des Erfolgs der Firma (vgl. Fackler et al. 2022: 2212), kann auch dieser Aspekt als solutionistischer Gedanken gelesen werden.

Da Gründer:innen in Start-ups »typischerweise nicht oder nur in Teilen auf Gelderwerb ausgerichtet« (Hepp 2022: 239) sind, ist die Arbeitsumgebung umso zentraler. Eine moderne flexible Arbeitsumgebung, in der es den Mitarbeiter:innen *gut geht*, trägt ebenso zur wahrgenommenen gesellschaftlichen

14 »Es ist kein Zufall, dass die Venture-Capital-Logik beim Investment von Start-Ups und Unicorns genau darauf setzt: auf die Skalierung, also das Versprechen einer maximal angenommenen Ausdehnung des Geschäftsmodells und der Nutzungszahlen [...]. Allerdings geht dieses Versprechen nur für die wenigsten Start-Ups (und ihre Investoren) auf [...].« (Pfeiffer 2021: 139)

Transformation bei, wie das Produkt der Firma. Im selben Zug kristallisiert sich heraus, wie die *Coding Elite* (Burrell/Fourcade 2021) die möglichen enormen Gewinne des Plattformgeschäfts als Chance interpretiert, die eigenen Arbeitsbedingungen frei zu gestalten und Teil der digitalen Transformation zu sein. Wie die Rechtfertigungen für Plattformen über monetäre Argumente hinaus vorgetragen werden, zeigt sich anhand der folgenden Analyse von Domäne und Stil.

Domäne und Stil: Selbstwirksamkeit, Anerkennung und soziale Netzwerke

Für Gründer:innen sind neben den Idealen der Weltverbesserung durch technologischen Fortschritt auch pragmatische Stories von Bedeutung, um sich für den Gründungsprozess zu entscheiden. Die angeführten Geschichten verdeutlichen, dass sich Stories nicht in kontextfreien Räumen bewegen, sondern stets eingebettet und umgeben sind, von anderen, teils konkurrierenden und grundsätzlicheren Domänen. Domänen und Stile kennzeichnen die Art und Weise, in der die Argumente vorgetragen werden, eingebettet sind und die zwischen den Zeilen mitschwingen. Da der Stil des Solutionismus eine Zuschreibung ist, sollen Themen aus angrenzenden Domänen identifiziert werden, die den Solutionismus flankieren.

Was für die Gründer:innen zählt, ist die Attraktivität des Start-up-Narratives, das mit Erwartungshaltungen gegenüber der angrenzenden Domänen der Arbeitsorganisation, der Selbstverwirklichung, der Verantwortung und dem Spaß in Verbindung steht. Als Anstoß zur Gründung berichten die Proband:innen von Anerkennungserfahrungen und der Würdigung ihres Konzepts:

»Wir haben vor zwei Jahren beim Projekt [Name Projekt] ein äh ein/das war ein Ideenwettbewerb/eine Idee eingereicht, haben den zweiten Platz gemacht und alle haben gesagt: »Super Konzept! Super Idee! Macht mal weiter!.« (17)

Im Gründungskontext arbeiten sie an der Verwirklichung von eigenen oder zumindest individuellen Ideen, sodass Spaß an/bei der Arbeit als Treiber funktioniert:

»Die das natürlich aus technischem Interesse, aus Spaß an der Sache machen. Aber auch aus nem gesellschaftlichen/Oder aus dem Wunsch, die ge-

sellschaftlichen Gegebenheiten n bisschen zu verändern. [...] Und das kriege ich auch von den andern Gründern oder Mitarbeitern von Start-ups mit. Und auch bei uns. Dass alles allen SUPERSPAß macht, an so nem Projekt mit zu/zu arbeiten. Sei es unser Projekt oder andere.« (I12)

Die Ideen und Gründungsnarrative schließen an eigene Interessen, Fähigkeiten und die Vorstellung von »guter Arbeit« an. Eine erfolgreiche Story eines Plattform-Start-ups ist entsprechend eng verwoben mit transformativen Ansprüchen an eine flexiblere Arbeitskultur, die als Narrativ besserer Arbeitsbedingungen ausgelegt wird. Als Stories werden die Ideale von Start-ups eingebettet in größere Kontexte, die häufig in Verbindung mit einer digitalen Transformation der Gesellschaft erzählt werden. Die Gründer:innen explizieren so ihre Vorstellung eines bestimmten Lebens- und Arbeitsstils, der z.B. den engen Kontakt zu Mitarbeiter:innen und anderen Gründer:innen betont sowie flache Hierarchien und Weiterbildungschancen verspricht.

Für Prozesse der Digitalisierung bieten Start-ups in der Phase der »Formierung« (Hepp 2022: 242) bzw. »Komposition« (Laux 2014: 163f.) einen experimentellen Raum. Hier können digitale Praxen und innovative technologische Infrastrukturen eingeführt und erprobt werden, ohne mit analogen Prozessen der Arbeitsorganisation zu konkurrieren oder sich an etablierten Pfadabhängigkeiten abzarbeiten.

»Für uns oder für mich auf jeden Fall, um selber Digitalisierung zu erfahren und äh Plattformen zu verstehen, wie funktionieren Plattformen [...]. Ob das jetzt hinterher ein Geschäftsmodell wird und ob das angenommen wird. Es wäre schön, wenn es so wäre, äh aber eben auch ähm Digitalisierung zu leben und zu sagen, ok, man hat sich schon mal mit Plattformen beschäftigt, man weiß, über was man redet. [...] aber auf der anderen Seite, wenn es erfolgreich wird, ist es natürlich auch schön.« (I7)

Eine weitere Facette des Start-up Narrativs ist daher besser beschrieben mit den Begriffen der Autonomie und der »kollektiven Selbstwirksamkeit«. Identitäten erfahren sich als kollektiv selbstwirksam, wenn sie sich gegenseitig und etwas erreichen, d.h. (gemeinsam) Einfluss auf ihre Umwelt und damit auf ihre Position (im Netzwerk) nehmen können (vgl. Rosa 2016: 270ff.). Mittels Selbstwirksamkeit lässt sich das Selbstverständnis von Start-ups als Pioniere nachvollziehen, die eine (nicht unbedingt von allen akzeptierte) Vorreiterrolle für ihre Domäne einnehmen. Durch das Erlernen, Erproben und Ausüben von

neuen Organisations- und Arbeitsformen führen Gründer:innen neue Praktiken in die Domäne ein und fungieren somit darüber hinaus als Intermediäre (vgl. Hepp 2022: 239). Entsprechend ließe sich ex negativo aus den Statements der Gründer:innen ableiten, dass ebenjener experimentelle Raum, die Möglichkeit »Digitalisierung zu leben« (I7) in klassischen Arbeitsverhältnissen nicht – oder zumindest nur in geringem Umfang – besteht.

Sobald das Start-up als Unternehmen wächst, entsteht in den Gründer:innen ein Verantwortungsgefühl gegenüber den Involvierten. Zum Anspruch der Selbstverwirklichung tritt nun der Anspruch hinzu, ein zeitgemäßes Unternehmen und ein stabiles Lebens- und Arbeitsumfeld zu etablieren.

»Deswegen bin ich nicht so einer, der irgendwie wie in Berlin irgendwie/ja, Du musst irgendwie ein Milliardenunternehmen gründen und dann muss alles super groß sein und so, damit du (unv.) bist, so ein Quatsch, ja? Das halt ich alles für nicht SINNVOLL. Ähm, und unter der Prämisse leben wir halt auch. Uns ist halt wichtig, dass wir ein langfristig STABILES Unternehmen aufbauen, [...] wo wir selber die KONTROLLE darüber haben, deswegen halt auch keine VC-Investoren und deswegen eigenfinanziert. [...] Äh, wo wir halt wirklich äh, selbst entscheiden können, was mit dem Unternehmen geschieht.« (I6)

Einer der Probanden formuliert, dass seine Mindesterwartung an die Gründung die erfolgreiche Institutionalisierung ist, sodass er und seine Mitarbeiter:innen »[...] gut damit LEBEN können [...]« (I6). Unmittelbar verbunden mit dem Ziel, seinen Mitstreiter:innen die existenziellen Unsicherheitserfahrungen der Gründungsphase zu ersparen, ist jedoch der Anspruch »[...] halt um den Faktor zehn ungefähr [zu] wachsen so.« (I6) Der solutionistische Anspruch der Weltverbesserung wird aufgrund der Skalierbarkeit von Plattformen in der Literatur häufig in globalen Effekten portraitiert. Hier zeigt sich, wie Firmen, die diesen Ansprüchen perspektivisch gerecht werden könnten, die Weltverbesserung zunächst in der Gründungsphase im überschaubaren Rahmen der Mitarbeiter:innen priorisieren. Viele der bisher genannten Aspekte von Stabilität und Einfluss auf die eigene Firmenzukunft, sind nicht unähnlich den Zielen regulärer kleiner und mittelständischer Unternehmen (Schweiger et al. 2022). Um eine Expansion zu ermöglichen und Stabilität zu sichern, sind die Gründer:innen darauf angewiesen, dass sie ihr Story-Set ausdifferenzieren, weitere Identitäten akquirieren und ihr Netzwerk um Mitarbeiter:innen und Investor:innen erweitern.

Neben die Polis der Solution treten Argumente, die eher der Polis der Netzwerks zuzuordnen wären.

»Dann du brauchst ein gutes Netzwerk. Viele Start-ups/Also viele Geschäftsideen kannst du halt auch nur mit einem externen Geld finanzieren. Dann brauchst du aber ein gutes Netzwerk, um ähm, das Geld zu bekommen. Brauchst eine gewisse VITA, damit du überhaupt die äh, potenziellen Investoren überzeugen kannst. [...] Ähm, und dann ist wahrscheinlich leider auch ein bisschen GLÜCK dabei.« (I6)

Die Gründer:in spricht hier von der narrativen Komposition der eigenen Vita, einer Geschäftsidee, eines guten Netzwerks und Glück, um potenzielle Kapitalgeber:innen zu überzeugen. Die Rolle der Kapitalgeber:innen als »Lebensnerv« (Pfeiffer 2021: 224) für KI- und Plattformunternehmen, ist folglich bereits institutionalisiert, was das bisher stilisierte Risiko der Start-up-Gründung zumindest teilweise relativiert (ebd.: 231).

In der Gründungsphase ist besonders das Feedback anderer Start-ups im Netzwerk richtungsweisend (vgl. Hepp 2022: 242f., 246). Im späteren Verlauf beziehen die Gründer:innen das Feedback mehrheitlich indirekt als Rückmeldung der plattformvermittelten Kundeninteraktion. Sie räumen der Kundenmeinung hohe Relevanz ein, deuten sie aus den zur Verfügung stehenden Daten und empfinden den kundenseitigen Fokus auf ein attraktives (»sexy« (I4)) Frontend als Einschränkung der eigenen Autonomie. Um die »viel größere Vision« (I4) zu vermitteln, Kundenbindung zu erreichen und die Kunden »glücklich zu machen, um eben eine möglichst große Kundenzahl zu haben« sind für Plattformunternehmen datenbasierte Feedbackschleifen letztlich unerlässlich:

»Die Höchste [Relevanz hat Datenhandling, T.F.]. Also, das ist die einzige Variante, schnell und die auf der Wahrheit basierend Änderungen vornehmen zu können, wo man weiß, dass/hier hängen Leute. Und das müssen wir fixen, das müssen wir anders machen. Ohne Daten geht das nicht. Und das hat in Deutschland halt auch noch niemand gecheckt[...]. Also, das ist die Grundlage dafür, dass man schnell sein kann. Sodass man Analysen und Daten auswerten kann.« (I3)

Qua Feedback setzen sich die Unternehmen selbst konstanten Kontrollversuchen von Kund:innen aus, die das Produkt und auch den Arbeitsprozess formen. Gründer:innen interpretieren ihre Beziehung zu Nutzer:innen so,

dass sie dazu beitragen, soziale oder pragmatische (Alltags-)Probleme der Nutzer:innen zu lösen. Kaum diskutiert wird hingegen der Aspekt der Story, dass die von Nutzer:innen produzierten Daten elementarer Bestandteil des digitalen Geschäftsmodells sind.

Start-ups beschreiten somit zwar neue Pfade in technologischer und arbeitsorganisationaler Hinsicht, ihre KI-Anwendungen weichen jedoch häufig nicht weit von etablierten Verfahren ab bzw. orientieren sich extensiv an den Möglichkeiten der großen Firmen: »Start-ups can find novel uses of machine learning and apply them to new products. Yet they will remain tenants on the clouds provided by the biggest AI companies, dutifully paying their rents to these American and Chinese companies« (Srnicek 2022: 254). In diesem Sinne sind die entwickelten, einzigartigen Geschäftsmodelle und Datentypen zwar innovativ, aber letztlich in gewissermaßen durch die Netzwerkförmigkeit des soziotechnischen Story-Sets an die Möglichkeiten des Mainstreams gebunden (vgl. ebd.: 243f.).

Story-Line: Don't be evil und die digitale Transformation

Stories stehen nicht für sich selbst und sind nicht als abgeschlossene Formationen zu betrachten, die – einmal erzählt – Ordnung stiften bzw. (Netzwerk-)Positionen ausweisen. »Any particular story by itself would quickly get so far out of step with the ongoing situation as to be unuseable.« (White/Godart 2007: 7) Stories sind »permanently beta« (Neff/Stark 2004), d.h. sie sind prozessual, dauerhaft im Werden begriffen und selten abgeschlossen. Stories enthalten also eine temporale Dimension, die White begrifflich in Story-Lines verdichtet: »Story-lines are explanation spread over time in time frames. And story-lines come at least in a pair as they offer alternative accounts for change and constitute the fabric of social time.« (White/Godart 2007: 7) Sie fokussieren den chronologischen Aufbau der Stories, deren retrospektiven Ausgangs- und prospektiven Fluchtpunkte. Das narrative Verknüpfen von Stories konstituiert soziale Temporalität und reduziert durch das Ordnen von kulturellen Deutungen situative Ambiguität. Letztlich begrenzen Stories soziale Kontexte insofern, als dass sich das Netzwerk auf eine gültige Story-Line verständigt. Durch die Festlegung werden alternative Stories ausgeschlossen und die Story kann aushärten und sich institutionalisieren: In einer genealogischen Rekonstruktion lässt sich der Solutionismus auf die kalifornischen Subkultur der späten 1960er Jahre und das *Versprechen der (technologischen) Dezentralisierung* sozioökonomischer Koordinationsformen zurückführen (vgl. Schrape 2019).

Dieses lässt sich als Story historischen Ursprungs auch in den Interviews mit den Gründer:innen wiederfinden. Im folgenden Abschnitt gilt es zu erörtern, inwiefern die Probanden auf vergangene Stories zurückgreifen und bestehende in die Zukunft projizieren.

Eine/r der Proband:innen bezieht sich auf eine historische Sequenz, in der Google den eigenen Slogan von *Don't be evil* in *do the right thing* änderte. Das Firmenmotto *Don't be evil* symbolisierte eine Art digitalen Verhaltenskodex, und wurde deshalb in- und außerhalb Googles diskutiert (vgl. z.B.: Pakalski 2018). Heute illustriert der Wandel des Firmenmottos den Wandel des Unternehmens selbst. Eine Gründer:in greift im Interview die Story auf, um sich bzw. die eigene Plattform explizit von Google zu distanzieren:

»Also äh, gerade dieser, diese Abwägung zwischen Einfachheit und Privatsphäre, das ist eine Sache, die uns sehr beschäftigt. Ähm, und ja, diese drei großen Player sind ja echt ähm, was das angeht, ähm stark in der Kritik auch, nicht. Und das auch nicht umsonst. Also allein was Google [...] die letzte Zeit gemacht hat, ähm, ich war früher Google-Fan, weil so viel Open Source, sehr transparent gemacht wurde. Der Wandel in den letzten Jahren ist halt katastrophal. Ich meine, das fing damals an, als die sich (umgemo-delt?) haben, zu Alphabet und ›Don't be evil‹ geborgt haben, nicht. Das war damals der Slogan von denen. Der wurde entfernt. Ähm, sagt ja schon viel aus, nicht.« (I4)

In den Interviews wird ebenfalls ein Bezug zu dem Motto hergestellt, das mit Facebook in Verbindung gebracht wird: *Move fast and break things*. Die Probanden distanzieren sich im Kontext der Privatsphäre ebenfalls von Facebook. Sie verweisen aber auch darauf, dass das Ausnutzen gesetzlicher Grauzonen im Sinne des Mottos kein Alleinstellungsmerkmal der Plattformökonomie darstellt:

»Plattformen haben auch natürlich ein paar negative Assoziationen, die bei MIR in den Kopf kommen. Weil ich denke immer sofort bei Plattform an Geschäftsmodelle, die aktuell funktionieren. Und das sind so Sachen wie Pizza.de oder Lieferando.de, die sich äh quasi Monopole schaffen. [...] und das fühlt sich manchmal ein bisschen komisch an, weil eigentlich diejenigen die Gewinner sind, wirtschaftlich, die keine Produkte anbieten, sondern nur die Infrastruktur zusammenbinden. [...] Ist glaube ich aber auch nicht nur äh digitale Plattform. Wir haben auch in der Autoindustrie, was immer mehr herauskommt auch, solche Sachen gehabt, wo bewusst Regeln gebro-

chen wurden. [...] Ich glaube das ist nicht nur die Digitalen, das sind auch die anderen. [...]« (14)

Anstatt sich auf die Erfolgsgeschichten der (meist) amerikanischen Plattformunternehmen zu berufen und als Musterbeispiel der digitalen Ökonomie anzuführen, grenzen die Gründer:innen sich eher ab. Sie stellen so ihre Individualität heraus und inszenieren sich als alternatives Plattformenmodell, das aus den Fehlern der Großen gelernt hat. Zumindest partiell entkoppeln sie sich von den klassisch-solutionistischen Erfolgstories der amerikanischen Plattformökonomie. Ein alternativer Erklärungsansatz wäre, dass sich seit Morozovs (2013) Diagnose der solutionistischen Ethik der gesellschaftliche Diskurs zu den Großen der Plattformökonomie geändert hat, da deren gesellschaftsweiten Implikationen zunehmend ans Licht der Öffentlichkeit gelangen.

Mit Blick auf zukünftige Geschäftsmodelle, skizzieren die Gründer:innen zwei weitere Fluchtpunkte der Story-Lines. Es erfolgt einerseits eine Abgrenzung gegenüber analogen Geschäftsmodellen. Analoge Geschäftsmodelle seien nicht in dem Maße dazu in der Lage, sich an den Wünschen der Kund:innen zu orientieren, wie digitale Unternehmen, die das Verhalten der Kund:innen tracken. Analoge sind somit stets im Nachteil gegenüber und unter Umständen abhängig von digitalen Unternehmen:

»Wer heute sich nicht mit Plattformen beschäftigt äh, der muss damit rechnen, dass äh er in Zukunft nur noch Lieferant ist für Plattformen und dann auf einer Seite von der Plattform steht. Entweder Anbieter ist oder Nutzer ist und äh, wenn man als Nutzer nicht mehr den Kontakt zum Kunden hat, dann hat man in Zukunft ein Problem.« (17)

Außerdem wird die Start-up-Story-Line mit einer digitalen Transformation in Verbindung gebracht, die mit der Polarisierung entlang digitaler und analoger Geschäftsmodellen einhergeht:

»Es wird eine Spaltung der Gesellschaft geben und ähm wie ich mir darüber bewusst bin, dass man äh nicht jedes Unternehmen digital transformieren kann, [...] man kann nicht jeden Mitarbeiter und jeden Menschen digital transformieren oder auf jeden Fall zur Bereitschaft bringen, Digitalisierung wirklich auch zu leben und zu verstehen. Äh, von daher wird es schon irgendwo eine Spaltung der Gesellschaft geben, aber die gibt es glaube ich

immer schon. Es gibt heute auch Leute, die sagen: »Internet fange ich gar nicht mehr mit an, will ich mich nicht mit beschäftigen!« (17)

Die digitalkapitalistischen Plattformen werden insofern verteidigt, als sie zukunftsfähig sind. Analoge Unternehmen seien Ausdruck einer unflexiblen Form der Arbeitsorganisation, die nicht dem Zeitgeist entspricht. An diese umfangreiche Prognose anknüpfend sollen anschließend die Ergebnisse in einen weiteren Kontext der solutionistischen Logiken gesetzt werden.

5. Diskussion

Die Entwicklung digitaler Plattformen durch Start-ups ist in komplexe narrative Story-Sets eingebettet, voraussetzungsreich und mit bestimmten Zukunftserwartungen verknüpft. Whites Theorie bietet sich ob seiner Komplexität als analytisches Tool an, um qualitative Daten zu sezieren und verschiedene Blickwinkel einzunehmen. Sie ist weniger voraussetzungsreich und normativ, als die Theorie der Rechtfertigungsordnungen, und somit der empirischen Analyse näher.¹⁵ In der Konstruktion digitaler Plattformen spielen narrative Elemente eine zentrale Rolle: Solutionistische Stories werden in Story-Sets eingewoben, ihr am Silicon Valley orientierter Stil lässt sich auch andernorts identifizieren, sie sind gleichzeitig durch historische Entwicklungslinien geprägt und prägen diese. Da Normativität nicht nur Teil solutionistischer Stories ist, sondern qua Konstruktionsprozess Teil digitaler Plattformen, bedarf es auch der kritischen Betrachtung ebenjener Konstruktionsprozesse und digitalen Plattformen. Wie von Mützel (2009) beschrieben, dienen

15 Am Beispiel der »Analytischen Soziologie« Hedströms verweist Baum (2010: 2) auf eine potenzielle »theorieimmanente Schwäche« analytischer und empirischer Forschung. Nämlich die »ausbleibende Thematisierung der eigenen Normativität« bzw. das unausgesprochene Verstehen der eigenen Arbeit als neutral, insbesondere, wenn die Möglichkeit besteht, dass die Ergebnisse in Entscheidungsfindungsprozesse einfließen. Dieser Kritik kann im Rahmen des Artikels nur bedingt Sorge getragen werden. Baum fordert eine »integrative Theoriebildung«, d.h. die »Einnahme verschiedener Perspektiven auf den Gegenstand«. (vgl. Baum 2010: 15). In dem vorliegenden Beitrag werden im Anschluss an Boltanski/Thévenot (2007) und White (2008) zwei Perspektiven eingenommen. Der eigene Standpunkt lässt sich zumindest so weit reflektieren, dass im Sinne der Science & Technology Studies nicht nur jede Konstruktion, sondern auch jede Dekonstruktion normative Züge aufweist.

Stories und Erfolgsgeschichten als Orientierungsnarrative, zu denen andere Marktteilnehmer:innen Position beziehen. Stories von globalen Playern wie Amazon und Co., die um Plattformen herum gestaltet werden, können so über den Kontext des Marktes hinausreichen.

Für Günther (2020: 525) ist die Entwicklung digitaler Plattformen kritisch zu betrachten: Gründer:innen digitaler Plattformen sind Produzenten *smarter Ordnungen*, also Ordnungen, »[...] die darauf ausgerichtet sind, durch intelligentes Design und mit Hilfe algorithmischer Operationen Abweichungen von ihren Normen zu minimieren oder ganz unmöglich zu machen [...]«. ¹⁶ Im Konstruktionsprozess verdinglichen Gründer:innen Stories, manifestieren und verschleiern Kontrollstrategien, von denen sie erwarten, dass sie im allgemeinen Interesse liegen (vgl. ebd.: 536). Dies schließt an ein grundlegendes Argument der *Science and Technology Studies* an, dass es keine wertfreien Konstruktionsprozesse geben kann. Einmal etabliert substituieren smarte Ordnungen Prozesse kritisch-reflexiver Normprüfung (vgl. ebd.: 545f.). Sie *verschließen Kontingenz- und Möglichkeitsräume*, indem Reflexivität und Normprüfung in digitaltechnologische Prozesse, in Code übersetzt werden (vgl. Reckwitz 2021: 72ff.).

Die Gründer:innen hätten zwar Einblick in die algorithmische Dimension smarter Ordnungen, jedoch nur aus der Perspektive numerischer Parameter, deren Korrektur schließlich der effizienten Normverwirklichung dient (vgl. Günther 2020: 549). Allerdings ist das Design digitaler Plattformen nicht allein Ausdruck des individuellen Interesses der Gründer:innen, sondern ebenfalls eingebettet in Trends, abhängig von bestehenden Code-Strukturen und den Präferenzen der Nutzer:innen.

Welcher Prägung solutionistische Stories folgen, hängt auch von gesellschaftlichen, geographischen und kulturellen Kontexten ab. Kalbermatter et al. untersuchen den digitalkapitalistischen Geist schweizerischer Prägung und fordern eine kulturelle Differenzierung von *Varieties of Spirits of Capitalism* (vgl. 2020: 46). Auf der Grundlage einer quantitativen Analyse untersuchen Sternberg et al. (2020) individuelle Motive zur Unternehmensgründung im inter-

16 »An die Stelle des kritischen Urteils der Adressat:innen qua Autor:innen der Norm, zu der aus Teilnehmer:innenperspektive Stellung genommen werden müsste, tritt die Verhaltensprognose aus der Beobachter:innenperspektive, über deren Wahrheit oder Falschheit ohne Stellungnahme der beobachtenden Objekte geurteilt wird. Der nächste Schritt wäre es, politische Entscheidungen gleich ganz von KI-Maschinen treffen zu lassen.« (ebd.: 547)

nationalen Vergleich, wobei das solutionistische Motiv »Um die Welt zu verändern« integriert wird. Etwa 40 % der in Deutschland Befragten geben an, dass für sie die Weltveränderung als Motiv die Unternehmensgründung beeinflusst hat. Zum Vergleich: In den USA und Polen geben über 60 % der Befragten das solutionistische Motiv als Grund an. Mit ca. 70 % Zustimmung wird in Deutschland deutlich häufiger das Motiv angegeben, eine Familientradition fortführen zu wollen. In den USA hingegen scheint es mit über 70 % Zustimmung ebenfalls als prominentes Gründungsmotiv zu sein, möglichst viel Geld verdienen zu wollen (vgl. ebd.: 34). Sowohl bei Sternberg et al. (2020) als auch bei Kalbermatter et al. (2020) wird deutlich, dass die Analyse solutionistischer Stories der Differenzierung bedarf, da es sich um *idealtypische* Konstrukte handelt, die auch von Institutionen wie z.B. dem Recht beeinflusst werden.

Weder aus diesen quantitativen Studien noch aus den hier analysierten Interviews geht hervor, inwiefern sozioökonomische und soziohistorische Kontexte die Abundanz solutionistischer Stories beeinflussen. Sollte die Reichhaltigkeit solutionistischer Stories zunehmen (Nachtwey/Seidl 2017), drängen sich bspw. Fragen nach Klimawandel, Generationenwandel, oder politisch-militärischen Konflikten auf, die kaum rein technologische Lösungen fordern. Wird die Polis der Solution dann abgelöst, etwa durch eine ökologische Polis?

Abschließend lässt sich sagen, dass auch die Erhebung solutionistischer Wertorientierung sowie das Forschungsdesign eine entscheidende Rolle spielen. Seidl (2021) fokussiert in seiner Analyse Zeitschriftenbeiträge, die im Magazin WIRED veröffentlicht werden. Ein solches Vorgehen wird dem Solutionismus gerecht, da es nicht unterstellt, dass es sich um ein klassen-, gesellschafts- und milieuübergreifend wirksame Story handelt. Gleichzeitig wird deutlich, dass es sich um eine Story unter vielen handelt, die in Reinform kaum zu identifizieren ist und mit vielen anderen Stories um diskursive Deutungshoheit kämpft. Das deduktive Voraussetzen der Rechtfertigungsordnungen als analytische Kategorien ist zwar vielversprechend, weil sie ein breites Spektrum potenzieller Stories abbilden. Jedoch ist es unwahrscheinlich, dass die Kategorisierung der Sequenzen oder Interviewpassagen disjunkt erfolgen kann. Die Ergebnisse von Sternberg et al. (2020) deuten darauf hin, dass die Befragten multiple Motive für ihre Gründung angeben, von denen das des Weltverbesserns eines unter vielen ist.

6. Fazit

Der Beitrag verfolgte das Ziel zunächst auf theoretischer Ebene die Konzepte der Rechtfertigungsordnung im Anschluss an Boltanski und Thévenot (2007) mit dem Konzept der Stories von White (2007) ins Gespräch zu bringen. Gegenstand der empirischen Analyse waren Interviews mit Gründer:innen digitaler Start-ups, die im Kontext des Forschungsprojektes »INDIZ« erhoben und einer Sekundäranalyse unterzogen wurden. Es wurde herausgestellt, dass mittels der White'schen Formationsbegriffe solutionistische Stories dechiffriert werden können: Mit Story-Sets etwa kann die Komposition einer Story aufgelöst werden, indem auf einzelnen Elemente des Story-Sets scharfgestellt wird. Die Betrachtung von Domänen und Stilen legt den Kontext der Story frei und Story-Lines fokussieren retrospektive Ausgangs- und prospektive Fluchtpunkte. Mit der TIK ließ sich zeigen, auf wie vielen Ebenen die Story digitaler Start-ups voraussetzungsreich ist, wie sie subtil das Mind-Set der Gründer:innen prägt, indem sie anverwandelt, rekombiniert und integriert wird, und somit die Erwartungshaltungen der Gründer:innen gegenüber ihrer Plattform, aber auch gegenüber einer gesellschaftlichen (digitalen) Transformation prägen. Die Leerstelle in Whites Theorie bzgl. dem Erzeugen von Stories durch Technologien, konnte zumindest in ersten Schritten diskutiert werden.

Das Besondere des Solutionismus ergibt sich unmittelbar, wenn es vor dem Hintergrund der großen digitalkapitalistischen Unternehmen gedacht wird: Primär kapitalistische Interessen werden in ein Gewand der Weltverbesserung gehüllt, das so fein gewebt ist, dass es für Laien nicht ohne Weiteres zu durchschauen ist. Die Dekonstruktion der solutionistischen Argumente lokaler Gründer:innen, zeichnet aber ein weniger brisantes Bild: Mit Blick auf die angeführten Story-Sets wird deutlich, dass ökonomische Interessen und Wohlstand häufig Teil des Gründungsmotivs sind. Auch bei den digitalen Start-ups besteht die Gefahr (zum Teil auch die Absicht), dass Daten, die in bestimmten Kontexten von Nutzer:innen produziert werden, Aufschluss über andere, unter Umständen intimere, wertvollere Aspekte der Lebensführung geben. In den meisten Fällen zeigen sich die lokalen Start-up-Stories aber wesentlich weniger polarisierend. Die Anwendung von maschinellem Lernen, das Produzieren smarterer Ordnungen oder das skrupellose Aggregieren von Nutzer:innendaten unter der Vortäuschung falscher Tatsachen ist in der Regel nicht zu erkennen – was ebenfalls an der Form der Erhebung liegen kann und sicherlich ein Teil der Limitationen der vorliegenden Analyse ist. Es stimmt

zwar, dass solutionistische Stories implementiert werden, und, dass sie überzufällig häufig den Aspekt der technioptimistischen Weltverbesserung aufweisen. Allerdings sind auch das Skalieren des Produkts, das Partizipieren an globalen Märkten, neue Arbeitsorganisationsformen und Autonomie zentrale Aspekte digitaler Dienstleistungen und Teil solutionistischer Story-Sets. Maschinelles Lernens scheint eher als marketingstrategische Story der eigenen digitalen Dienstleistung relevant zu sein, weniger als technologische Implikation einer KI. Speziell die großen Plattformunternehmen werden von den Gründer:innen aufgrund ihres zu Grunde liegenden Datenhandlings kritisch betrachtet.

Zwei weitere Aspekte sind zu erwähnen, die die Annahme eines digitalkapitalistischen Silicon-Valley-Geistes auf der Ebene der Start-ups entschärfen und kontextualisieren. Erstens werden in Kap. 5 unterschiedliche Motivstrukturen entlang kultureller Gesellschaftsformationen aufgezeigt. Somit können die von uns analysierten Interviews einen standortbezogenen Pragmatismus aufweisen, etwa weil in Deutschland rasantes Wachstum mittels des massiven Einsatzes von Risikokapital als ökonomisches Motiv eher unwahrscheinlich ist. Zweitens geraten die Stories der idealtypischen Solutionist:innen der Metaplattformen (Amazon, Facebook, Google und Co.) zunehmend unter Rechtfertigungsdruck – sowohl in der öffentlichen Meinungsbildung als auch im Kontext einer sich verschärfenden Gesetzeslage.

»While solutionism provided a powerful normative defense at a time when capitalists were increasingly criticized for producing, rather than solving social problems, it's proponents have recently themselves come under criticism for producing all sorts of social problems, from creating addiction to spreading misinformation. It would be unwise however, to write solutionism off [...]«. (Seidl 2021: 207)

Als Ergebnis der analytischen Dekonstruktion solutionistischer Argumente in Deutschland lässt sich festhalten, dass nicht jedes digitale Unternehmen ausschließlich Solutionist:innen beheimatet, dass nicht jede digitale Plattform anstrebt, die Welt zu verändern, und, dass nicht jeder Datensatz sich eignet, mittels maschineller Lerner in smarte Ordnungen überführt zu werden. Im Anschluss an die diskutierte Literatur strebt diese Analyse an den Solutionismus differenziert zu betrachten. Was allerdings nicht verleugnet werden kann – und sich auch in den Interviewpassagen der lokalen Gründer:innen widerfindet – ist die ökonomische Attraktivität und Perspektive

auf potenzielle (globale) Einflussnahme, die digitale Geschäftsmodelle auch außerhalb des Silicon Valleys ausstrahlen. Für die Proband:innen sind diese Aussichten scheinbar Motivation genug, um das Risiko einer Unternehmensgründung in diesen stark fluktuierenden Märkten einzugehen, auch wenn sie die Machtkonzentration seitens der großen digitalen Infrastrukturen kritisch betrachten. Das Ziel einer analytischen Dekonstruktion der Start-up-Story besteht daher in der Entzauberung des Narratives, nicht in dem Leugnen jeglichen Potenzials (des Empowerment) digitaler Plattformen.

7. Literatur

- Agre, Philip. E. (1997): »Toward a Critical Technical Practice: Lessons Learned Trying to Reform AI«, in: Geoffrey C. Bowker/Susan Leigh Star/Les Gasser/William Turner (Hg.), *Toward a Critical Technical Practice: Lessons Learned Trying to Reform AI*, Mahwah/New Jersey: Lawrence Erlbaum Associates, S. 131–157.
- Baum, Markus (2010): »Der soziale Wandel der analytischen Soziologie im Lichte kritischer Reflexion«, in: Thomas Kron/Thomas Grund (Hg.), *Die Analytische Soziologie in der Diskussion*, Wiesbaden: VS Verlag für Sozialwissenschaften, S. 133–149.
- Beckert, Jens (2013): »Imagined futures: fictional expectations in the economy«, in: *Theory and Society* 42(3), S. 219–240.
- Boltanski, Luc/Chiapello, Ève (2003): »Der neue Geist des Kapitalismus«, in: Franz Schultheis (Hg.), *Klassische und zeitgenössische Texte der französischsprachigen Humanwissenschaften (= édition discours, Band 38)*, Konstanz: UVK Verlagsgesellschaft mbH.
- Boltanski, Luc/Thévenot, Laurent (2007): *Über die Rechtfertigung. Eine Soziologie der kritischen Urteilskraft*, Hamburg: Hamburger Edition.
- Bucher, Taina (2021): *Facebook*, Cambridge: Polity Press.
- Burrell, Jenna/Fourcade, Marion (2021): »The Society of Algorithms«, in: *Annual Review of Sociology* 47(1), S. 213–237.
- Daub, Adrian (2021): *Was das Valley denken nennt: über die Ideologie der Techbranche*, Berlin: Suhrkamp.
- Dolata, Ulrich (2015): »Volatile Monopole. Konzentration, Konkurrenz und Innovationsstrategien der Internetkonzerne«, in: *Berliner Journal für Soziologie* 24(4), S. 505–529.

- Fackler, Daniel/Hölscher, Lisa/Schnabel, Claus/Weyh, Antje (2022): »Does working at a start-up pay off?«, in: *Small Business Economics* 58(4), S. 2211–2233.
- Fuhse, Jan/Schmitt, Marco (2015): Zur Aktualität von Harrison White. Einführung in sein Werk, Wiesbaden: Springer VS.
- Günther, Klaus (2011): Die Herausbildung normativer Ordnungen. Interdisziplinäre Perspektiven, Frankfurt a.M./New York: Campus Verlag.
- Günther, Klaus (2020): »Von normativen zu smarten Ordnungen?«, in: Rainer Forst/Klaus Günther (Hg.), *Normative Ordnungen*, Berlin: Suhrkamp Verlag, S. 523–553.
- Häußling, Roger (2006): »Ein netzwerkanalytisches Vierebenenkonzept zur struktur- und akteursbezogenen Deutung sozialer Interaktionen«, in: Bettina Hollstein/Florian Straus (Hg.), *Qualitative Netzwerkanalyse. Konzepte, Methoden, Anwendungen*, Wiesbaden: Verlag für Sozialwissenschaften, S. 125–153.
- Helmond, Anne (2015): »The Platformization of the Web: Making Web Data Platform Ready«, in: *Social Media + Society* 1(2).
- Hepp, Andreas (2022): »Jenseits der Disruption: Zum Lebenszyklus von Pioniergemeinschaften und ihrer Rolle beim Entstehen einer ›digitalen Gesellschaft‹«, in: *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie* 74 (S1), S. 231–255.
- Jaton, Florian (2020): *The constitution of algorithms: ground-truthing, programming, formulating*, Cambridge/Massachusetts: MIT Press.
- Jaton, Florian (2021): »Assessing biases, relaxing moralism: On ground-truthing practices in machine learning design and application«, in: *Big Data & Society* 8(1).
- Kalbermatter, Jacqueline/Nachtwey, Oliver/Truffer, Johannes (2020): »Der digitalkapitalistische Geist in der Schweiz. Muster normativer Orientierungen in einem Traditionsunternehmen und einem transnationalen Digitalunternehmen«, in: *Behemoth* 13(1), S. 35–46.
- Karafilidis, Anthanasios (2019): »White, Harrison C. (1995): Network Switchings and Bayesian Forks: Reconstructing the Social and Behavioral Sciences. *Social Research* 62 (4), S. 1035–1063«, in: Boris Holzer/Christian Stegbauer (Hg.), *Schlüsselwerke der Netzwerkforschung*, Wiesbaden: Springer, S. 587–591.
- Köbis, Nils/Mossink, Luca D. (2021): »Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-gener-

- ated from human-written poetry«, in: *Computers in Human Behaviour* 114.
- Laux, Henning (2014): *Soziologie im Zeitalter der Komposition. Koordinaten einer integrativen Netzwerktheorie*, Weilerwist: Velbrück Wissenschaft.
- Mayring, Philipp (2015): *Qualitative Inhaltsanalyse: Grundlagen und Techniken*. 12. überarbeitete Auflage, Weinheim Basel: Beltz Verlag.
- Medjedović, Irena (2014): »Qualitative Daten für die Sekundäranalyse«, in: Jörg Blasius/Nina Baur (Hg.), *Handbuch Methoden der empirischen Sozialforschung*, Wiesbaden: Springer Fachmedien Wiesbaden, S. 223–232.
- Metzger, Georg (2018): »KfW-Gründungsmonitor 2018. KfW Research, KfW Bankengruppe, Frankfurt«, in: KfW vom Mai 2018, <https://www.kfw.de/PDF/Download-Center/Konzernthemen/Research/PDF-Dokumente-Gr%C3%BCndungsmonitor/KfW-Gruendungsmonitor-2018.pdf>.
- Morozov, Evgeny (2013): *To save everything, click here: the folly of technological solutionism* (First edition.), New York: PublicAffairs.
- Morozov, Evgeny (2014): »The rise of data and the death of politics«, in: *The Guardian* vom 20.06.2014, <https://www.theguardian.com/technology/2014/jul/20/rise-of-data-death-of-politics-evgeny-morozov-algorithmic-regulation>.
- Morozov, Evgeny (2022): »Potenziale der Digitalisierung jenseits von Markt und Staat«, in: *APuZ* 10–11, S. 4–9.
- Mützel, Sophie (2007): »Marktkonstitutionen durch narrativen Wettbewerb«, in: *Berliner Journal für Soziologie* 17, S. 451–464.
- Mützel, Sophie (2009): »Koordinierung von Märkten durch narrativen Wettbewerb«, in: *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 49, S. 87–106.
- Mützel, Sophie (2009): »Networks as Culturally Constituted Process. A Comparison of Relational Sociology and Actor-network Theory«, in: *Current Sociology* 57(6), S. 871–887.
- Nachtwey, Oliver/Seidl, Timo (2017): »Die Ethik der Solution und der Geist des Kapitalismus«, in: *IFS Working Paper 11*, Frankfurt a.M.: Institut für Sozialforschung.
- Neff, Gina/Stark, David C. (2002): »Permanently Beta: Responsive Organizations in the Internet Era«, in: *Institute for Social and Economic Research and Policy ISERP Working Paper 02–05*, CA: Sage.
- Pakalski, Ingo (2018): »Google verabschiedet sich von ›Don't be evil‹«, in: *Golem* vom 21.05.2018, <https://www.golem.de/news/verhaltenskodex-google-verabschiedet-sich-von-don-t-be-evil-1805-134479.html>.

- Pfeiffer, Sabine (2021): Digitalisierung als Distributivkraft: über das Neue am digitalen Kapitalismus, Bielefeld: transcript Verlag.
- Phillip, Tobias (2017): Netzwerk zwischen Physik und Soziologie. Perspektiven der Netzwerkforschung mit Bruno Latour und Harrison White. Dissertation, Wiesbaden: Springer VS.
- Pongratz, Hans J. (2022): »Plattformen auf dem Arbeitsmarkt: Digitalisierung und Diversifizierung in der Beschäftigungsindustrie«, in: KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie 74(S1), S. 133–157.
- Pongratz, Hans J./Bernhard, Stefan/Abbenhardt, Lisa (2014): »Fiktion und Substanz. Praktiken der Bewältigung zukunftsbezogener Ungewissheit wirtschaftlichen Handelns am Beispiel der Gründungsförderung«, in: Berliner Journal für Soziologie 24(3), S. 397–423.
- Reckwitz, Andreas/Rosa, Hartmut (2021): Spätmoderne in der Krise. Was leistet die Gesellschaftstheorie?, Berlin: Suhrkamp Verlag.
- Rieder, Bernhard/Skop, Yarden (2021): »The fabrics of machine moderation: Studying the technical, normative, and organizational structure of Perspective API«, in: Big Data & Society 8(2), S. 1–16.
- Rosa, Hartmut (2016): Resonanz. Eine Soziologie der Weltbeziehung, Berlin: Suhrkamp.
- Sadowski, Jathan (2020): Too smart: how digital capitalism is extracting data, controlling our lives, and taking over the world, Cambridge/Massachusetts: MIT Press.
- Sadowski, Jathan (2022): »Planetary Potemkin AI: The Humans Hidden inside Mechanical Minds«, in: Mark Graham/Fabian Ferrari (Hg.), Digital Work in the Planetary Market, Cambridge/Massachusetts: MIT Press, S. 229–240.
- Sadowski, Jathan/Selinger, Evan (2014): »Creating a taxonomic tool for technocracy and applying it to Silicon Valley«, in: Technology in Society 38, S. 161–168.
- Schrape, Jan-Felix (2019): »The Promise of Technological Dezentralization. A Brief Reconstruction«, in: Society 56, S. 31–36.
- Schreier, Margrit (2012): Qualitative content analysis in practice, Los Angeles: Sage.
- Schweiger, Stefan/Sowka, Sarita/Schmierl, Klaus/Roth, Philip/Pfahl, Sebastian/Heidling, Eckhard/Georges, Fadi/Fellmann, Michael/Eber, Sarah/Bohn, Gloria/Böhme, Torsten (2022): »Analyse organisationaler Prozesse in klein- und mittelständischen Unternehmen«, in: Verena Nitsch/Christopher Brandl/Roger Häußling/Jacqueline Lemm/Thomas Gries/

- Bernhard Schmenk (Hg.), *Digitalisierung der Arbeitswelt im Mittelstand 1*, Berlin: Springer Vieweg, S. 29–54.
- Seidl, Timo (2020): *Ideas, Politics, and Technological Change. Essays on the Comparative Political Economy of Digital Capitalism*. Doctoral Thesis, Florence: European University Institute.
- Somers, Margaret R. (1994): »The narrative constitution of identity: A relational and network approach«, in: *Theory and Society* 23(5), S. 605–649.
- Srnicek, Nick (2017): *Platform capitalism*, Cambridge/UK/Malden/MA: Polity.
- Srnicek, Nick (2022): »Data, Compute, Labor«, in: Mark Graham/Fabian Ferrari (Hg.), *Digital Work in the Planetary Market*, Cambridge/Massachusetts: MIT Press, S. 241–262.
- Sternberg, Rolf/Gorynia-Pfeffer, Natalia/Wallisch, Matthias/Baharian, Armin/Stolz, Lennard/Schauer, Julia (2020): *Global Entrepreneurship Monitor – Unternehmensgründungen im weltweiten Vergleich. Länderbericht Deutschland 2019/20*, Eschborn: RKW-Kompetenzzentrum.
- van Lente, Harro/Rip, Arie (1998): »Expectations in Technological Developments: An Example of Prospective Structures to be Filled in by Agency«, in: Cornelis Disco/Barend van der Meulen (Hg.), *Getting New Technologies Together*, Berlin/Boston: DE GRUYTER, S. 203–230.
- Vogl, Joseph (2021): *Kapital und Ressentiment. Eine kurze Theorie der Gegenwart*, München: C.H. Beck Verlag.
- White, Harrison C. (2008): *Identity and Control. How Social Formations Emerge*, Princeton: Princeton University Press.

Künstliche Intelligenz in der Zukunftsforschung

Axel Zweck, Thomas Werner

1. Einleitung¹

Seit über 100 Jahren verfolgen Denker:innen, Forscher:innen und Entwickler:innen das Ziel, Menschen und Maschinen miteinander zu verbinden. Mit zunehmender Anwendungsbreite steht heute nicht mehr die gesellschaftliche Utopie, sondern die Rolle, die Künstliche Intelligenz auf unserem Weg in eine digitale Wirtschaft und Gesellschaft einnimmt, im Mittelpunkt. Künstliche Intelligenz ändert gegenwärtig unsere Lebens- und Arbeitswelt wie jede grundlegend neue Technologie, die sich in kurzer Zeit so ubiquitär durchsetzt. Bisher ist die Frage, welche Auswirkungen dieser Wandel haben wird, noch nicht entschieden. KI gilt als zentraler Treiber der jetzigen Phase der Digitalisierung und als entscheidende Zukunftstechnologie für die wirtschaftliche Entwicklung nicht nur unseres Landes, sondern als ein künftiger Wirtschaftsmotor aller Industrienationen. Für das Jahr 2025 wird geschätzt, dass mit KI-Anwendungen weltweit ein Umsatzvolumen von ca. 30 Milliarden USD erreicht werden wird.² Länder wie die USA, China und auch Deutschland fördern KI-Forschung und deren Umsetzung in die Anwendung mit Milliardenbeträgen. China sieht in der KI eine »strategische Gelegenheit«, um die Wirtschaft des Landes in eine wissensbasierte Ökonomie zu transformieren (Groth et al. 2018). Die KI-Strategie Pekings sieht vor, dass China bis 2030 eine Vormachtstellung einnimmt und die mit KI verbundenen Industriezweige einen Wert 1,2 Billionen Euro erlangen (ebd.). Auch Deutschland will im Rahmen der »Strategie Künstliche Intelligenz« bis 2025 insgesamt etwa fünf

1 Der Redaktionsschluss für diesen Beitrag war Ende 2022, weswegen die rasante Entwicklung von ChatGPT nicht mehr berücksichtigt werden konnte.

2 Statistiken zum Thema Künstliche Intelligenz, Statista 2021, <https://de.statista.com/themen/3103/kuenstliche-intelligenz/#dossierKeyfigures>, abgerufen am 02.12.2021.

Milliarden Euro für die Umsetzung zur Verfügung stellen.³ Dies macht offensichtlich, dass dieses Thema auch für die Zukunftsforschung von besonderem Interesse ist. Wobei es in den weiteren Ausführungen weniger um die Frage geht, wie sich KI aus Perspektive der Zukunftsforschung entwickelt, als um die Frage, wie der Einsatz von KI Zukunftsforschung und Zukunftsforscher:innen beeinflussen wird.

Im vorliegenden Text wird gezeigt, dass sowohl der Begriff der Künstlichen Intelligenz als auch der Begriff der Intelligenz inhärent unscharf sind. Nach Gardners Theorien der multiplen Intelligenzen, hat menschliche Intelligenz zahlreiche Dimensionen, sodass der Intelligenzbegriff differenziert und im Kontext des jeweiligen Anwendungsfeldes betrachtet werden muss. So kann erahnt werden, wie wir Leistungen von KI-Systemen hinsichtlich ihrer »Intelligenz« zu bewerten haben. Es reicht nicht aus, nur Chancen und Risiken beim Einsatz von KI kritisch zu hinterfragen, sondern vor allem unsere eigene Erwartungshaltung einer solchen Technologie gegenüber.

Einem Computersystem bzw. einer Künstliche Intelligenz wird schnell intelligentes Verhalten zugeschrieben, sobald es menschliches Verhalten erfolgreich nachahmt. Dass diese Projektion auf KI (und das Verhalten) historische Wurzeln hat, zeigen wir anhand des »Siegeszuges der schwachen KI«.

Während »starke KIs« noch Gegenstand der Forschung sind, dringen »schwache KIs« (also spezifisch problemlösende Intelligenzen) zunehmend in menschliche Problemdomänen vor, für deren Lösungen bisher Kreativleistungen oder gar »echte Intelligenzleistungen« vermutet wurden.

Künstliche Intelligenz ist ein Instrument, um bisher als unlösbar oder sehr aufwändig gesehene Probleme über neuartige Wege zu lösen und zu denken. Sie führt in neue, auch disruptive Nutzungskontexte und verändert zugleich den technischen und methodischen Kern wissenschaftlicher Arbeitsweisen. Damit geht nicht nur ein Wandel des Wissenschaftsverständnisses einher, sondern auch des wissenschaftlichen Erkenntnisprozesses selbst.

KI macht Komplexität handhabbarer und schafft zugleich neue Komplexität. Sie erhält durch ihren bisher vorherrschenden Black Box-Charakter (Entscheidungs-)Autonomie und wandelt sich zunehmend vom Gegenstand der Betrachtung zum Instrument für Wissenschaft und Forschung. Sie ändert hierdurch auch ihre Rolle vom Instrument zur Akteur:in.

3 KI-Strategie der Bundesregierung, 2020, <https://www.bmwi.de/Redaktion/DE/Artikel/Technologie/kuenstliche-intelligenz.html>, abgerufen am 02.12.2021

KI-Systeme in Wissenschaft und Forschung charakterisieren sich immer weniger ausschließlich in ihrer Werkzeugfunktion. Sie ermöglichen fundamental neue Denkansätze und eröffnen neue Zugänge des Denkens und des Forschens. Auch die heute verfügbare schwache KI führt bereits zu neuen, sogar disruptiven Nutzungskontexten und dringt in die Zweckdimension forschenden Handelns ein. Der Mensch und auch die Wissenschaftler:in sind damit nicht mehr alleinige Akteur:in. Sie definieren ihre Rolle als handelnde Akteur:in im Zentrum von Wissenschaft und Forschung neu. Dennoch werden KIs nicht als eigenständige Akteur:innen in Begründungsdiskursen auftreten, weil ihnen die Merkmale einer Handlungsurheberschaft fehlen. Dies kann mit dem Angebot Whites für ein gestuftes Identitätsmodell nachvollziehbar gemacht werden (White 2008; Schmitt/Fuhse 2015), mit dem eine genaue Nachvollziehbarkeit der graduellen Übertragung von Handlungsträgerschaften (als Träger von identifizierbaren Kontrollprojekten, die nicht allein aus dem KI-Verfahren stammen) an KI-Verfahren möglich wird. Sie zeigt sich in einer anderen Positionierung der technischen Bestandteile in Netzwerken und Erzählungen, ohne auf die Komplexitätsstufe der Person (als integriertes Bündel von multiplen Positionierungen und Erzählungen) zu gelangen. Kontrolle wird in dieser Analyse dann an KI-Verfahren übertragen, die in deren Umsetzung dann auch mit steigenden Freiheitsgraden ausgestattet sind.

2. Der Weg zur Künstlichen Intelligenz

Doch was steckt hinter dieser wichtigen Schlüsseltechnologie? Wie wird KI einzelne Branchen, künftige Arbeitsweisen, Forschung und auch die Zukunftsforschung verändern? Wie ist eine Technologie bzw. die dahinterliegenden Algorithmen und trainierten Modelle⁴ in der Lage, mit nahezu

4 Der klassische Weg (Rechen-)Probleme mit dem Computer zu lösen, ist die Analyse des Problems derart, dass zunächst Regeln und Zusammenhänge identifiziert werden. Diese werden dann genutzt, um ein Schema zu entwickeln – den Algorithmus –, mit dem (wiederholbare, also deterministische) Rechenvorgänge durchgeführt werden, die das Ausgangsproblem lösen. Machine Learning-Modelle werden nicht programmiert, sondern zunächst mit Trainingsdaten trainiert und berechnen zu einer Eingabe auf Basis des trainierten Modells eine passende Ausgabe – auch wenn das zugrundeliegende Machine Learning-System durch Algorithmen abgebildet wird, arbeiten KI-Systeme in ihrer Wirkweise daher nicht wie klassische Algorithmen.

übermenschlicher Genauigkeit, die Leistung des Menschen in vielen Bereichen zu übertreffen? Definitionen für Künstliche Intelligenz haben eine große inhaltliche Streubreite. Im vorliegenden Text wird Künstliche Intelligenz als die Fähigkeit eines Computers verstanden, Aufgaben zu lösen, für die üblicherweise Intelligenz erforderlich ist. Solche Systeme sind in der Lage, sich wie ein Mensch »intelligent« zu verhalten und selbständig zu lernen. Die KI-Forschung hat mittlerweile eine Leistungsstärke erreicht, die selbst bei komplexeren Aufgabenstellungen beeindruckende Ergebnisse erzielt: Smartphones übersetzen Fremdsprachen in Echtzeit in die eigene Sprache und autonomes Fahren soll die Mobilität verändern sowie Sicherheit wie auch Komfort auf ein neues Niveau heben. Im Gesundheitssektor ist durch KI eine Verbesserung in Diagnostik, Früherkennung von Krankheiten und bei der Entwicklung neuer Medikamente zu erwarten (Häußling et al. 2021). Auch die rasche Entwicklung von mRNA-Impfstoffen gegen das Covid-19-Virus wurde durch den Einsatz von KI erst ermöglicht (Baidu 2020). Facebook entwickelt eine KI, die eine eigene, für Menschen unverständliche »KI-Sprache« erfindet und Wissenschaftler:innen entwickeln einen Algorithmus, der automatisch Fake News generiert, sich aber aus ethischen Gründen weigern, diesen zu veröffentlichen – gleichzeitig kann dieser Algorithmus dazu eingesetzt werden, um Fake News aufzudecken (Giansiracusa 2021).

Doch nicht überall, wo KI draufsteht, ist auch KI drin und nicht alles was intelligent erscheint (sei es durch die geschickte Nutzung regelbasierter Systeme oder das Erzeugen überraschender Ergebnisse), ist wirklich intelligent im menschlichen Sinne. Nach einer Untersuchung von MMC Ventures (Kelnar/Kostadinov 2019) haben 40 Prozent aller europäischen Neugründungen, die als »KI-Unternehmen« klassifiziert werden, in Wirklichkeit nichts mit Künstlicher Intelligenz zu tun. Sie nutzen das Label KI zur Erlangung von Fördermöglichkeiten und Kund:innen. Zur besseren Einordnung solcher Beobachtungen und wie künstliche mit menschlicher Intelligenz verglichen werden kann, lohnt ein vertiefter Blick auf die Geschichte und die Technologie hinter der Künstlichen Intelligenz.

Was ist Intelligenz?

Die inhärente Unschärfe in der Definition des Begriffes »Künstliche Intelligenz« liegt zum einen in der Historie, zum anderen im Verständnis des Begriffes von Intelligenz im Allgemeinen. Der Psychologe und Erziehungswissenschaftler Howard Gardner (Gardner 1983) vertritt eine Theorie multipler

Intelligenzen. Über die Beobachtung von Menschen mit Inselbegabungen und der Untersuchung historisch herausragender Persönlichkeiten beschreibt Gardner Intelligenz als Zusammenspiel von acht Dimensionen: Die *sprachlich-linguistische Intelligenz* umfasst die Sensibilität für gesprochene und geschriebene Sprache sowie die Fähigkeit, Sprachen zu lernen und zu bestimmten Zwecken zu gebrauchen. Berufsgruppen wie Rechtsanwält:innen, Redner:innen, Schriftsteller:innen und Dichter:innen wird eine hohe sprachlich-linguistische Intelligenz nachgesagt. Zur *logisch-mathematischen Intelligenz* zählt die Fähigkeit, Probleme nach logischen Gesichtspunkten zu analysieren und mathematische Operationen einzusetzen, um wissenschaftlichen Fragen nachzugehen. Mathematiker:innen, Logiker:innen, Programmierer:innen und Naturwissenschaftler:innen machen von dieser Form der Intelligenz besonders Gebrauch. *Musikalische Intelligenz* bedeutet die Begabung zu Musizieren, zu Komponieren und das Verständnis für musikalische Prinzipien. Der *räumlichen Intelligenz* wird der theoretische und praktische Sinn für die Strukturen großer Räume zugeordnet. Seeleuten und Pilot:innen, Bildhauer:innen, Chirurg:innen, Schachspieler:innen, Ingenieur:innen, Graphiker:innen oder Architekt:innen nutzen besonders diese Form der Intelligenz. Die *körperlich-kinästhetische Intelligenz* beschreibt die Fähigkeit den Körper und Teile davon zur Problemlösung oder zur Gestaltung von Produkten einzusetzen. Neben Vertreter:innen technischer Berufe wie Handwerker:innen, Chirurg:innen und Mechaniker:innen zählen hierzu auch Tänzer:innen, Schauspieler:innen und Sportler:innen. Die *naturalistische Intelligenz* beschreibt die Fähigkeit, Naturphänomene zu beobachten, zu unterscheiden und zu erkennen. Typische Vertreter:innen für den Einsatz dieser Fähigkeiten sind Naturforscher:innen, Umweltspezialist:innen, Tierärzt:innen und Köch:innen. Die *interpersonale Intelligenz*, die nach den Psychologen Edward Lee Thorndike und David Wechsler auch als soziale Intelligenz (Wechsler 1964) bezeichnet wurde, bezeichnet die Fähigkeit, unausgesprochene Gefühle und Motive anderer Menschen nachzuempfinden und deren Emotionen zu beeinflussen. Eine Fähigkeit, die wesentliche Voraussetzung für den erfolgreichen Umgang mit anderen Menschen ist und laut Gardner bei politischen oder religiösen Führer:innen, Eltern und Lehrer:innen sowie bei beratenden oder heilenden Berufen besonders stark ausgeprägt ist.

Als letzte der acht Dimensionen der Intelligenz wird die *intrapersonelle Intelligenz* angeführt. Die Fähigkeit, die eigenen Gefühle und Motive zu verstehen und letztlich auch zu beeinflussen. Eine Art »interne Selbsterkenntnis« die Menschen zu richtigen Entscheidungen verhilft und besonders bei Schriftstel-

ler:innen, Schauspieler:innen und Künstler:innen ausgeprägt ist. Gerade die beiden letzten Intelligenztypen sind Basis der Theorien der Psychologen John D. Mayer und Peter Salovey zur »Emotionalen Intelligenz« (Mayer et al. 1990: 772f.; Mayer/Salovey 1990: 185f.), die später auch durch Daniel Golemans Buch »Emotionale Intelligenz« (Goleman 1995) popularisiert wurde.

Obgleich Gardners Theorie der multiplen Intelligenz von der wissenschaftlichen Intelligenzforschung aufgrund empirischer Mängel kritisch gesehen wird (Süß 2011: 97f.), zeigt dieser Exkurs die Vielfältigkeit des Intelligenzbegriffs. Es wird deutlich, dass es keine universelle Antwort auf die Frage nach der Messbarkeit von Intelligenz geben kann. Die zahlreichen existierenden Testverfahren zur Messung von Intelligenz und die damit einhergehende Kritik an den einzelnen Verfahren stützen diese Einschätzung.

Eine weit größere Herausforderung scheint die Beurteilung, ob ein Computersystem intelligent ist. Der britische Mathematiker Alan Turing hat 1950 ein Verfahren – den Turing-Test – vorgeschlagen (Turing 1950: 433f.), um zu entscheiden, ob ein System sich intelligent verhält bzw. ob eine Maschine ein dem Menschen gleichwertiges Denkvermögen hat. Der Testablauf kann im Wesentlichen wie folgt beschrieben werden: Eine menschliche Fragesteller:in führt über eine Tastatur und einen Bildschirm, also ohne Sicht- und Hörkontakt, eine Unterhaltung mit zwei ihm unbekanntem Gesprächspartner:innen. Ein(e) Gesprächspartner:in ist ein Mensch, die andere eine Maschine. Kann die Fragesteller:innach intensiver Befragung nicht entscheiden, welcher der Gesprächspartner:innen die Maschine ist, hat die Maschine den Turing-Test bestanden. Es wird der Maschine dann ein dem Menschen gleichrangiges Denkvermögen unterstellt. Turing prophezeite in seiner Schrift von 1950, dass es in 50 Jahren nicht mehr als 70 Prozent der Fragesteller:innen möglich sei, zwischen Programm und Mensch zu unterscheiden. Aus der Interpretation dieses Postulats von Turing resultiert die mutmaßliche 30-Prozent-Grenze, nach der der Turing-Test als bestanden gilt, wenn mindestens 30 % der menschlichen Teilnehmer:innen getäuscht werden können. Der klassische Turing-Test und die bisherigen Versuche, ihn zu bestehen, lässt allerdings vermuten, dass dieses Verfahren eher das Potenzial hat, erfolgreiches Täuschen zu bewerten, er teste »eher menschliche Leichtgläubigkeit als echte künstliche Intelligenz« (Tegmark 2017: 138f.).

Legendär wurde auch das von Joseph Weizenbaum 1966 entwickelte Computerprogramm ELIZA. Der Informatikprofessor am Massachusetts Institute of Technology (MIT) wollte mit einem einfachen, regelbasierten System die Möglichkeiten der Kommunikation zwischen Mensch und Maschine über

natürliche Sprache demonstrieren. Er simulierte (aus verschiedenen Gründen der Vereinfachung und Generalisierung) einen Psychotherapeuten als Gesprächspartner. Zu Weizenbaums Überraschung war ein Großteil der Versuchspersonen überzeugt, dass ihr Gesprächspartner ein tatsächliches Verständnis für ihre Probleme entwickelt habe. Da das Programm aufgrund der zugrundeliegenden Regeln konstruiert wurde, um menschlich, intelligent und einfühlungsvermögend zu wirken – also zu täuschen –, weigerten sich viele Proband:innen nach dem Experiment zu akzeptieren, dass ihr Gesprächspartner lediglich ein Computersystem war. Weizenbaum, der ursprünglich zeigen wollte, dass ein echter, vertiefter Dialog zwischen Mensch und Maschine nicht möglich sei, sondern stets oberflächlich bliebe, war von diesem Ergebnis erschüttert (Schanze 2010). Das Programm löste statt einer kritischen Haltung einen gegenteiligen Effekt aus: Menschen begannen der Maschine Gefühle und Verständnis zuzuschreiben. Der so genannte ELIZA-Effekt war geboren. Weizenbaum schrieb dazu später: »I had not realized [...] that extremely short exposures to a relatively simple computer program could induce powerful delusional thinking in quite normal people.« (Weizenbaum 1976: 7f.)

2008 traten bei einem Experiment an der University of Reading sechs Computerprogramme an, um den Turing-Test zu bestehen. Dem besten Programm gelang es, 25 Prozent der menschlichen Versuchsteilnehmer:innen zu täuschen. 2011 konnte die Software Cleverbot beim »Technique 2011« am indischen Institut IIT Guwahati sogar 59 Prozent der Teilnehmer:innen täuschen.

2014 titulierte die internationale Presse, dass der Chatbot Eugene Goostman auf einer Veranstaltung der Royal Society in London den Turing-Test bestanden hat. 33 Prozent der Jury-Mitglieder:innen konnten die Handlungen des Chatbots nicht von den Handlungen einer menschlichen Person unterscheiden. Ein denkwürdiger Meilenstein in der KI-Forschung galt damit als überwunden. Doch das Experiment warf Fragen auf. So wurde Kritik am Versuchsaufbau wie auch »kognitiven Kniffen und Programmiertricks«, die verwendet wurden, um die Prüfer:innen zu täuschen, laut (Kühl 2021).

Mit Blick auf das primäre Ziel zu täuschen, den grundsätzlichen Mängeln des Turing-Test zu begegnen und die bereits beschriebenen multiplen Dimensionen von Intelligenz zu berücksichtigen, wurden im Folgenden Konzepte entwickelt, den Turing-Test zu erweitern. Zu diesen gehören der Lovelace-Test (Bringsjorden et al. 2003) mit der Forderung, eine KI müsse »Kreativität beweisen und originäre Leistungen erbringen«. Beim Metzinger-Test (Met-

zinger 2001), erfüllt die KI den Intelligenzanspruch, wenn sie mit eigener Argumentation in die Diskussion um künstliches Bewusstsein eingreift und überzeugend für ihre eigene Theorie des (eigenen) Bewusstseins argumentieren könne. Bemerkenswert ist zudem, dass heute auch »umgekehrte« Turing-Tests eingesetzt werden. Mit den sogenannten CAPTCHAs (»Completely Automated Public Turing test to tell Computers and Humans Apart«) (von Ahn et al. 2003) werden Menschen im Internet bei der Nutzung von Eingabefeldern aufgefordert den Beleg zu erbringen, dass sie keine Maschine bzw. keine Künstliche Intelligenz sind.

Es scheint ein fundamentaler Denkfehler des Menschen zu sein, einen Computer wie einen Menschen zu betrachten und zu behandeln, ihn nach menschlichen Maßstäben messen zu wollen und ihm ein menschenähnliches Verständnis der Welt zuzuschreiben. Laut dem Philosophen John Searle kann die formale Struktur der Mathematik niemals den informellen Charakter menschlicher Gedanken erfassen, oder, einfacher ausgedrückt, menschliche Intelligenz kann grundsätzlich nicht durch ein Computerprogramm simuliert werden. Searle illustrierte diese Auffassung anhand eines Gedankenexperiments mit dem Namen »Das chinesische Zimmer« (Searle 1980). In diesem hypothetischen Experiment sitzt eine Person mit einigen Texten auf Chinesisch in einem geschlossenen Raum. Diese Person kann Chinesisch weder sprechen oder schreiben noch ist sie in der Lage, die chinesischen Schriftzeichen als solche zu verstehen. Durch einen Schlitz in der Wand werden der Person auf Papier Geschichten auf Chinesisch zugeschickt, ergänzend erhält sie Fragen zu der Geschichte, ebenfalls auf Chinesisch. Außerdem erhält diese Person ein »Handbuch« in ihrer Muttersprache. Diese Anleitung erlaubt es der Person anhand der erhaltenen Symbole, also der Geschichte und der Fragen, eine Antwort ebenfalls auf Chinesisch zu schreiben. Sie folgt hierbei aber ausschließlich den Anweisungen, also den Regeln aus der Anleitung und versteht die Antworten, die sie anschließend durch den Schlitz wieder nach draußen schiebt, nicht. Außerhalb des Raumes nimmt eine chinesische Muttersprachler:in die Antworten auf ihre Fragen zur Geschichte entgegen und kommt zwangsläufig zu dem Schluss, dass sich im Raum jemand befindet, der ebenfalls chinesisch spricht. Anhand dieses Aufbaus argumentiert Searle, dass ein Programm, das den Turing-Test besteht, nicht zwangsläufig auch intelligent sein muss, es ist nur scheinbar intelligent. Ein korrektes Ausführen von Anweisungen setzt daher weder Verständnis noch Bewusstsein voraus.

Siegeszug der schwachen KI

Was verleitet nun den Menschen dazu, Computer hinsichtlich unserer Wahrnehmung wie Menschen zu behandeln und Intelligenz in sie hineinzuprojizieren? Diese Frage ist auch interessant in Zusammenhang mit dem Wunsch des Menschen, ebenbürtige künstliche Systeme zu schaffen.

Historisch lässt sich ein Muster des Strebens nach künstlichen Kopien des Menschen erkennen. Sei es der Homunculus, der Golem, Frankenstein, später mechanische Automaten zur Mechanisierung des Menschen und seines Verhaltens bis hin zum Human Brain Project⁵, einem Forschungsprojekt der Europäischen Kommission, welches das gesamte verfügbare Wissen über das menschliche Gehirn zusammenfassen und mittels computerbasierter Modelle in Simulationen nachbilden will. Also eine ultimative Reproduktion und Imitation menschlichen Denkens.

Die Frage ist aber: Ist eine Abbildung menschlicher Denkweisen in einer Maschine überhaupt möglich? Ein Blick auf die Entstehungsgeschichte der KI zeigt, dass hierfür mindestens zwei Aspekte notwendig sind: eine formale Sprache (Denkweise/Logik/Algorithmus), um kognitive Prozesse zu beschreiben, und ein Medium (Rechenmaschine), um diesen Formalismus zu vollziehen. Gemäß dem Motto »Wer über die Zukunft reden will, sollte die Vergangenheit kennen« (Zweck 2021: 69), lohnt daher ein detaillierterer Blick auf die Historie der Künstlichen Intelligenz:

Meilensteine in der Entwicklung der Künstlichen Intelligenz

17. Jhrd.: Wilhelm Leibniz vertrat als erster die Auffassung, dass alles in der Welt, einschließlich Gehirn und Denken, den physikalischen Gesetzen folgt. Er baute als erster ein mechanisches Gerät, das geistige Operationen simulieren sollte.

1763: Das Bayes-Theorem von Thomas Bayes wird veröffentlicht; es handelt sich um eine Formel, die für das Schließen unter unsicheren Wissensbedingungen für die KI bedeutsam ist.

5 Human Brain Project, <https://www.humanbrainproject.eu/en/>, abgerufen am 20.03.2022

1769: Wolfgang von Kempelen formuliert die Idee einer intelligenten Maschine und entwickelt den »Schachtürken«, einen scheinbar, mechanischen Schachspielerroboter, in dessen Inneren jedoch ein menschlicher Schachspieler versteckt war.

19. Jhrd.: Philosophen und Mathematiker wie Boole oder Frege entwerfen mit der Logik einen formalen Apparat, mit dem sich auch kognitive Prozesse abbilden lassen.

1921: Das Deutsche Patentamt segnet eine »Vorrichtung zur Beantwortung von Fragen« ab; Carl Spierer entwarf darin ein durchdachtes Logiksystem für das was heute Künstliche Intelligenz heißt.

1931: Kurt Gödel zeigt die Grenzen der Logik (Berechenbarkeit) auf und weist gleichzeitig nach, dass jedes berechenbare Problem lösbar und damit das menschliche Gehirn mit Computern abbildbar ist. Es gibt jedoch fundamentale Grenzen dessen, was für Rechner entscheidbar ist.

1936: Alan Turing stellt sein universelles Rechenmodell vor, mit dem jedes berechenbare Problem lösbar ist. Die Turing-Maschine bietet eine Basis, auf der heutige Computer noch arbeiten.

1948: Norbert Wiener prägt mit seinem Werk »Cybernetics or Control and Communication in the Animal and the Machine« den Begriff der Kybernetik und schuf damit die mathematischen Grundlagen für die Kontrolltheorie und Regelungstechnik zur Berechnung der Dynamik und Stabilität von rückgekoppelten Systemen.

1950: Alan Turing veröffentlicht in der englischen Philosophiezeitschrift »Mind« den Artikel »Computing Machinery and Intelligence« (Turing 1950) und stellt die Frage, ob Computer denken können. In seinem Aufsatz beschrieb Turing eine Prüfung ihrer geistigen Fähigkeiten und entwickelt den beschriebenen Turing-Test.

1952: Arthur Samuel (IBM) entwickelt Computerprogramme, die Dame spielen. Im Zuge der Übertragung einer Partie im Fernsehen, stieg der Kurswert der IBM-Aktie um 15 Punkte.

1956: Auf der Dartmouth-Konferenz prägt J. McCarthy den bis heute verwendeten Begriff »Künstliche Intelligenz«.

1957: F. Rosenblatt entwickelt das Perzeptron, ein vereinfachtes künstliches neuronales Netzwerk⁶, das nach Training Muster erkennen kann.

1957: Noam Chomsky entwickelt eine Universalgrammatik, eine mathematische Beschreibungsform für die natürliche Sprache und postuliert, dass sich menschliche Sprache streng regelbasiert, wie ein Computerprogramm/Algorithmus verhält.

1960: Karl Steinbuch erfindet die Lernmatrix, ein frühes neuronales Netz mit analoger Elektronik.

1964: Joseph Weizenbaum programmiert den schon beschriebenen Chatbot Eliza.

1965: Edward Feigenbaum entwickelt das weltweit erste Expertensystem.

1969: Marvin Minsky weist nach, dass Perzeptrons keine nichtlinearen Zusammenhänge darstellen können und leitet damit den so genannten KI-Winter ein, eine Phase in der die Auseinandersetzung mit KI in einem quasi-Winterschlaf erstarre und kaum ernst zu nehmende Fortschritte erreicht werden.

1980: Fukushima schlägt mit dem Neocogitron das erste Convolutional Neural Network (CNN) vor. Die Struktur von CNNs ist von biologischen Prozessen inspiriert und gilt als Grundlage heutiger Deep Learning-Verfahren für maschinelles Lernen.

1984: David Rumelhart und James McClelland erzielen mit Backpropagation, einem Verfahren zur Fehlerrückführung bei künstlichen neuronalen Netzen, einen Durchbruch bei der Entwicklung von neuronalen Netzen.

1986: Geoffrey Hinton setzt Backpropagation zur Mustererkennung ein und begründet den Konnektionismus, also die Entwicklung und Analyse informationsverarbeitender Systeme, die wesentliche Eigenschaften kognitiver Prozesse nachbilden.

1987: Sepp Hochreiter entwickelt das LSTM (Long short-term memory). Ein Kurzzeitgedächtnis für neuronale Netze, das den Durchbruch der künstlichen Intelligenz, wie wir sie heute verstehen, erst ermöglichte.

1987: Yann LeCun setzt Backpropagation erfolgreich beim Training von Convolutional Networks ein.

2000: Jürgen Schmidhuber und Sepp Hochreiter verbessern RNN (»Rekurrente neuronale Netzwerke«) mit Hilfe des LSTM und bezeichnen diese RNNs als Deep Learning-Netzwerke⁷.

2014: Yoshua Bengio et al. stellen das Konzept der Generativen Adversarialen Netzwerke (Goodfellow et al. 2014) vor – die Methode wird auch als Turing-Lernen bezeichnet.

Der Blick auf diese Historie wesentlicher Meilensteine zeigt, dass die Entwicklung der Künstlichen Intelligenz mehr als nur eine technische Realisierung informationsverarbeitender Maschinen ist. Es geht um ein Zusammenspiel von Mathematik/Logik, Sprache und Philosophie. Die Philosophie leistete und leistet zur KI-Entwicklung insofern einen wichtigen Beitrag, als sie die Auffassung vertritt, dass Denken etwas Maschinelles ist, das auf Wissen basiert. Wissen, das in einer internen Sprache repräsentiert ist und das Denkprozesse ermöglicht, die gewünschte Handlung zu vollziehen. Mathematik und Logik liefern Werkzeuge zur Handhabung von sicheren wie von unsicheren, wahrscheinlichkeitstheoretischen Aussagen. Sie schufen erst Grundlagen

6 Künstliche neuronale Netzwerke bilden, entsprechend ihrem biologischen Vorbild, Netze aus künstlichen Neuronen nach. Sie stellen, als Forschungsgegenstand der Neuroinformatik, einen Zweig der künstlichen Intelligenz dar.

7 Deep Learning ist eine Methode zur Informationsverarbeitung und verwendet künstliche neuronale Netze (KNN) die mehrere Zwischenschichten zwischen Eingabe- und Ausgabeschicht einsetzen.

für das Schlussfolgern mit Hilfe von Algorithmen, d.h. regelbasierten Abläufen. Aus der Psychologie kommt die Vorstellung, dass Menschen und Tiere als informationsverarbeitende Maschinen betrachtet werden können – eine Betrachtung, die auch von der Linguistik geteilt wird. Elektronik und vor allem die Computertechnik lieferten die Hardwaregrundlagen für KI-Anwendungen und seit einigen Dekaden die notwendigen Infrastrukturen, um die hinsichtlich Speicher und Rechenzeit immer größer werdenden KI-Programme ausführen zu können. Die historische Entwicklung zeigt die Fortschritte in der KI-Forschung und scheint auf den ersten Blick kontinuierlich zu verlaufen. Tatsächlich verlief sie eher wellenartig in Form von Fortschritten und Rückschlägen. So waren die Jahre 1950–1970 geprägt von Aufbruchstimmung und Euphorie. Die Jahre 1965–1975 waren zunehmend von Rückschlägen und ungelösten Herausforderungen mit der konkreten Realisierung einer KI geprägt. In der Zeit von 1980–1990 entstanden völlig neuartige Zugänge, Anwendungen und KI-Industrien. Mit der Verfügbarkeit günstigerer Speicher und Rechenzeiten erfuhren die neuronalen Netze seit 1985 eine Renaissance bis zur Gegenwart. Fortschritte in jüngster Zeit ermöglichen ein besseres Verständnis der theoretischen Grundlagen von Intelligenz und brachten verbesserte Fähigkeiten realer Systeme mit sich. Heute ergeben sich daraus zwei grundsätzlich unterschiedliche Perspektiven bezüglich des Verständnisses und der Zielsetzung für KI. Die eine Strategie versucht zu verstehen, wie das menschliche bzw. das biologische Denken funktionieren und diese Einsichten nach bionischen Prinzipien abzubilden. Die andere stellt die Frage, ob von einer grundlegenden Neuformulierung des KI-Konzeptes ausgegangen werden muss, die bezogen auf das Verständnis von Intelligenz einen gänzlich neuen, eigenständigen Weg geht.

Der vorsichtige Versuch, die bisherige Historie in die Zukunft fortzuschreiben, lässt absehbar längerfristig weitere Fortschritte erwarten. Doch bereits gegenwärtig sind KI-Systeme – insbesondere künstliche neuronale Netze – zu außergewöhnlichen (Kreativ-)Leistungen fähig. Das gilt, obwohl heutige Künstliche Intelligenz lediglich den Kriterien einer sogenannten »schwachen KI« entsprechen. »Schwache KIs« besitzen per definitionem keine Kreativität nach menschlichem Verständnis und keine explizite Fähigkeit, autonom im universellen Sinne zu lernen. Die Lernfähigkeiten solcher Systeme sind in der Regel auf das Trainieren von Erkennungsmustern oder das Abgleichen und Durchsuchen großer Datenmengen reduziert. Sie bewältigen komplexe, aber klar definierte Aufgabentypen, sie folgen einer festgelegten Methodik und sind in der Lage, wiederkehrende und genau

spezifizierte Probleme zu lösen. Zu den Einsatzbereichen schwacher KI zählen die Automatisierung von Prozessen aber auch Spracherkennung und -verarbeitung. Weitere Beispiele sind Text- und Bilderkennung, Spracherkennung, Übersetzungen, Navigationssysteme etc., aber auch verbreitete digitale Assistenzsysteme wie Alexa, Siri oder Google Assistant.

Im Gegensatz zur spezifisch problemlösenden Intelligenz schwacher KIs kann eine starke KI Aufgabenstellungen selbstständig erkennen und definieren und hierfür »eigenständig« Wissen aus der entsprechenden Anwendungsdomäne zusammentragen. Sie untersucht und analysiert Probleme, um adäquate neue oder kreative Lösungen für Herausforderungen zu finden. Obwohl heutige KI-Systeme den Anschein erwecken, bereits in die Domäne der starken KI vorgedrungen zu sein, ist ihre Realisierung keineswegs in greifbarer Nähe und noch Gegenstand der Forschung.⁸

Frei nach Arthur C. Clarke formuliert: »Jede hinreichend fortschrittliche Technologie ist von Zauberei nicht zu unterscheiden.« (Clarke 1962: 14) zeigt die »Magie« heutiger, spezifisch problemlösender Intelligenzen im Sinne schwacher KIs aber, wie schnell ihnen nicht nur Kreativleistungen, sondern unberechtigterweise und verfrüht »echte Intelligenzleistungen« zugesprochen werden. Dies geschieht insbesondere, wenn sie menschliche Fähigkeiten in bestimmten Aufgabenstellungen übertrumpfen. Sie fallen dann auch mehr und mehr aus dem Bereich bio-physikalischer Regularitäten heraus, den White ja auch explizit aus sozialwissenschaftlicher Analyse ausschließen möchte.

In den letzten 50er Jahren des vorherigen Jahrhunderts besiegten Computerprogramme menschliche Spieler:innen im Brettspiel Dame. Nicht einmal 50 Jahre danach, 1977, wurde der amtierende Schachweltmeister Garri Kasparov vom IBM-Supercomputer DeepBlue im Schach geschlagen. Die verwendete Technik – ein speziell für diesen Zweck gebautes »Rechenmonster«, dass in jeder Sekunde des Spiels durchschnittlich 126 Millionen Stellungen durchrechnet – besiegte den Menschen in einem Spiel, dem von je her nachgesagt⁹

8 Generalisiert anwendbare KI: Hype und Stand der Forschung, science media center germany, URL: <https://www.sciencemediacenter.de/alle-angebote/science-responses/details/news/generalisiert-anwendbare-ki-hype-und-stand-der-forschung/>, abgerufen am 04.12.2021.

9 The mathematical complexity of Go, <https://senseis.xmp.net/?ComplexityOfGo>, abgerufen am 04.12.2021.

wurde, nur mit logischem Denken, Intelligenz, Wagemut und strategischem Scharfsinn beherrschbar zu sein.

Wiederum 20 Jahre später besiegte das auf maschinellem Lernen basierende Computerprogramm AlphaGo den amtierenden Go-Weltmeister in einem Spiel, das als das älteste und komplexeste Brettspiel der Welt gilt. Go ermöglicht mehr Züge als Atome im bekannten Universum existieren und mit traditionellen Strategien/Algorithmen (wie sie noch DeepBlue verwendet hatte) wäre kein Computer in der Lage gewesen, dem Menschen auch nur annähernd ebenbürtig zu sein. AlphaGo siegte mit völlig neuen Spielweisen und Lösungsstrategien, die im Nachhinein als kreativ empfunden wurden. In einem berühmten Spiel gegen den möglicherweise besten Go-Spieler aller Zeiten (den Südkoreaner Lee Sedol) verwendete AlphaGo einen Spielzug, den menschliche Beobachter:innen zunächst als Fehler abtaten (Menick 2016), der sich jedoch im weiteren Spielverlauf als für den Sieg entscheidend entpuppte. Der später als »Zug 37« bekannt gewordene Zug, hat viele Expert:innen überrascht, neben der internationalen Go-Gemeinschaft auch die Erschaffer:innen von AlphaGo selbst. Dieser Spielzug lag so weit außerhalb dessen, was seit über tausend Jahren gewohnte Wege waren, Go zu spielen, dass er zum Inbegriff von Kreativität in der KI wurde. Zahlreiche wissenschaftliche Abhandlungen untersuchen seitdem diesen Spielzug und das spezifische Verhalten von AlphaGo in der damaligen Spielsituation.¹⁰

2017, zwei Jahre nach AlphaGo, stellte das Unternehmen DeepMind den Nachfolger von AlphaGo, AlphaZero vor. War AlphaGo noch auf das Spielen des Brettspiels Go trainiert (zur Erinnerung, schwache KIs sind in der Regel auf eine dedizierte Problemklasse trainiert), ist AlphaZero ein autodidaktisches Computerprogramm, dessen Algorithmus mehrere komplexe Brettspiele einzig anhand der Spielregeln und Siegbedingungen sowie durch intensives Spielen gegen sich selbst erlernt. AlphaZero ist in der Lage, auch andere Strategiespiele wie Schach oder Shogi zu beherrschen und wurde mit Methoden des sogenannten *Reinforcement Learning* (kurz RL und zu Deutsch: bestärkendes Lernen oder verstärkendes Lernen) entwickelt. Es handelt sich um einen Ansatz des maschinellen Lernens, bei dem die KI eigenständig eine Strategie erlernt, um »Belohnungen« aus einer programmierten Belohnungsfunktion zu maximieren. Anders als beim überwachten Lernen (Supervised Learning) sind beim RL im Vorfeld keine Daten erforderlich. Stattdessen erfolgt die

10 AlphaGo's Games, <https://www.alphago-games.com/view/eventname/leesedol/game/1/move/37>, abgerufen am 09.01.2022.

Bildung der Datenbasis durch ausführliche Trial-and-Error-Abläufe innerhalb eines eigens angelegten Simulations-Szenarios. Die Trainingsdauer, um AlphaZero das Schachspielen beizubringen, betrug nur wenige Stunden – auf das komplexe Brettspiel Go wurde das System mit weniger als 35 Stunden Rechenzeit trainiert.

Die Reaktionen verschiedener Schachweltmeister:innen war entsprechend überschwänglich. Schachweltmeister Garry Kasparow kommentierte: »It's a remarkable achievement, even if we should have expected it after AlphaGo. We have always assumed that chess required too much empirical knowledge for a machine to play so well from scratch, with no human knowledge added at all.« (Kasparov 2017)

Andere Großmeister:innen beschrieben die ungewohnte Spielweise als »insane«, »bahnbrechend« oder »phänomenal«. Professor Wattenhofer von der ETH Zürich kommentierte die Leistungen mit:

»AlphaZero spielt sehr ungewöhnlich, nicht wie ein Mensch, aber auch nicht wie ein typischer Computer, sondern mit ›echter künstlicher‹ Intelligenz. [...] AlphaZero's Leistung zeigt, dass wir die Künstliche Intelligenz ernst nehmen müssen. Sie beweist, dass wir Zeitzeugen eines gewaltigen Wandels sind. [...] AlphaZero & Co. werden die Gesellschaft nachhaltig verändern.« (Wattenhofer 2018)

Ähnlich erstaunliche Leistungen können beim Einsatz von KI im Bereich der pharmazeutischen Wirkstoffforschung beobachtet werden. Zu den fundamentalsten Herausforderungen der Molekularbiologie zählt das Proteinfaltungsproblem. Proteine bestehen aus Hunderten von Aminosäuren, die sich zu komplexen Strukturen zusammenfalten. Welche Aminosäuren zu einem Protein gehören, wird zwar in der DNS codiert, aber diese Codierung gibt keinen Aufschluss über die räumliche Struktur, in die sich die Aminosäurekette zu einem Protein mit spezifischer Form faltet. Diese 3D-Struktur des Moleküls ist jedoch entscheidend für seine Funktion im Organismus. Sofern bekannt, wird die Funktionsweise des Proteins nachvollziehbar. Die Frage mit welchen Wirkstoffen (Medikamenten) das Protein ggf. beeinflussbar ist, ist dann deutlich leichter zu beantworten. Ein bloßes Ausprobieren dauert bei der Strukturvorhersage von Proteinen sehr lange, obwohl sich Proteine in der Natur innerhalb weniger Minuten formieren. Im Dezember 2020 überraschte DeepMind – eine KI-Tochter von Google – die Fachwelt der Biologie, als sie mit AlphaFold, einem KI-Tool zur Vorhersage von Proteinstrukturen,

den schon seit 50 Jahren laufenden wissenschaftlichen Wettbewerb CASP (»Critical Assessment of Techniques for Protein Structure Prediction«) gewann. AlphaFold schnitt bei dem alle zwei Jahre stattfindenden Event so gut ab, dass der Mitbegründer des Wettbewerbs erklärte, dass »[...] in gewisser Weise das Problem als gelöst betrachten werden kann.« (Callaway 2020). Das Fachmagazin Science schätzte den neuen Ansatz für so bedeutsam ein, dass es ihn zum wissenschaftlichen Durchbruch des Jahres kürte (Service 2021). Eine Künstliche Intelligenz erzielt mit enorm hoher Geschwindigkeit und Genauigkeit einen wissenschaftlichen Durchbruch und revolutioniert die Wirkstoffforschung. Ein Erfolg jedoch ohne die Mechanismen geklärt zu haben, nach denen die KI dies bewerkstelligt, geschweige denn die Natur dieses komplexe Problem in nur wenigen Minuten lösen kann. So lässt sich vermuten, dass wir eine Wissenschaftsvorstellung ohne Theorie erleben und künftig möglicherweise auf ein massives Theoriedefizit zusteuern. Die Nutzer:in:erkennt, dass die Dinge funktionieren, verliert aber die Fähigkeit zu erklären, warum dies so ist. Wissenschaft produziert hier also weniger eine Entzauberung (Weber 1919) als vielmehr eine erneute Verzauberung der Welt – um erneut auf Clarke zu rekurrieren und den Kreis zu schließen: »Jede hinreichend fortschrittliche Technologie ist von Zauberei nicht zu unterscheiden.« (Clarke 1962) Unklarheiten wie diese Ergebnisse beim Einsatz von KI zu Stande kommen, stellen eine neue Dimension des Umgangs mit wissenschaftlichen Verfahren dar und erfordern eine intensivere Auseinandersetzung mit den daraus sich ergebenden Fragen, insbesondere in Bezug auf unser Wissenschaftsverständnis.

3. KI in der Wissenschaft

KI-Anwendungen sind im privaten wie industriellen Umfeld bereits ubiquitär im Einsatz. Auch in Wissenschaft und Forschung finden sich zahlreiche Einsatzmöglichkeiten, bei denen KI als Werkzeug etabliert ist, wie

- Mustererkennung und Simulationen in der Klimaforschung
- Data Mining, wie z.B. beschrieben in »Big Data für die Sozialforschung« (Scheiermann/Zweck 2014)
- Wirkstoffforschung und Medikamentenentwicklung in der Pharmazie
- Prognosen und Simulationen in der Epidemiologie

- Theorieentwicklung in Mathematik und Teilchenphysik unter anderem mit Hilfe automatisierter Theorembeweiser.

Der gesteigerte Einsatz von KI lässt neue Möglichkeiten für die Forschung erwarten. So können bestimmte Forschungsaktivitäten, die bislang als nicht praktikabel galten, erst mit Hilfe von KI effizient durchgeführt werden. Hierzu zählt die Analyse großer Textmengen oder (Forschungs-)Datenbanken, um per Extraktion und Korrelation aus Daten neue Erkenntnisse zu gewinnen. Außerdem erlaubt der Einsatz von KI Analysen, die bislang grundsätzlich nicht möglich waren. Die bereits beschriebene Proteinfaltung kann hier ebenso als Beispiel dienen, wie die individualisierte Genomanalyse. KI unterstützt bei Modellbildung und Hypothesenfindung, sie deckt Muster auf, die über bisher verborgene kausale Erklärungen zur Erkenntnisbildung beisteuern.

Wie verändert sich das Wissenschaftsverständnis durch KI- und Machine Learning-Verfahren? Steht eine methodologische Revolution bevor, oder wird die bisherige Kontinuität im Wissenschaftsbetrieb – nur mit modifizierten Instrumenten – fortgesetzt?

Neue datengetriebene Verfahren dienen hier quasi als Katalysatoren. KI verändert die Wissensgenese und zugleich das, was Wissenschaft ist. Aktuell zeichnet sich für das künftige Wissenschaftsverständnis eine Verschiebung des wissenschaftlichen Arbeitsstils (im Sinne Whites als einer Veränderung von stochastischen Verteilungen hin zu neuen Dominanzen mit erkennbarer Signalwirkung, vgl. Schmitt/Fuhse 2015) weg von einer Erklärungsorientierung hin zu einer pragmatischen Fokussierung auf die reine Erkenntnisgewinnung ab. Und zwar hin zu Nutzen und Anwendbarkeit sowohl für die Forschung als auch für die Gesellschaft. Das Beispiel AlphaFold zeigt, dass das Leistungsmerkmal der Prognose (hier die Vorhersage der räumlichen Struktur eines Eiweißes auf der Basis bekannter Aminosäuresequenzen) in den Mittelpunkt gestellt wird, indem bei KI- und Machine Learning-Verfahren (verborgene) schwache Kausalitäten oder Regelmäßigkeiten und starke Korrelationen zugelassen werden, um diese für praktische Zwecke nutzbar zu machen. KI- und Machine Learning-basierte Prognoseverfahren basieren auf statistischen bzw. probabilistischen Prognosen, die auf Basis einer Datenanalyse vorgenommen werden. Sie arbeiten auf einer Datenbasis, ohne deren kausale Regelmäßigkeiten zu kennen, sie »wetten« auf Kausalität, die in den Mustern der Daten verborgen sind.

Traditionell zielt Wissenschaft darauf ab, Phänomene durch Formulierung fundamentaler Theorien, Modelle oder Gesetze zu erklären und deduktiv ab-

zuleiten. Vorrangiges Ziel angewandter Forschung ist es, Anwendungsoptionen zu realisieren, wobei Erklärungen ein Mittel hierfür sein können. KI zeigt in ihren Möglichkeiten eine Nähe zu den Zielen angewandter Wissenschaften.

2008 schrieb Chris Anderson unter dem Titel »The End of Theory: The Data Deluge Makes the Scientific Method Obsolete«, dass vor allem Korrelationen ein hinreichendes Kriterium seien, während Kausalität nicht als notwendig gesehen werden müsse. Das ist insofern interessant, als dass in vielen anwendungsrelevanten KI-Anwendungen wenig über Ursache-Wirkungs-Zusammenhänge bekannt ist – man hat es mit einer »Black Box« zu tun (Heckwolf/Schmitt 2022). Für die Medizin liegt hier eine große Herausforderung, denn für Diagnose und Therapie vieler Krankheiten ist solides, konsistentes Kausalwissen über Ursache-Wirkungs-Zusammenhänge elementar, um die Wahl der Therapie kausal belegen zu können (Häußling et al. 2021). Gleichzeitig wohnt solchen Forschungsfeldern eine natürliche Komplexität inne, die traditionelle, lineare Ursache-Wirkungsprämissen kaum mehr als analytisches Kriterium zulassen. Vor diesem Hintergrund können KI- und Machine Learning-Verfahren als methodische Ansätze zur Bewältigung von Komplexitätsproblemen betrachtet werden. Sie eröffnen so einen (wie auch immer gearteten) Zugang zu überkomplexen Objektsystemen. Es liegt in der Natur komplexer nicht-linearer dynamischer Systeme, dass für sie aufgrund von Nichtlinearitäten, Instabilitäten und Bifurkationspunkten keine weitreichenden (theorie- und gesetzesbasierten) Prognosen möglich sind.

Nun ist es wesentliches Ziel von KI-Systemen, Formen der Regelmäßigkeit, also verborgene Muster oder Abhängigkeiten in komplexen (vermeintlich regelschwachen) Objektsystemen zu finden, und auch schwache Kausalitäten zu erkennen. Ziel ist dabei prognostischen Nutzen zu ziehen, ohne die zugrundeliegenden deterministischen oder probabilistischen Gesetze zu kennen oder verstanden zu haben.

Wird KI als methodischer Ansatz verstanden, um den Umgang mit komplexen Systemen zu ermöglichen, stellt sich die Frage, wie transparent nachvollziehbar und wie reproduzierbar das Verhalten von KI-Systemen selbst ist. Der Philosoph Paul Humphreys beschrieb 2007 in seinem Vortrag »Methodological Novelties in Computer Simulations« (Humphreys 2007) eine tiefgreifende Veränderung in der Bedeutung von »Techniken« im Wissenschaftsein-satz. Computermethoden bedeuten eine grundlegende (arbeits-)philosophische Neuerung in folgendem Sinne: Sie schieben den Menschen aus dem Zentrum des Forschungsprozesses der Erkenntnisgewinnung heraus, allerdings ohne, dass er vollständig verdrängt wird.

Zugleich führe der zunehmende Einsatz computerbasierter Verfahren, von Algorithmen und Simulationen naturgemäß zu einer Intransparenz. Die Forscher:in sei, nach Humphreys, kognitiv nicht mehr in der Lage, zu durchdringen, was in Simulationen geschehe und wie sich die Ergebnisse aus den Daten ableiten. Einzeloperationen des Algorithmus' ließen sich noch nachvollziehen, aber das Verhalten eines komplexen Modells in seiner Gesamtheit entzieht sich dem Verständnis des Forschenden. Eine Erklärung dessen, was innerhalb der »Black Box« von Computermodellen vor sich gehe, sei kaum mehr möglich, womit zwangsläufig Probleme der Validierung entstehen (Knight 2017). Richtiges oder falsches Verhalten dieser Modelle können allenfalls beobachtet, aber kaum bewiesen werden. Es entsteht das beunruhigende Gefühl von unkontrollierbarer Handlungsautonomie bei solchen Systemen. Dieses Problem verschärft sich bei der Verwendung tiefer neuronaler Netze. Bei ihnen kann von außen nur noch mit großem Aufwand reproduziert werden, auf welche Muster das System reagiert und welche Merkmals-Korrelationen gezogen werden. Kommt bei den Eingangsdaten ein geringes Rauschen hinzu, führt dies zu strukturellen Instabilitäten des Modells. Kleinste Störungen oder Artefakte in den Eingangsdaten können zu vollständig anderen Resultaten führen. Das hat in der Praxis weitreichende Konsequenzen: Bei Bilderkennungssystemen autonomer Fahrzeuge wird aus einem Luftballon ein Stopp-Schild oder Geschwindigkeitsschilder werden bei überlagerten Artefakten nicht als solche erkannt. Der Ruf nach sich erklärenden Systemen »Explainable AIs« (Holzinger 2018: 138), die offenlegen können wie Entscheidungen z.B. in neuronalen Netzen entstanden sind, werden lauter (Beuth 2017). Dies gilt insbesondere, wenn sich durch diese KI-Entscheidungen direkte Konsequenzen für den Menschen ergeben. Daraus ergibt sich mit Bezug auf White die Frage nach der Neuverteilung von Kontrollbeziehungen. Wer kontrolliert hier wie durch welche Verfahren und ergeben sich aus den Kontrollschleifen neue Positionierungen in den Netzwerken, die mit dem Autofahren oder medizinischen Entscheidungen vergleichbar sind.

Doch die Forderung nach Transparenz stößt an Grenzen, wenn die Erklärung des Entscheidungsprozesses ihrerseits nur noch durch einen Computer überprüft werden kann. Ein rekursives Paradoxon: Um mit den neuen komplexen Herausforderungen in der Wissenschaft umgehen zu können, werden Verfahren und Instrumente mit ebenfalls hinreichender Komplexität benötigt. Komplexität soll durch Komplexität beherrschbar gemacht werden. Die modernen Instrumente, die neue Technik der KI-Verfahren, bekommen so einen »Eigensinn«, ein eigentümliches, scheinbar magisches Verhalten (vgl. im vor-

hergehenden Kapitel das Verhalten von AlphaGo). In der Mathematik und der theoretischen Physik haben sich Machine Learning-Verfahren als Beweisverfahren bereits etabliert. So konnte das »Vier-Farben-Theorem«, für das bisher kein analytischer Beweis gefunden wurde, mit KI-Hilfe zumindest ein numerisch-algorithmischer Beweis gefunden werden. In den 1970er Jahren proklamierte der Mathematiker und Träger der Fields-Medaille, Paul Cohen, dass »[...] zu einem unbestimmten Zeitpunkt in der Zukunft Mathematiker durch Computer ersetzt werden«. Cohen war davon überzeugt, dass sich die gesamte Mathematik aufgrund ihrer formalisierten Natur automatisieren ließe, einschließlich des Führens von Beweisen.

Mit diesem Wandel wissenschaftlicher Instrumente verändert sich auch die Form der Zusammenarbeit der Forscher:in mit ihrem Analyseinstrumentarium und damit auch der wissenschaftliche Erkenntnisprozess selbst. Es zeichnen sich neuartige Herangehensweisen, unkonventionelle Lösungen außerhalb etablierter Denkmuster ab. Es entstehen neue Forschungskonzepte, Denkweisen und Denkstile (Fleck 1980). Vor dem Hintergrund der beispielhaft beschriebenen Anwendungsfälle von KI-Systemen, wird diesem neuen Techniktyp sowohl in der Gesellschaft als auch in der Wissenschaft ein Quasi-Akteursstatus zuerkannt. Dadurch zeichnen sich grundlegende Verschiebungen in Forschungspraktiken, wie auch beim Selbstverständnis von Erkenntnisprozessen, ab. In ihnen entstehen bei der wissenschaftlichen Wissensproduktion neue und möglicherweise weitreichende Interaktionsformen zwischen menschlichen und nicht-menschlichen Akteur:innen.

4. Die Rolle der KI in der Zukunftsforschung

Es ist offensichtlich, dass KI die Experimentieroptionen und Handlungsmöglichkeiten der Forscher:in erweitern. Ihr Einsatz führt schon heute zu einer veränderten Rolle der wissenschaftlichen Wissensarbeiter:innen und der Forscher:innen. Sie verändern auch das prognoseorientierte Wissenschaftsverständnis – und zwar zunehmend in Richtung einer Prognostik ohne Gegenstandswissen. Es scheint gar nicht mehr notwendig zu sein, Wissen zu erwerben. Prognosen scheinen auch ohne Gegenstandswissen (theoriefrei) möglich, Daten sind hinreichend. Big Data (Scheiermann/Zweck 2014; Bosse

et al. 2018) und die Datengetriebene Prognostik¹¹ sind auch Bestandteil der Arbeit der Zukunftsforscher:in geworden. Mit ausreichend großen und qualitativ hochwertigen Daten scheint Realität abbildbar zu werden. Mit einem dazu passenden geeigneten Daten- und Prognosemodell werden Vorhersagen möglich, wie die automatisierten Prognosen von Maschinenausfällen, Prozessabweichungen oder sogar das Verhalten von Menschen, Märkten und Gesellschaften. Dies erinnert an Isaac Asimovs Psychohistorik (Nevala-Lee 2018), einem fiktionalen, hypothetischen Konzept, nach dem auf Basis mathematischer, statistischer und soziologischer Verfahren allgemeine und dennoch präzise Voraussagen über das zukünftige Verhalten großer Gruppen von Menschen gemacht werden können. Wie ein ganzes Land mit Hilfe einer computergestützten Planwirtschaft gesteuert werden sollte, wurde bereits in den 1970er Jahren untersucht. Unter der Regierung von Salvador Allendes unternahm Chile den ehrgeizigen Versuch, mit dem Projekt Cybersyn («cybernetic synergy») die Verwaltungswirtschaft des Landes in Echtzeit durch einen zentralen Computer zu kontrollieren. Die verstaatlichten Fabriken und Unternehmen sollten sich selbst steuern, wie Zellen und Organe eines Lebewesens, indem Lagerstände, Absatzzahlen und Produktionsmengen im ganzen Land in Echtzeit erfasst werden. So sollte zentral mit Algorithmen berechnet werden, wo Knappheiten drohen und wie Ressourcen am effektivsten verteilt werden können. Nach den Prinzipien des Konnektionismus und der Kybernetik, also der Wissenschaft von der Steuerung und Regelung von Systemen, sollte sich das Wirtschaftssystem Chiles selbst im Gleichgewicht halten. Ziel dieses technologisch-politischen Projekts war eine gerechtere Gesellschaft auf Basis einer rechnergestützten Wirtschaft (Beer 1973). Das Projekt konnte erste Erfolge nachweisen, wurde jedoch 1973 im Zuge eines Militärputsches zerstört.

Heute, ein halbes Jahrhundert später, erfahren Konzepte wie Kybernetik, System Dynamics und einige der ursprünglichen Ideen hinter Projekt Cybersyn durch Künstliche Intelligenz, dem Machine Learning und den Möglichkeiten von Big Data eine Renaissance. Methoden wie »Predictive Analytics« haben aktuell Hochkonjunktur (Zweck/Braun 2021). Die prädiktive Analyse wird dabei nicht nur eingesetzt, um künftige Bedarfe zu prognostizieren und frühzeitig die zielgerichtete industrielle Produktion zu planen, sondern auch

11 Prognose zum Volumen der jährlich generierten digitalen Datenmenge weltweit in den Jahren 2018 und 2025, <https://de.statista.com/statistik/daten/studie/267974/umfrage/prognose-zum-weltweit-generierten-datenvolumen/>, abgerufen am 20.03.2022.

als Methode zur »(Voraus-)Berechnung« einer mehr oder weniger einfachen, linearen Zukunft mit geringem Zeithorizont. In »Predictive Analytics: Sind Zukunftsforscher:innen ein Auslaufmodell?« (ebd.) zeigen die Autoren, inwieweit KI bereits heute Zukunftsforschende unterstützen kann und plädieren dabei auch für eine stärkere Vernetzung der Zukunftsforschung mit anderen Disziplinen, die sich schon länger mit Big Data Analysen beschäftigen. Zugleich bezweifeln sie ausdrücklich, dass wir über Big Data und Predictive Analytics die Zukunft auch nur annähernd im Griff haben. Die zugrundeliegenden Analyseverfahren und Algorithmen aus Machine Learning und Statistik prognostizieren ein Bild der Zukunft, das auf Daten basiert, die die Vergangenheit abbilden oder im Idealfall sogar auf Echtzeitdaten der Gegenwart beruht. Vorhersagen auf Basis vergangener Ereignisse gelingen jedoch nur in stabilen und abgrenzbaren Umwelten einigermaßen gut: in Systemen, in denen keine komplexen Einflüsse oder gravierende interne wie externe Ereignisse eintreten, die das Gesamtsystem wesentlich beeinflussen können. Sie funktionieren also nur in den Fällen, in denen komplexe Wechselwirkungen gesellschaftlicher Dynamik wie auch unwahrscheinliche Ereignisse mit hoher Wirkungskraft (Wildcards) keinen Faktor darstellen. Extrapolationen und Prognosen gelingen umso schlechter, je dynamischer die betrachteten Systeme sind und je stärker exogene Ereignisse Einfluss wirken – wie beispielsweise der Ausbruch einer Pandemie. Im Frühjahr 2020, also im ersten Jahr der Corona-Pandemie, hatten viele Prognosealgorithmen (insbesondere aus dem Bereich der Predictive Analytics) erhebliche Probleme mit Strukturbrüchen in Folge des Corona-Lockdowns (Moorstedt 2020). Das bisher prognostizierte Konsumverhalten der Kund:innen in Supermärkten hatte sich grundlegend verändert und die massiv ansteigenden Verkaufszahlen bei Toilettenpapier und Teigwaren führten dazu, dass die Absatzprognosen nach einem Einbruch den Trend vollkommen überzeichneten. Extrapolationen und Prognosen sind also nur geeignet, um Aussagen über Systeme zu gewinnen, die sich wie in einem Laborexperiment unter klar definierten und oft künstlich geschaffenen experimentellen Konditionen abspielen. Letztlich geht dieses Konzept von einer Fortführung von Tendenzen aus, die sich aus den bisherigen Daten im Sinne »einer prognostizierten Zukunft« ableiten lässt.

Um die Zukunft in Situationen hoher Unsicherheit oder komplexen Umfeldern zu antizipieren, sind andere Methoden der Zukunftsforschung erforderlich, die keinen Anspruch von Prognosen im Sinne einer vorausberechneten Zukunft haben, sondern in einer Vielzahl möglicher Zukünfte denken. Es geht hier um Prospektionen, d.h. einer Auslotung möglicher

Zukunftsoptionen, um strategische Entscheidungen mit größerem Zeithorizont zu unterstützen. Dementsprechend wird hier ein Strauß von möglichen Entwicklungspfaden einer offenen Zukunft angeboten (Geschka/Hammer 1997). Dabei geht es weniger darum, einzelne Ereignisse mit maximaler Präzision vorherzusagen, wie bei den o.g. operativen Prozessen, als darum, strategische Handlungsalternativen anhand möglicher Zukunftsszenarien aufzuzeigen. Zu diesen Methoden gehört z.B. die Szenarioanalyse (Kosow/Gaßner 2008), eine der bekanntesten Methoden der Zukunftsforschung. Ziel der Szenariomethode ist es, mögliche Entwicklungen der Zukunft zu analysieren und zusammenhängend und konsistent darzustellen. Nach der Definition und Abgrenzung von Betrachtungsgegenstand und Zeithorizont werden zunächst Einflussfaktoren ermittelt, die das Thema charakterisieren. Durch Befragungen werden paarweise Wechselwirkungen bewertet, anschließend per Wechselwirkungsanalyse Schlüsselfaktoren ermittelt, zu denen dann, bezogen auf den betrachteten Zeithorizont, Projektionen durchgeführt werden. Die Projektionen werden über die Schlüsselfaktoren paarweise hinsichtlich ihrer Konsistenz bewertet. Rechnergestützt werden dann jene Projektionsbündel (Rohszenarien) berechnet, die eine ausreichend hohe Konsistenz aufweisen. Auf Basis der Rohszenarien werden dann alternative zukünftige Situationen (Szenarien) sowie Pfade, die zu diesen zukünftigen Situationen führen können, ausformuliert. Szenarien stellen evidenzbasierte, hypothetische Folgen von Ereignissen auf, um auf kausale Prozesse und Entscheidungsmomente aufmerksam zu machen. Szenarien stellen mögliche Zukunftsbilder einer offenen Zukunft dar und geben Hinweise darauf, welche Faktoren wie ausgeprägt sein müssen, damit ein bestimmtes Szenario möglich werden könnte. Neben der Darstellung, wie eine hypothetische Situation in der Zukunft zustande kommen kann, werden also ebenso mögliche Varianten und Alternativen dargestellt und aufgezeigt, welche Entscheidungs- und Handlungsoptionen es für verschiedene Akteur:innen gibt.

Durch das Einbeziehen von Wildcards, kann darüber hinaus das Eintreten auch unvorhergesehener oder unwahrscheinlicher Ereignisse mit hoher Wirkungsintensität auf das Gesamtsystem in die Szenario-Methode integriert und modelliert werden. Meistens werden in Verbindung mit Risikomanagementinitiativen Worst-Case-Szenarien wie Kriegausbrüche, Naturkatastrophen oder Pandemien herangezogen, um vorbereitende Strategien für solche Herausforderungen zu entwickeln. Unternehmen haben, wenn eines dieser Ereignisse von geringer Wahrscheinlichkeit aber großer Reichweite eintritt, bereits eine passende Strategie in der Schublade. Sie können sich so in Kri-

sensituation rasch von der taumelnden Masse im Markt abheben. Eine umfassende Vorbereitung wichtiger politischer Entscheidungen erfordert nicht nur Analysen auf Basis bestehenden Wissens, sondern auch eine Auseinandersetzung mit möglichen Zukunftsperspektiven.

Ein Gedankenspiel ist, dass bei der Prädiktion der Zukunft »nur« spezielle, isolierte Teilprobleme berechnet werden, die durch neuronale Netze gezielt trainiert werden können. So liefert die KI Watson von IBM Antworten auf Basis von Wahrscheinlichkeiten – Prädiktion per Wahrscheinlichkeit. Auch Klassifikationsalgorithmen (wie sie beispielsweise in Spam-Filtern verwendet werden) liefern Ergebnisse nach dem Muster: Antwort A ist mit 87 % Wahrscheinlichkeit wahr, Antwort B mit 9 %, Antwort C mit 2 % usw. Für viele Anwendungen ist das ausreichend. Hinsichtlich der kausalen Argumentation mutet dies unvollständig an und erscheint im Kontext evidenzbasierter Forschung unbefriedigend, denn die Forderung nach Nachvollziehbarkeit, Erklärbarkeit und Transparenz sind ihrerseits Gütekriterien seriöser Zukunftsforschung (Gerhold et al. 2015). Es existieren aber Problemklassen, deren Lösungen zwar »schwierig« (im mathematischen Sinne z.B. der NP-Vollständigkeit, Exponentialität usw.) zu finden, aber leicht (in konstanter Zeit) zu überprüfen sind. Zu diesen Problemen zählen beispielsweise bestimmte Optimierungsprobleme der Statik, Logistik, Proteinfaltung oder Existenzbeweise. Ähnlich verhält es sich mit bestimmten Methoden der Zukunftsforschung. Beispielsweise erklärt die Szenario-Methode selten (das ist oft auch gar nicht notwendig) die Ursachen und Gründe bzw. die historischen Entwicklungspfade, die zu einem zu betrachteten Szenario geführt haben. Es wird vielmehr ein Szenario als konsistentes Gesamtereignis innerhalb eines zukünftigen Möglichkeitsraumes formuliert. Der Wahrheitswert eines solchen Szenarios steht nicht zur Diskussion, solange die Konsistenz ausreichend und das Bild insgesamt schlüssig ist. KI als Akteurin der Prädiktion wäre hier keine Rechenschaft pflichtig, sondern präsentiert der Entscheider:in eigenständig ermittelte, konsistente Zukunftsszenarien, die betrachtet und berücksichtigt werden sollten. Wie dies funktionieren kann, wird in der Folge an einem Beispiel gezeigt.

Ansätze wie das Projekt Cassandra des Bundesverteidigungsministeriums und des Weltethos-Instituts zur Krisenfrüherkennung und Gewaltprävention zeigen¹², wie immer häufiger nicht nur algorithmenbasierte Modelle für

12 <https://weltethos-institut.org/forschung/cassandra-projekt/>, abgerufen am 20.03.2022.

Prognosen eingesetzt werden, sondern auch bereits existierende Weltbilder zielgerichtet für neue Erkenntnisse ausgewertet werden können. In diesem Fall wird bestehende Literatur ausgewertet, um auch das Unerwartete, wie Pandemien, Klimakrisen, Kriminalistik oder geopolitische Spannungsfelder zu durchdenken. Im Projekt Cassandra dient Literatur als besondere Form des Wissens. Die Literatur zeigt die verdeckten Seiten der Wirklichkeit und dient hier als Instrument für das »Worldbuilding« (Gavins 2016), für die Prospektion und für das, was eintreten könnte.

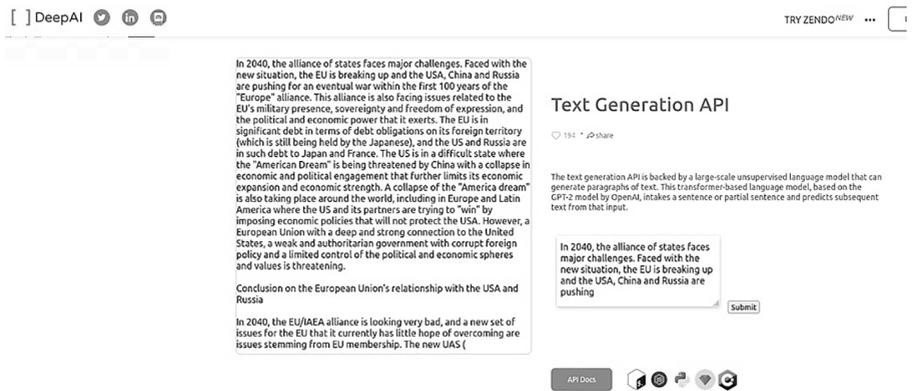
Wie Künstliche Intelligenzen bzw. Machine Learning-Verfahren im Detail als Instrumente in der Zukunftsforschung eingesetzt werden können, überstiege den Umfang dieses Textes und bedürfte einer umfassenderen Untersuchung, die an dieser Stelle nicht möglich ist. Dennoch kann an einem Beispiel gezeigt werden, wie Sprachmodelle als Instrument für eine fiktive Narration dienen können, indem sie auf bestehendem Weltwissen (wie z.B. Literatur oder wissenschaftlichen Texten) trainiert wird. Künstliche Intelligenzen verwenden in den meisten Fällen statistische Verfahren, die auf Trainingsdatensätze angewendet werden, um auf die in Daten enthaltenen Strukturen und Muster zu reagieren (auch wenn diese augenscheinlich versteckt sind). KI-Systeme werden trainiert – und zwar ausschließlich mit dem »Weltbild« der Trainingsdaten. Das ist zwar ein Bias, aber gleichzeitig kann diese Prägung auch Teil der Lernstrategie sein. Eine Art »narrative Prospektion« auf Basis von Sprachmodellen (Mayer 2021; Knight 2021) liegt nahe. So stellt das Trainieren von KI-Systemen in gewisser Weise die Schaffung eines Abbildes der Vergangenheit dar – die Gegenwart oder die Zukunft werden auf die Daten der Vergangenheit normiert. Ein KI-basiertes Instrument für Worldbuilding stellen schon heute umfangreiche Sprachmodelle (Heaven 2017) wie GPT-2, GPT-3 oder Aleph Alpha dar.

Auch derartige Sprachmodelle werden auf Grund ihres jeweiligen Trainings in ihrer Eigenschaft die Formalität der Sprache abzubilden nicht nur zu Instrumenten der Narration im Sinne eines erzählenden deskriptiven Charakters, sondern zu einem Werkzeug der Vermittlung dahinterstehender offen oder nicht offen gelegter impliziter Narrative. Die »narrative Prospektion« (Eggert/Zweck 2021: 424f.) hat in diesem Sinne also einen doppelzüngigen Charakter.

Wie Sprachmodelle und Textgeneratoren für die narrative Prospektion bzw. für das Szenario-Building eingesetzt werden können und zu welchen Leistungen diese Modelle heutzutage in der Lage sind, auch um neue, unkonventionelle Denkwege zu beschreiten, zeigt das folgende Beispiel: Es wird als

Grundannahme definiert, dass in einer alternativen, hypothetischen Zukunft im Jahr 2040 die EU gescheitert ist. Das Staatenbündnis bricht auseinander. Welches Szenario zeichnet ein KI-basiertes Sprachmodell für eine solche Zukunft? Der Textgenerator Deep AI von OpenAI, wird mit dieser hypothetischen Situation in der Zukunft »konfrontiert«. Mit welcher Geschichte und mit welchen Erklärungen »konfrontiert« uns die KI in der Folge? Abbildung 2 zeigt die Situation, wie sie auch die Leser:in ausprobieren kann.¹³

Abbildung 2: Texterstellung fiktionaler Texte mit Hilfe eines KI-gestützten Textgenerators



Der folgende Eingangstext wurde dem Textgenerator als Starttext gegeben, die KI vervollständigt die Geschichte dann selbstständig.

»In 2040, the alliance of states faces major challenges. Faced with the new situation, the EU is breaking up and the USA, China and Russia are pushing[...].«

Von der KI produzierte Geschichte oder in diesem Kontext, das produzierte Szenario:

»In 2040, the alliance of states faces major challenges. Faced with new situations, the EU is breaking up and the USA, China and Russia are pushing for

13 DeepAI, Textgenerator, <https://deepai.org/machine-learning-model/text-generator>, abgerufen am 20.03.2022.

an eventual war within the first 100 years of the ›Europe‹ alliance. This alliance is also facing issues related to the EU's military presence, sovereignty and freedom of expression, and the political and economic power that it exerts. The EU is in significant debt in terms of debt obligations on its foreign territory (which is still being held by the Japanese), and the US and Russia are in such debt to Japan and France. The US is in a difficult state where the ›American Dream‹ is being threatened by China with a collapse in economic and political engagement that further limits its economic expansion and economic strength. A collapse of the ›America dream‹ is also taking place around the world, including in Europe and Latin America where the US and its partners are trying to ›win‹ by imposing economic policies that will not protect the USA. However, a European Union with a deep and strong connection to the United States, a weak and authoritarian government with corrupt foreign policy and a limited control of the political and economic spheres and values is threatening.

Conclusion on the European Union's relationship with the USA and Russia: In 2040, the EU/IAEA alliance is looking very bad, and a new set of issues for the EU that it currently has little hope of overcoming are issues stemming from EU membership. The new UAS [...].«

Der Textgenerator beendet diesen spannenden Ausblick auf das Jahr 2040 mit einem wahren »Cliffhanger«, in Form eines nicht weiter erwähnten Staatenbündnisses namens »UAS«. Doch so »kreativ« oder »intelligent« KI-Modelle und insbesondere die modernen Sprachmodelle wie GPT-3 und Aleph Alpha auch wirken, am Ende sind sie doch nur »stochastische Papageien«, wie es amerikanische Forscherinnen genannt haben (Bender et al. 2021): »Sie verstehen nicht, sondern plappern nach, besser oder schlechter, je nachdem, wie man sie füttert«. Dennoch, neuronale Netze finden Lösungen zu Problemen, die Expert:innen überraschen und erstaunen. Sie gehen Wege, verwenden Strategien, und liefern Lösungen, formulieren approximierete Wahrheitswerte in Wahrscheinlichkeiten und entwickeln Zukunftsentwürfe, die verblüffen. Wie bereits festgestellt, sind wahrheitsgetreue (im Sinne der Aussagenlogik) Aussagen über die Zukunft inhärent unscharf. Diese naturgegebene Unschärfe bei der Prognostik muss daher auch Künstlichen Intelligenzen bei der Berechnung von Zukünften zugestanden werden. Auch aus dieser Perspektive lässt sich die Einschätzung rechtfertigen, dass KI eine hilfreiche Unterstützung für die Zukunftsforschung sein kann. Der daraus entstehende Wechselbezug zwischen KI und Zukunftsforschung und wie die gegenseitige Beeinflussung wirkt (Rückkopplungseffekte und Technikfolgenabschätzung), wird in der

Zukunftsforschung selbst intensiv diskutiert (Steinmüller et al. 2022) und kann an anderer Stelle weiter vertieft werden. Es kann beobachtet werden, dass, wenn KI bzw. Prognostik verstärkt zum Einsatz kommen, dies auch das Umfeld verändert – wie das Beispiel Predictive Policing oder flächendeckende Videoüberwachung zeigen (Gierke 2013) – und zu wechselseitigen, fast schon rekursiven, selbstreferenziellen Überlegungen führt. KI-Prognostik verändert das Umfeld und damit auch die Zukunftsforschung.

5. Neuere Künstliche Intelligenz – weder Akteur:in noch reines Instrument?

Es ist deutlich geworden, dass KI-Systeme in Wissenschaft und Forschung weit mehr als nur Werkzeugcharakter haben. Sie ermöglichen fundamental neue Denkansätze und eröffnen neue Zugänge des Denkens und des Forschens. Es liegt daher nahe zu fragen, wie die Zukunft der KI aussehen könnte. Der vorliegende Text jedoch befasst sich mit der Perspektive des Einsatzes von KI in der Zukunftsforschung. KI führt zu neuen, auch disruptiven Nutzungskontexten, und dringt in die Zweckdimension forschenden Handelns ein. Auf der anderen Seite wird die o.g. besondere Leistungsfähigkeit von KI-Systemen neue, bislang kaum zugängliche Forschungsfelder erschließbar machen. KI wird Teil des Forschens, sie wird Impulsgeber:in, Lösungsgehilfin und Sparringspartner:in. Aus dieser Perspektive erscheint es berechtigt und zweckmäßig, KI sowohl in der Soziologie als auch in der Zukunftsforschung nicht nur als Gegenstand der Betrachtung zu sehen, sondern als agierende Akteur:innen. Ob und in welcher Weise Computer – wenn überhaupt – »denken« können, rückt zunehmend in den Hintergrund gegenüber der Frage in welchem Sinne Computer (z.B. als humanoide Roboter) »handeln« können. Diese Aufmerksamkeitsverschiebung verläuft parallel zur technischen Entwicklung, die intelligente Systeme nicht nur als reine Computersysteme mit hoher Informationsverarbeitungskapazität versteht, sondern als Quasi-Akteur:in, die menschliches Handeln nicht nur nachahmen oder übertreffen, sondern zunehmend »autonom« Handeln. Es stellt sich somit die Frage, in welchem Sinne diese neue Form der Akteur:innen überhaupt handelt und in welchem Umfang »autonome« Systeme für ihr Handeln verantwortlich gemacht werden können. Eine vertiefte Reflektion hierüber ist an dieser Stelle nicht möglich, aber es wird deutlich, dass mit der digitalen, autonom handelnden Akteur:in die Akteursrolle in das Instrument selbst verlegt wird, das

begrifflich genau in diesem Moment aufhört, nur noch Instrument zu sein. Es ist nachvollziehbar, dass eine Handlungsurheberschaft, also die Empfindung alleine und autonom verantwortlich für die Veränderung des eigenen Wesens oder der umgebenen Umwelt zu sein – der bewusste Wille – solchen autonomen Systemen nicht zugestanden werden kann.

Bruno Latour wies auf eine zunehmend unklare Rollenverteilung zwischen Mensch und Maschine hin (Latour 2001), nach der die Welt voller Hybride, Quasi-Subjekte ist, die nicht nur konstruiert sind, sondern Welt durch Vermittlung und Delegation selbst konstruieren und in Folge dessen weder beherrschbar noch zwangsläufig nachvollziehbar sind. Latour stellt die Außengrenze des Menschen in Frage, sodass beim »Duell« Mensch gegen Maschine die Rollen nicht mehr klar verteilt sind.

Die Relationale Soziologie um Harrison White konzipiert Netzwerke als zwischenmenschliche Konstruktionen. Als Ausgangspunkt seiner Betrachtung sieht White sogenannte Transaktionen, soziale Prozesse in denen sich Netzwerke und andere soziale Formationen ergeben (White 1995: 1035f.). Der Begriff lenkt damit den Blick weg von den Akteur:innen und hin zu den Prozessen, die zwischen den Akteur:innen bzw. »Identitäten« ablaufen. Bei White sind Identitäten die Einheiten, die in den Netzwerken um gegenseitige Kontrolle ringen (White 1992).

Andererseits verweist der Identitätsbegriff auch auf die konstruierende Beobachtung dieser Einheiten im Netzwerk. Der Begriff Identität darf den Blick nicht auf menschliche Individuen verengen, da White unter dem Begriff der Identität auch kollektive Einheiten wie Staaten, Unternehmen, soziale Bewegungen versteht. Da (soziale) Netzwerke im Sinne von White sowohl zwischen Personen als auch zwischen Organisationen und anderen kommunikativ konstruierten Identitäten zu finden sind, ist also der Identitätsbegriff für eine theoretische Grundlegung sozialer Netzwerke besser geeignet als der der Person – dies auch, wenn das Personenkonzept genauer zwischen den internen Prozessen von psychischen Systemen und deren Beobachtung als Person in der Kommunikation unterscheidet. Nach White sind soziale Strukturen notwendig, denn individuelle Identitäten werden erst durch soziale Strukturen oder Netzwerke konstituiert und die sozialen Strukturen werden die Umgebung für Identität, die deren Entstehen und Entwicklung formieren und kontrollieren (vgl. White 1992; White et al. 1976; White et al. 2007).

6. Ausblick

Zukunftsforschung reduziert sich in der Praxis von Foresight, Technologiefrüherkennung oder Technikfolgenabschätzung (Zweck 2002: 55f.) nicht auf einen rein rechnerbasierten, technokratisch-organisierten Beratungsprozess (Zweck/Braun 2021). In der Praxis handelt es sich um einen sozialen Prozess: um einen in wohldefinierten, mehreren Schritten verlaufenden Einsichtsprozess der Beteiligten. Die Schritte dienen aber zugleich dazu, eine iterative Interaktion zwischen den Akteur:innen, seien es Experten:innen, Betroffene, oder anderweitig Beteiligte sicher zu stellen. Denn während des Zukunftsforschungsprozesses ist es neben dem Erkenntnisgewinn essenziell, eine begleitende und über das Projekt hinausgreifende Akzeptanz und Verbreitung gewonnener Einsichten sicher zu stellen. Vor dem Hintergrund, der im vorliegenden Beitrag reflektierten Fragen, bleibt offen, ob, wann und in welchem Grad sich KI über die diskutierten Vorteile bezüglich Datenanalyse etc. hinaus zu einer im Prozess eigenständigen und sozialen Akteur:in entwickeln könnte, einer Akteur:in die im Zukunftsforschungsprozess zukünftig Gehör findet. Gegenwärtig erreicht KI diese Qualität im sozialen Prozess in keiner Weise. KI tritt allenfalls als Unterstützung in zweiter Reihe an, also als Hintergrundakteur:in zur neuen impulsgebenden Verbesserung des Wissensstandes einzelner menschlicher Beteiligter im Zukunftsforschungsprozess, die die entsprechenden Ergebnisse einbringen. Wenn auch gegenwärtig noch aus zweiter Reihe, wird dieser datengetriebene Beitrag der KI zum Zukunftsforschungsprozess zunehmend wichtiger werden. Wie der vorliegende Beitrag zeigt, zeichnet sich dies sowohl was kreative Impulse anbelangt, in dem unerwartete Zusammenhänge oder übersehene Wildcards erfasst werden, wie auch bei der Analyse großer Datenbestände komplexer Fragestellungen ab. In einer KI-getriebenen Zukunftsforschung wird die Vernetzung der Akteur:innen aus verschiedenen Disziplinen zu einem noch zentraleren Erfolgsfaktor. Die Erschließung der Vielseitigkeit disziplinspezifischer Erfahrungen und Datensammlungen durch KI-gestützte Vernetzung und Auswertung dürften zukünftig wichtige Impulse für eine bessere Erfassung und Charakterisierung des Umfeldes des gewonnenen Zukunftsbildes bieten und neue und verbesserte Möglichkeiten der Qualitätssicherung eröffnen.

Technische Systeme, die mit sog. künstlicher »Intelligenz« ausgestattet sind, können grundsätzlich nicht als Akteur:in in Begründungsdiskursen auftreten, da ihnen die Merkmale von Handlungsurheberschaft und Zurechenbarkeit fehlen und allenfalls vom Menschen zugerechnet werden. Die

wissenschaftliche Forschung ist und wird auch im Zeitalter von KI, Machine Learning und Big Data weiterhin auf den Menschen als Wissenschaftler:in und Forscher:in angewiesen sein. In der bestehenden Phase schwacher KI bleibt der Mensch als handelnde Akteur:in im Zentrum von Wissenschaft und Forschung. Ein etwaiger Ersatz des Menschen durch »starke KI« ist nicht zu erwarten, da diese zumindest auf absehbare Zeit nicht erreichbar erscheint. Immerhin ist es mit den abgestuften Konzepten von Handlungsträgerschaften – wie bei White oder Latour – nachvollziehbar, an welchem Punkt der soziotechnische Entwicklungsprozess steht und welche Implikationen damit für die sozialen Formationen einhergehen, etwa wenn Eliza tatsächlich für Therapiesitzungen geeignet wäre oder Programme den Turing-Test bestehen. Forscher:innen müssen sich auch im Zeitalter von KI als Urheber:in des Forschungshandelns begreifen. Die Frage, ob es gelingt, dieser Einsicht treu zu bleiben, wenn Wissenschaftler:innen vor Bewunderung von durch schwache KI gefundenen Erkenntnissen – wie im Falle von Eliza – intelligente und menschliche Eigenschaften und Fähigkeiten hineinprojizieren, wird die Zukunft zeigen.

Seit der Endredaktion dieses Beitrages hat sich die Entwicklung in der KI, insbesondere im Bereich der großen Sprachmodelle rasant beschleunigt. Nur wenige Wochen nach Fertigstellung dieses Beitrages präsentierte OpenAI mit ChatGPT sein heute bekanntestes Sprachmodell zur Verarbeitung natürlich sprachlicher Computeranweisungen, das inzwischen von mehreren hundert Millionen Menschen weltweit benutzt wird. In den darauf folgenden Monaten entstanden zahlreiche weitere, leistungsfähige Sprachmodelle von Google, Amazon, Meta, Nvidia, Tencent, Aleph Alpha und vielen anderen großen Akteur:innen/Plattformen. Basierend auf einigen dieser KI-Modelle wurden mehrere frei verfügbare Open Source Modelle entwickelt, die in zahlreichen Anwendungsbereichen (wie z.B. Question Answering, Information Retrieval oder Logical Reasoning) auch und vor allem in verschiedenen Forschungsbereichen zur Unterstützung der Forschenden eingesetzt werden. Mit Stand 13.02.2024 sind ca. 505.000 KI-Modelle verfügbar, d.h. täglich entstehen weltweit ca. 1.000 neue Modelle – eine Entwicklungsgeschwindigkeit, mit der Publikationen in gedruckter Form kaum noch mehr Schritt halten können.

7. Literatur

- Baidu (2020): »How Baidu is bringing AI to the fight against coronavirus«, in: MIT Technology Review vom 11. März 2020, <https://www.technologyreview.com/2020/03/11/905366/how-baidu-is-bringing-ai-to-the-fight-against-coronavirus/>.
- Beer, Stafford (1973): *Fanfare for Effective Freedom: Cybernetic Praxis in Government*, Wales: Cwavel Isaf Institute.
- Bender, Emily M./Gebru, Timnit/McMillan-Major, Angelina/Shmitchell, Shmargaret (2021): »On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?«, in: *Conference on Fairness, Accountability, and Transparency (FACt '21)*, New York: Association for Computing Machinery, S. 610–623.
- Beuth, Patrick (2017): »Die Automaten brauchen Aufsicht«, in: *Die Zeit* vom 25.10.2017, S. 3.
- Bosse, Christian K./Hoffmann, Judith/van Eist, Ludger (2018): »Potenzialeinschätzung von Big Data Mining als methodischer Zugang für Foresight«, in: *Zeitschrift für Zukunftsforschung* 2018(1), S. 2195–3155.
- Bringsjorden, Selmer/Bello, Paul/Ferrucci, David (2001): »Creativity, the Turing Test, and the (Better) Lovelace Test«, in: *Minds and Machines* 11, S. 3–27.
- Callaway, Ewen (2020): »It will change everything: DeepMind's AI makes gigantic leap in solving protein structures«, in: *nature* 588, S. 203–204.
- Clarke, Arthur C. (1962): *Hazards of Prophecy: The Failure of Imagination*, London: Gollanz.
- Eggert, Michael/Zweck, Axel (2021): »Die Analyse naturwissenschaftlich-technischer Narrative als Element zukunftsbezogener Beratungs- und Entscheidungsprozesse«, in: Jürgen Gausemeier (Hg.)/Wilhelm Bauern (Hg.)/Roman Dumitrescu (Hg.), *Vorausschau und Technologieplanung: 16. Symposium für Vorausschau und Technologieplanung. 2. und 3. Dezember 2021*, Berlin/Paderborn: Universität Paderborn, S. 424–448.
- Fleck, Ludwig (1980): *Entstehung und Entwicklung einer wissenschaftlichen Tatsache*, Frankfurt a.M.: Suhrkamp.
- Gardner, Howard (1983): *Frames of Mind. The theory of multiple intelligences*, New York: Basic Books.
- Gavins, Joanna/Lahey, Ernestine (2016): *World Building: Discourse in the Mind*, London: Bloomsbury.

- Gerhold, Lars/Holtmannspötter, Dirk/Neuhaus, Christian/Schüll, Elmar/Schulz-Montag, Beate/Steinmüller, Karlheinz/Zweck, Axel (2015): Standards und Gütekriterien der Zukunftsforschung: Ein Handbuch für Wissenschaft und Praxis, Salzburg: Springer.
- Geschka, Horst/Hammer, Richard (1997): »Die Szenario-Technik in der strategischen Unternehmensplanung«, in: Dietger Hahn/Bernard Taylor (Hg.), Strategische Unternehmensplanung/Strategische Unternehmensführung, Heidelberg: Physica, S. 464–489.
- Giansiracusa, Noah (2021): How Algorithms Create and Prevent Fake News. Exploring the Impacts of Social Media, Deepfakes, GPT-3, and More, Acton/MA/USA: Apress.
- Gierke, Sebastian (2013): »Wie intelligent sind Videoüberwachungssysteme?«, in: Süddeutsche Zeitung vom 31.06.2013, S. 1–2.
- Goleman, Daniel (1995): Emotional Intelligence. Why It Can Matter More Than IQ, New York: Bantam.
- Goodfellow, Ian/Pouget-Abadie, Jean/Mirza, Mehdi/Xu, Bing/Warde-Farley, David/Ozair, Sherjil/Courville, Aaron/Bengio, Yoshua (2014): »Generative Adversarial Nets«, in: NIPS'14, S. 2672–2680.
- Groth, Olaf J./Nitzberg, Mark/Zehr, Dan (2018): Vergleich nationaler Strategien zur Förderung von Künstlicher Intelligenz, Berlin: Konrad-Adenauer-Stiftung.
- Häußling, Roger/Franke, Tim/Härpfer, Claudius/Roth, Philip/Schmitt, Marco/Strüver, Niklas/Zantis, Sascha (2021): Mendelian und das Erklärungspotential der Theorie von Identität und Kontrolle: Ein techniksoziologischer Blick auf Recommender-Systeme, Aachen: RWTH-Aachen IFS.
- Heaven, Will D. (2021): 2021: »Das Jahr der KI-Monstermodelle«, in: heise online vom 31.03.2021, <https://www.heise.de/hintergrund/2021-Das-Jahr-der-KI%20Monstermodelle6305269.html>.
- Heckwolf, Christoph/Schmitt, Marco (2022): KI zwischen Blackbox und Transparenz– Das Koppeln und Entkoppeln von Kontrollprojekten, Aachen: RWTH-Aachen IFS.
- Holzinger, Andreas (2018): »Explainable AI (ex-AI)«, in: Informatik-Spektrum 41, S. 138–143.
- Humphreys, Paul (2009): »The Philosophical Novelty of Computer Simulation Methods«, in: Synthese 169, S. 615–626.
- Kasparov, Garry (2017): »AlphaZero AI beats champion chess program after teaching itself in four hours«, in: Garry Kasparov vom 12.08.2017, <https://>

- www.kasparov.com/blogpost/alphazero-ai-beats-champion-chess-program-after-teaching-itself-in-four-hours/.
- Kelnar, David/Kostadinov, Asen (2019): *The State of AI 2019: Divergence*, London: MMC Ventures Research, S. 99.
- Knight, Will (2017): »The U.S. Military Wants Its Autonomous Machines to Explain Themselves«, in: MIT Technology Review vom 4. März 2017, <https://www.technologyreview.com/2017/03/14/243295/the-us-military-wants-its-autonomous-machines-to-explain-themselves/>.
- Knight, Will (2021): »This AI Can Generate Convincing Text – and Anyone Can Use It«, in: Wired vom 29.03.2021, <https://www.wired.com/story/ai-generate-convincing-text-anyone-use-it/>.
- Kosow, Hannah/Gaßner, Robert (2008): *Methoden der Zukunfts- und Szenarioanalyse Überblick, Bewertung und Auswahlkriterien. Werkstattbericht, Bericht Nr. 103*, Berlin: Institut für Zukunftsstudien und Technologiebewertung.
- Kühl, Eike (2014): »Ein Trickser namens Eugene Goostman«, in: Die Zeit vom 10.06.2014, <https://www.zeit.de/digital/internet/2014-06/turing-test-eugene-goostman-kritik>.
- Latour, Bruno (2001): *Das Parlament der Dinge: Für eine Ökologie der Dinge*, Frankfurt a.M.: Suhrkamp.
- Mayer, Boris (2021): »Textgenerator GPT-3 auf Deutsch getestet.«, in: golem.de IT-New für Profis vom 03.11.2021, <https://www.golem.de/news/textgenerator-gpt-3-auf-deutsch-getestet-genau-wahrscheinlich-sie-sind-wie-die-ameisen-2111-160468.html>.
- Mayer, John D./Salovey, Peter (1990): »Emotional intelligence«, in: *Imagination, Cognition, and Personality* 9, S. 185–211.
- Mayer, John D/DiPaolo, Maria/Salovey, Peter (1990): »Perceiving affective content in ambiguous visual stimuli: A component of emotional intelligence«, in: *Journal of Personality Assessment* 54, S. 772–781.
- Menick, John (2016): »Move 37: Artificial Intelligence, Randomness, and Creativity«, in: *Mousse Magazine* 55/53, <https://johnmenick.com/writing/move-37-alpha-go-deep-mind.html>
- Metzinger, Thomas (2001): »Postbiotisches Bewusstsein: Wie man ein künstliches Subjekt baut – und warum wir es nicht tun sollten«, in: *Paderborner Podium/20 Jahre Heinz Nixdorf Museums Forum*, S. 87–113.
- Moorstedt, Michael (2020): »Vorhersage-Algorithmen: Wie Corona künstliche Intelligenzen verwirrt«, in: *Süddeutsche Zeitung* vom 19.05.2020, <https://>

- www.sueddeutsche.de/digital/corona-kuenstliche-intelligenz-vorhersage-algorithmus-1.4910460.
- Nevala-Lee, Alec: »What Isaac Asimov Taught Us About Predicting the Future«, in: *The New York Times* vom 31.10.2018, <https://www.nytimes.com/2018/10/31/books/review/isaac-asimov-psychohistory.html>.
- Scheiermann, A./Zweck, Axel (2014): »Big Data für die Sozialforschung. Innovations- und Technikanalyse«. Kurzstudie, VDI-Technologiezentrum (Hg.), Schriftenreihe Zukünftige Technologien Consulting Nr. 98, Düsseldorf.
- Schmitt, Marco/Fuhse, Jan (2015): Zur Aktualität von Harrison White. Einführung in sein Werk, Wiesbaden: Springer Fachmedien.
- Searle, John R. (1980): »Minds, Brains, and Programs«, in: *Behavioral and Brain Sciences* 3(3), S. 417–424.
- Service, Robert (2021): »2021 Breakthrough of the year – Protein structures for all«, in: *Science* vom 16.12.2021, <https://www.science.org/content/article/breakthrough-2021>.
- Steinmüller, Karlheinz/Burchardt, Aljoscha/Gondlach, Kai/von der Gracht, Heiko/Kisgen, Stefanie/Ellermann, Kai/Martini, Melanie/John, Marcus (2022): »Kann Künstliche Intelligenz Zukunftsforschung? – Ein spekulativer Impuls«, in: *Zeitschrift für Zukunftsforschung* 2022(1), S. 2195–3155.
- Stuß, H.-M.; Beauducel, A. (2011): »Intelligenztests und ihre Bezüge zu Intelligenztheorien«, in: Lutz F. Hornke/Manfred Amelang/Martin Kersting (Hg.), *Leistungs-, Intelligenz- und Verhaltensdiagnostik*, Göttingen/Bern/Toronto/Seattle: Hogrefe, S. 97–234.
- Tegmark, Max (2017): *Leben 3.0: Mensch sein im Zeitalter Künstlicher Intelligenz*, Berlin: Ullstein.
- Turing, Alan M. (1950): »I. Computing Machinery and Intelligence«, in: *Mind* LIX(236), S. 433–460.
- von Ahn, Luis/Blum, Manuel/Hopper, Nicholas J./Langford, John (2003): »CAPTCHA: Using Hard AI Problems For Security«, in: Eli Biham (Hg.), *EUROCRYPT 2003: Advances in Cryptology*, Berlin/Heidelberg: Springer, S. 294–311.
- Wattenhofer, Roger (2018): »AlphaZero will doch nur spielen!«, in: *ETH Zürich* vom 23.02.2018, <https://ethz.ch/de/news-und-veranstaltungen/ethnews/news/2018/02/wattenhofer-alphazero.html>.
- Weber, Max (1919): *Wissenschaft als Beruf*, Berlin: Duncker & Humblot.
- Wechsler, David (1964): *Die Messung der Intelligenz Erwachsener*, Stuttgart: Hueber.

- Weizenbaum, Joseph (1976): *Computer Power and Human Reason: From Judgment to Calculation*, New York: W.H. Freeman and Company, S. 7.
- White, Harrison C. (1992): *Identity and Control: A Structural Theory of Social Action*, Princeton: Princeton University Press.
- White, Harrison C. (1995): »Network Switchings and Bayesian Forks: Reconstructing the Social and Behavioral Sciences«, in: *Social Research* 62, S. 1035–1063.
- White, Harrison C./Boorman, Scott/Breiger, Ronald (1976): »Social Structure from Multiple Networks. I. Blockmodels of Roles and Positions«, in: *American Journal of Sociology* 81, S. 730–780.
- White, Harrison C./Fuhse, Jan/Thiemann, Matthias/Buchholz, Larissa (2007): »Networks and Meaning: Styles and Switching«, in: *Soziale Systeme* 13(1-2), S. 543–555.
- White, Harrison. C. (2008): *Identity and Control. How social formations emerge*. Second edition, Princeton: Princeton University Press.
- Zweck, Axel (2002): »Three perspectives for one future in economy and society«, in: *Futures Research Quarterly* 18(1), S. 55–66.
- Zweck, Axel (2012): »Gedanken zur Zukunft der Zukunftsforschung«, in: Reinhold Popp (Hg.), *Zukunft und Wissenschaft: Wege und Irrwege der Zukunftsforschung*, Salzburg: Springer, S. 59–80.
- Zweck, Axel/Braun, Matthias (2021): *Predictive Analytics: Sind Zukunftsforscher*innen ein Auslaufmodell?*, Düsseldorf: VDI Technologiezentrum GmbH.

8. Filme

PLUG & PRAY (2010) (D, R: Jens Schanze)

Künstliche Intelligenz und relationale Techniksoziologie

Ein Schluss

Roger Häußling, Marco Schmitt

1. Einleitung

Die in diesem Sammelband verbundenen Beiträge nähern sich dem Phänomen aus einer gemeinsamen theoretischen Perspektive (relationale Techniksoziologie mit besonderer Berücksichtigung von Konzepten aus Harrison Whites Theorie von Identität und Kontrolle (White 1992, 2008)) und vor dem Hintergrund unterschiedlich gelagerter Expert:inneninterviews aus dem Feld. Dabei wird jeweils auf unterschiedliche Problemlagen scharf gestellt, wie Intransparenz oder Rollenwandel oder spezifische Anwendungsfelder wie Gesundheit oder Bilderkennung. Dadurch lässt sich ein soziologisch scharf gestelltes Bild der sozialen Bedeutung der Entwicklungen im Feld der KI der letzten Jahre gewinnen und gleichzeitig auch die techniksoziologische Perspektive einer an White orientierten relationalen Soziologie vorstellen und weiterentwickeln. Wir müssen an dieser Stelle jedoch mit einem Hinweis beginnen, da die neueren bzw. neuesten Entwicklungen, um auf Deep Learning basierenden sehr großen Sprachmodellen, in unserer Diskussion innerhalb dieses Bandes noch keine große Rolle spielen. Dies liegt daran, dass die Beiträge vor diesen, durch die von Open AI mit der Veröffentlichung von ChatGPT angestoßenen, Entwicklungen konzeptualisiert, empirisch unterfüttert und verfasst worden sind. Wir teilen die Einschätzung, dass diese Entwicklungen maßgeblich sind und Reichweite und Erwartungen an die KI nochmals massiv verändert haben; denken aber, dass die theoretischen Konzepte und soziologisch identifizierten Problemlagen aus diesem Band weiterhin Bestand haben, zumal sie technisch auf den selben Machine und Deep Learning-Verfahren basieren, die hier nur auf das spezifische Problem der Sprachnutzung angewendet werden

und damit auf der Basis massiver Datensätze eine deutlich allgemeinere Problemlösungsreichweite vermitteln. Diese Reichweiten-Phänomene sind sicherlich ein soziologisch relevantes Phänomen, doch wir gehen davon aus, dass eine relationale Techniksoziologie sich auch weiterhin gut zur Analyse des Phänomenbereichs eignet. Daher soll hier nochmal resümierend ein Fazit gezogen werden, wie nach den erfolgten Diskussionen die Grundzüge einer solchen Theorie zu skizzieren wären. Zu diesem Zweck sollen nochmal die wichtigsten Lernergebnisse aus den Beiträgen zusammengefasst werden, in ihrer Relevanz für eine an White orientierte Techniksoziologie geprüft und dann solche Grundzüge kurz vorgestellt werden.

2. Roter Faden und Lernergebnisse aus den Beiträgen

Der zentrale rote Faden, der alle Beiträge des Bandes vereint, ist die Frage nach den gesellschaftlichen Implikationen der technischen Entwicklungen im Bereich der KI, die ständig neue Möglichkeiten und auch Vorstellungswelten aufspannt und deren Beschleunigung hier sicher zu immer stärkeren Veränderungen führt. Diese gesellschaftlichen Implikationen richten sich auf unterschiedliche Bereiche, wie etwa die Positionen von bestimmten Akteur:innen in einem soziotechnischen Arrangement oder der Kontrolle über bestimmte Aspekte des Arrangements und wie sich diese Kontrolle verschiebt. Auch im Bereich der neueren KI ergibt sich hierbei eine komplexe Lage von Verschiebungen, die nicht ohne weiteres in einfachen Thesen ausgebreitet werden kann. Es kommt auf die Binnenstruktur eines gesellschaftlichen Bereichs ebenso an, wie auf spezifische Anwendungsbedingungen und die hier vorherrschenden Möglichkeiten technische Systeme zu trainieren. Klar ist, die Systeme sind meistens nur so gut wie ihre Passung für den jeweiligen Bereich und ihre Unterfütterung mit ausreichendem Material (in qualitativer wie quantitativer Hinsicht), um diese Passung zu erreichen. Gleichzeitig verschieben sich damit auch die Möglichkeiten oder Einschränkungen für andere Beteiligte. Ein weiteres Lernergebnis betrifft die Auswirkungen der Opakheit der Systeme. Wenn unklar ist, worauf sich ein Lernsystem stützt, sind auch seine Ergebnisse nur schwer nachzuvollziehen und daraus entsteht letztlich auf sozialer Ebene eine Vertrauensproblematik, die durch mehr oder weniger aufwendige begleitende Verfahren dann wiederhergestellt werden muss. Dieser Befund verschlimmert sich, je komplexer die Lage im soziotechnischen Arrangement sich gestaltet, also je mehr einzelne Schritte und Entitäten an

der Ergebnisproduktion beteiligt sind und dort eventuell undokumentierte Entscheidungen treffen. Für alle hier aufkommenden Problemlagen ist die Einbettung der KI-Verfahren in konkrete soziotechnische Arrangements entscheidend und für die Einbettungsfrage bietet eine relationale Techniksoziologie eine fruchtbare Grundlage, die hier einmal kurz zusammengefasst werden soll.

3. Warum sind gerade Erkenntnisse aus der Diskussion um ML für eine Techniksoziologie mit White interessant?

Bevor wir zu diesen Grundzügen kommen, soll hier jedoch noch einmal eine etwas andere Perspektive eingenommen werden, die sich der Frage stellt, warum die Diskussion um MachineLearning-Verfahren für die Entwicklung einer an White orientierten Techniksoziologie so gut fruchtbar gemacht werden kann. Techniksoziologisch ist jedes technische Artefakt eingebettet in einen soziotechnischen Kontext, welcher aus anderen Artefakten und Akteur:innen besteht. An den Verfahren aus dem Machine Learning, vor allem den an neuronalen Netzen orientierten Deep Learning-Verfahren, ist nun ihre explizite Einbettung und Angewiesenheit auf Kontext das Besondere. Schon durch die Einspeisung von Lerndaten ist ein expliziter, wenn auch beschränkter Kontextzugang gewährleistet, der das Blackboxing der Technik grundlegend problematisch macht, was man bei abgeschlosseneren Technologien eventuell eher ignorieren kann. Wie auch die Beiträge in diesem Band deutlich gemacht haben, ist es aber genau die Diskussion dieser Einbettungsproblematik und ihre Offenlegung, die besondere Stärken der relationalen Zugangsweise hervorhebt und es zulässt diese Stärken, die für jegliches soziotechnische Arrangement gelten, deutlich zu machen. Daher bieten diese Verfahren einen sehr guten Einstiegspunkt für eine an Harrison White orientierte relationale Techniksoziologie.

4. Ausblick: Grundzüge einer relationalen an White orientierten Techniksoziologie

Aus dieser Durchsicht der Problemlagen, sowie der Beschreibungs- und Erklärungsangebote einer dezidiert relational operierenden Theorie, wie derjenigen von White, für die neue KI ergeben sich auch einige generelle theore-

tisch-konzeptuelle Bausteine einer relationalen Techniksoziologie, die wir an dieser Stelle abschließend noch kurz darlegen wollen.

Kontrolle/Identität in soziotechnischen Netzwerken

Ausgangspunkt ist hier die Idee der Dualität von Identität und Kontrolle, die sich auch auf materiale Gegebenheiten und technische Artefakte ausdehnen lässt. Auch hier geht es darum, dass diese nicht einfach bio-physikalische Gegebenheiten in sozialen Zusammenhängen sind, sondern Identitäten, die auf ihre eigene Weise sozial relevant werden können, indem sie Kontrolle ausüben, sich mit anderen Kontrollprozessen verstricken und in Erzählungen ihre soziale Bedeutung verhandelt wird.

Die Actor-Network Theory (ANT) hat eindringlich aufgezeigt, dass soziale Akteur:innen nicht ohne die Verknüpfung mit technischen Aktanten existieren können. Angefangen von der Kleidung (Textiltechnik), über Schreibutensilien und Computer bis zu so genannten »large technological systems« (LTS) wie dem Energiesystem und dem Internet sind wir als positionierte soziale Identitäten mit technischen Prozessen verknüpft. Sie stellen eine *conditio sine qua non* unserer Existenz als soziale Akteur:innen dar, sodass beispielsweise wir – die beiden Autoren dieses Beitrags – als Sozialwissenschaftler gerade agieren können. Würde man uns alle diese Technologien nehmen, wären wir als Wissenschaftler schlichtweg nicht mehr aktionsfähig. Mit anderen Worten: Das Handeln- und Kommunizierenkönnen in Gesellschaft ergibt sich aus der Verknüpfung menschlicher und nicht-menschlicher Komponenten. Wir leben immer schon in soziotechnischen Wirklichkeiten.

Korrekter wäre es zu sagen, dass mit der Relationierung auch stets eine wechselseitige Positionierung einhergeht. Dies wird schon an ganz einfachen Beispielen sinnfällig: Das Schwert in der Hand meines stürmenden Körpers macht mich zu einem Angreifer, wie ich aus ihm eine Angriffs- waffe mache. Dieses einfache Beispiel zeigt deutlich, dass die Kopplungen von Identitäten in den Blick zu nehmen sind. Die Positionierungen erfolgen in der Regel unmerklich und doch spüren wir mit jedem Technisierungsschub (insbesondere bei den Informations- und Kommunikationstechnologien, kurz: IKT), dass wir uns grundlegend anders positioniert wiederfinden.

Der Aspekt der Positionierung leitet zum Aspekt der Identitätskonstruktion über. Da White ja verschiedene Identitätsstufen differenziert und hier recht basal auf der ersten Stufe mit einer Verortung im Raum beginnt, lässt sich ein solches Identitätsverständnis sehr wohl auf jegliche Technik anwenden. Denn

selbst ein behauener Stein im steinzeitlichen Kontext weist, so gesehen, eine Identität auf: Er ist beispielsweise für das Abziehen von Fell bei erjagten Tieren vorgesehen. Verortung ist dabei eine wichtige Basiskategorie für die Zuschreibung von Identität. Relational-soziologisch gesprochen, geht es um Positionierung in relationale, mithin netzwerkartige Verhältnisse. Neben dieser basalen Form der Identitätsbildung durch Positionierung in einem Netzwerk differenziert White (2008: 17f.), wie bereits dargelegt, vier andere Identitäten, die wie eine Stufenfolge der Komplexitätssteigerungen zu begreifen sind. Identitäten sind bei alledem permanent bedrohte Entitäten, welche »Identitätsarbeit« in Form von Kontrollprojekten verrichten müssen, das heißt, selbst beginnen müssen, Relationen zu knüpfen, um ihre erlangte Position in einem Netzwerk zumindest vorübergehend zu sichern.

Zwar hat White Technik nicht explizit als mögliche Kandidat:in für derartige Identitätskonstruktionen in Erwägung gezogen, jedoch ist seine auf Skalenfreiheit abzielende Begrifflichkeit hinreichend abstrakt, um sie auch auf diese Phänomene anzuwenden. So ist klar, dass jeglicher Technik zumindest eine Positionierung im Netzwerk attestiert werden kann. Aber auch die zweite Form ist üblich, wenn es sich zum Beispiel um Ordnungsphänomene handelt: alle Gewehre in Abgrenzung zu allen anderen Waffengattungen. Insbesondere bei Techniken mit mehr Freiheitsgraden ist aber auch die dritte Form der Identitätskonstruktion zu finden: Das Auto wird in dem einen Kontext als Fortbewegungsmittel genutzt, im anderen als Statussymbol und im dritten als Schlafmöglichkeit. Diese Genese der Autoverwendung kann natürlich auch in einem Narrativ kulminieren, welches für die vierte Form der Identität nach White steht. Bei selbstlernenden Systemen (und Robotersystemen) sind rudimentäre Formen der Selbstidentifizierung gegeben. Sie kommen der fünften Form der Identitätskonstruktion nahe, etwa wenn ein humanoider Roboter über ein eigenes Körpermodell verfügt, um in fremden Räumen zu navigieren und dabei unterscheiden zu können, ob es sich um ein gegebenenfalls störendes anderes Objekt handelt, oder um ein eigenes Bauteil.

Diese fünf Formen der Identitätskonstruktion erlauben einen differenzierenden Zugriff auf die Welt der Technik und verdeutlichen, dass der janusköpfige (Kontroll-)Prozess des Relationierens und Deutens mehrere abstrahierende Schleifen in der Welt der soziotechnischen Netzwerke durchlaufen kann.

Wir schlagen vor, soziotechnische Relationen als Kontrollprojekte zu begreifen. Augenscheinlichste Beispiele für das Wirksamwerden der Technik als Kontrollprojekt bilden Großtechnologien, wie das Elektrizitätsnetz oder das Internet, von denen sich Gesellschaften in umfassender Weise abhängig ge-

macht haben. Und in umgekehrter Richtung bilden Küstenschutzmaßnahmen wie Deiche und Dämme Kontrollmöglichkeiten gegenüber der Natur, um die Gesellschaft durch Technik vor den Gezeiten und Sturmfluten zu schützen. Aber auch jeder Einbau von Technik in den sozialen Alltag, angefangen von der Waschmaschine bis hin zum Privat-PC, bewirkt eine grundlegende Änderung eingespielter sozialer und kultureller Verhaltensweisen. Spätestens seit den SCOT-Ansätzen¹ ist augenscheinlich geworden, dass diese Beeinflussung bi-direktional ist (vgl. Bijker/Pinch 1984). D.h., was eine technische Neuerung ist, wie sie in die Alltagspraxen der Menschen eingebaut wird, ist mehr oder weniger variabel. Und erst diese Aushandlungs- und Aneignungsprozesse (die ja mit unseren Worten nichts anderes als Kontrollprojekte bilden) bringen Technik im sozialen Kontext überhaupt zum Funktionieren. An diese Überlegungen anknüpfend kann postuliert werden, dass ein wechselseitiges Kontrollprojekt am Werke ist, dessen Wirksamwerden dazu führt, dass sich soziale und technische Identitäten in einem (oder sogar mehreren) soziotechnischen Arrangement(s) positionieren.

Um soziotechnische Kontrollprojekte besser an die Terminologie von White anschließen zu können, sollen diese eine Erweiterung erfahren, indem von Sociotechnical Ambage, Technocultural Ambiguity und Sociotechnical Decoupling bzw. Blocking Action die Rede sein soll. Wie hat man sich derartige Kontrollprojekte nun konkret bei soziotechnischen Konstellationen vorzustellen? Die auf soziotechnische Interaktivität ausgerichteten technischen Operationen (= »sociotechnical ambage₁«, kurz: sa₁) resultieren aus einer verkörperten Funktionsdisposition (z.B. beim Hammer) und/oder einer Aktorik (z.B. bei einem Fließband) und/oder verkörperten Kognition (z.B. bei Multi-Agenten-Systemen). Sie müssen von der sozialen Identität als Versuche, sich in das laufende Geschehen einzuklinken, gedeutet werden (= »technocultural ambiguity₁«, kurz: ta₁), um als relevante Beiträge zu erscheinen. Umgekehrt müssen Aktivitäten von sozialen Identitäten (= »sociotechnical ambage₂«, kurz: sa₂) seitens der Technik »identifiziert«, »analysiert« und »bewertet« werden (= »technocultural ambiguity₂«, kurz: ta₂). Handelt es sich bei diesen sozialen Identitäten um menschliche Akteur:innen, dann bringen diese körperliche und bewusstseinsmäßige Spezifika in die Prozesse mit ein. Wenn sich diese Kontrollprojekte miteinander verkoppeln, kommt eine soziotechnische Relation zustande (vgl. Abbildung 1). Eine solche Relation bildet also eine Verkopplung technischer Operationen (zur Leistungserbringung)

1 SCOT steht für »Social Construction of Technology«.

mit sozialen Prozessen (z.B. Deutungsschemata), mit körperlicher Motorik (z.B. Bedienung einer Maschine), mit Kognitionen (z.B. Verstehen des gerade wahrgenommenen Vollzugs technischer Operationen) und mit Emotionen (z.B. Gefühl der Überforderung).

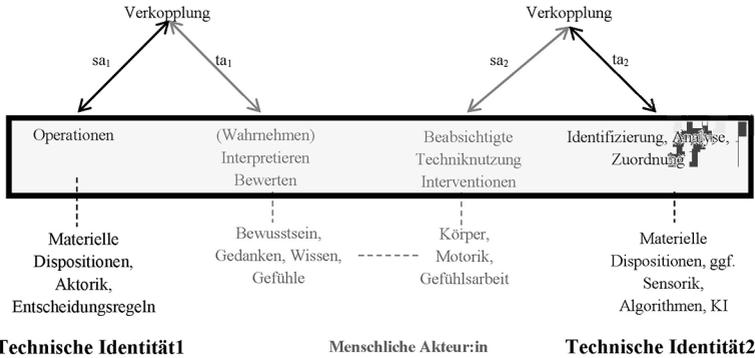


Abbildung 1: Verschränkung der Kontrollprojekte sa_x und ta_x

An diesen Beschreibungen sollte zum einen deutlich geworden sein, dass nicht-menschliche Identitäten und menschliche Identitäten Unterschiedliches in die Netzwerkprozesse einbringen. Während bei Personen so etwas wie »mindfulness«² ein Alleinstellungsmerkmal darstellt – also Empfinden zu können, ein Zusammenhangs- und schließendes Denken zu besitzen, mit komplexen Situationen umzugehen und kreative Entscheidungen in Unsicherheit zu treffen –, sind bei technischen Identitäten die immense Rechenleistung, eine Unermüdlichkeit ihres Operierens sowie die hohe Präzision ihres Bewirkens als einige wesentliche Spezifika hervorzuheben. Es gilt nun, genau diese heterogenen Fähigkeiten in den ablaufenden soziotechnischen Prozessen so aufeinander zu beziehen, dass Anschlussfähigkeit erzeugt wird. Zum anderen geht es um eine Reformulierung des Sozialen schlechthin als relationales Ereignis: Egos Beitrag zu den ablaufenden Prozessen (sa_1 oder

2 Weick/Sutcliffe (2003): Das Unerwartete managen. Wie Unternehmen aus Extremsituationen lernen.

sa₂), der durch Alter gedeutet wird (ta₁ oder ta₂), und der Beitrag Alters, der durch Ego gedeutet wird, verschränken sich nur dann dauerhaft anschlussfähig zu Interaktionen, wenn beide Beitragstypen ihren jeweiligen Part spielen. Interaktionen bzw. Interaktivitäten³ haben mit anderen Worten unauflöslich eine kognitive und eine materielle Seite⁴, sodass sich eine soziologische Theorie der Interaktion und Interaktivität ›zwischen‹ beiden aufzuhalten hat – und nicht etwa die eine Seite zugunsten der anderen unterschlägt. Kein Prozess hat per se den Status, eine Interaktion bzw. Interaktivität zu sein. Hierzu ist eine geglückte Relationierung dieser beiden Beitragstypen (sa_x und ta_x) erforderlich, die als Keimzelle (im Sinne einer ›Urrelation‹) des Sozialen bzw. genauer des Soziotechnischen überhaupt begriffen werden kann.⁵ Es handelt sich um komplementäre Hälften, die sich als konstruktive Leistungen von verschiedenen Identitäten verkoppeln, sodass ein Beitrag als Handlung, Kommunikation oder Operation erscheint. Dabei ist zu berücksichtigen, dass jede beteiligte Identität im Netzwerk eine eigene Position einnimmt, eine spezifische ausschnittshafte (Deutungs-)Perspektive besitzt und ›nur‹ über einen bestimmten Satz an Eingriffsmöglichkeiten verfügt.⁶ In einer Identität kann Leben oder Strom ›toben‹; sie kann mit menschlichen Sinnen und Bewusstsein oder spezieller Sensorik und Algorithmik ausgestattet sein; das Verarbeitete setzt sich in ihr in Expressionen oder Aktorik um. Maßgeblich sind identifizierbare Formen von Kontrolle, die sich miteinander verbinden und damit Positionierung ermöglichen.

Kontext/Einbettung/Entbettung

Netzwerktheoretisch geht es immer um die Frage der Einbettung von soziotechnischen Konstellationen in weitere Zusammenhänge und die Möglichkeiten sich von diesen unabhängig zu machen. In Whites Theorie stellt Kontext eine Beziehung her, deren Bedeutung ebenso durch Kontrollprojekte konfiguriert wird. Auch Kontext wird damit gleichzeitig skalierbar, situationsspe-

3 Unter Interaktion wird hier Face-to-Face-Interaktion verstanden, während Interaktivität Kommunikation mittels eines dazwischengeschalteten technischen Mediums (z.B. Buch, Fernseher, Internet) darstellt.

4 Vgl. auch *Kreckel* (1992): Politische Soziologie der sozialen Ungleichheit.

5 Vgl. *Häußling* (2006): Interaktionen in Organisationen. Ein Vierebenenkonzept des Methodologischen Relationalismus und dessen empirische Anwendung, S. 105ff.

6 Vgl. *Granovetter* (1985): »Economic Action and Social Structure: The Problem of Embeddedness«, in: *American Journal of Sociology* 91, S. 481–510.

zifisch und für Kontrollversuche und Identitätskonstruktionen zugänglich. Technik steht in einem mehrfachen Verhältnis zu diesen Aspekten der Einbettung: Erstens geht es immer um ein soziotechnisches Netzwerk, das in weitere Kontexte eingebettet ist. Die Vorstellung, das Soziale sei feinsäuberlich vom Nicht-Sozialen, insbesondere vom Technischen separierbar (so etwa bei Durkheims Soziologieverständnis), hat sich spätestens mit der Hoffähigwerdung der ANT in der Mainstream-Soziologie verabschiedet. Vielmehr wird das »Soziale« in vielfältiger und substantieller Weise von technischen Prozessen durchzogen, sodass alle soziologischen Grundbegriffe, wie die Akteur:in, die Handlung, die Gruppe etc. zu einem Gutteil auf der Wirkung von Techniken gründen. Etwa: »Der Rennradfahrer X nimmt mit seiner Mannschaft an einem Radrennen teil.« Es ist offensichtlich, dass Technik (auch im engeren Sinne) immanent für dieses soziale Ereignis ist.

Zweitens ist auch der Kontext soziotechnisch. D.h. jedes betrachtete soziotechnische Netzwerk ist abhängig von Technik, die sich in seinem Kontext lokalisieren lässt. Fast schon trivial ist, darauf hinzuweisen, dass jede soziale Interaktion in der Moderne die Existenz Großer Technischer Infrastruktursysteme voraussetzt, wie zum Beispiel das Energienetz, die Wasserversorgung, das Straßennetz, das Flugnetz, das Internet etc. Aber auch niedrigskaligere Kontexte sind für je konkret im Fokus stehende Netzwerke anzusetzen: Wie zum Beispiel für eine Fußballmannschaft die entsprechende Liga, in der man spielt. An derartigen Beispielen wird auch klar, dass die Kontexte die gleiche netzwerkartige Struktur aufweisen, wie die jeweils forschersich fokussierten Netzwerke. Mit anderen Worten unterscheidet sich auf dieser strukturellen Ebene nicht der Forschungsgegenstand von seinem Kontext. Vielmehr haben Netzwerkforscher:innen immer wieder betont, dass es streng genommen nur ein einziges soziales, genauer soziotechnisches bzw. soziomaterielles Netzwerk gibt, welches die ganze Welt umspannt – auch wenn es löchrig ist und eine sehr geringe Dichte aufweist (siehe z.B. die *small world*-Ansätze, oder Baecker 2007). Insofern ist Einbettung eine analytische Unterscheidung, die aus bestimmten inhaltlichen Gründen Grenzen in diesem *einen* Netzwerk zieht. Oder anders formuliert: Um etwas, als etwas untersuchen zu können, muss es zunächst von seinen weiterreichenden Relationierungen entbettet werden, sodass der Begriff Einbettung nur auf den Cut, den man als Forscher:in gemacht hat und machen musste, hinweist. Damit will nicht gesagt werden, dass alles gleich ist. Vielmehr können im soziotechnischen Globalnetzwerk Netzwerkareale beobachtet werden, deren jeweiligen Elemente wirkungsvoll aufeinander abgestimmt sind, wie zum Beispiel politische Bewegungen. Sie weisen In-

stitutionen und Stile auf, die sie von dem Rest unterscheidbar hält. Aber es ist verwickelter, als ein Schubkastendenken oder ein Denken in Reinheitsgeboten. Das Unterscheidbare ist doch mannigfaltig miteinander relationiert, damit ähnlich dem untereinander Ähnlichen und doch anders. Hier kann eine analytische Betrachtung der Types of Ties, wie sie konsequent in der Netzwerkforschung zur Anwendung kommt, beträchtliche Detailierungs- und Differenzierungsgewinne gegenüber anderen Ansätzen vorweisen.

Drittens wird häufig unter Zuhilfenahme von Technik diese Entbettung vollzogen, wie beispielsweise die Straßenschilder, die den Wechsel in ein anderes Bundesland markieren. Noch technischer wird der Prozess der Entbettung, wenn man an die unterschiedlichen Communities in den Social Media denkt, bei denen bereits technische Funktionalitäten (wie zum Beispiel die Zeichenbeschränkung für Tweets bei X, ehemals Twitter) einen Unterschied gegenüber anderen Plattform-Communities markieren. Und auch die Einbettung wird, insofern sie für einen abgegrenzten sozialen Ausschnitt mitkommuniziert wird, in der Regel nicht ohne Technik erfolgen. Man denke an Links, welche Inhalte verschiedener Social Media-Seiten mit- und zueinander relationieren und gleichzeitig die Grenze, die vorzugsweise durch Entbettung vollzogen wird, aufrechterhalten.

Viertens können diese soziotechnischen Netzwerkareale jeweils aus eigenen Kräften heraus wirkungsvolle Kontrollprojekte in Richtung der Kontexte, in die sie eingebettet sind, vollziehen. Auch diese Kontrollprojekte greifen oftmals in essentieller Form auf Technik zurück, wie zum Beispiel die Berliner Mauer zu Zeiten des Kalten Krieges, oder die NSA bei der Ausspionierung fremder Staaten bzw. Gesellschaften. Gerade die klar vollzogene Entbettung führt hierbei zu neuen Möglichkeiten der kontrollierenden Einbettung des Entbetteten. Der anti-kapitalistische Schutzwall etwa markiert den Feind ganz augenscheinlich und gewinnt mit dem dadurch auch symbolisch vollzogenen Narrativ die Möglichkeit, jede Aktivität des Feindes als kapitalistisch und damit das eigene politische System (der DDR) gefährdend hinzustellen und neue Handlungsmöglichkeiten aus der eigenen Logik heraus zu generieren, die anderenfalls kaum möglich gewesen wären. Oder in Bezug auf das zweitgenannte Beispiel: Gerade weil die NSA im Verborgenen operiert, sich damit vor den Beobachtungen der Beobachteten, oder Dritten entbettet, hat sie ungeahnte Kontrollmöglichkeiten über deren Aktivitäten gewonnen. So wäre auch das Blackboxing von KI-Verfahren als eine solche Entkopplung verstehbar (siehe auch den Beitrag von Heckwolf/Schmitt in diesem Band).

Aus diesem mindestens vierfachen Technikbezug von Einbettung/Entbettung ergeben sich für die (technik)soziologische Forschung spannende Perspektiven, bei denen die Netzwerkforschung ihr volles Potenzial an methodischer und konzeptueller Prägnanz entfalten kann. Etwa wenn man den jeweiligen Technisierungsgrad und das Portfolio an eingesetzten Techniken im Vergleich zwischen dem jeweils im Forschungsfokus stehenden Netzwerk und seinem unmittelbaren Kontext analysiert. Des Weiteren kann der Forschungsschwerpunkt auch darauf gelenkt werden, inwieweit Techniken der Einbettung die ›offiziellen‹ soziotechnischen Kontrollprojekte eines zu untersuchenden sozialen Netzwerkareals unterlaufen. Mithin geht es aber auch ›schlichtweg‹ um die Erforschung des technologischen Stils und korrespondierender Institutionalisierungen via Technik, die bei einem sozialen Gebilde – wie zum Beispiel bei der Ausgestaltung von Metaverse – beobachtbar sind. Dies sind nur wenige und kursorische Beispiele für spezifische (technik)soziologische Forschungsfragen, die aus einer konzeptuellen Scharfstellung auf Entbettungs- und Einbettungsprozessen gewonnen werden können.

Relationale Reichweite

Das Reichweite über den Aufbau von Relationen realisiert wird, ist eine der zentralen Beobachtungen aus der Theorie von White. Dabei ist es bedeutsam, Identitäten als Spuren transportabel zu machen, um über Kontexte hinweg Beziehungen aufzubauen. Dass hierbei Technologien und technische Artefakte eine wichtige Rolle spielen, ist techniksoziologisch nicht neu, dass dies aber über Kontrollversuche verwirklicht werden muss, die sich in der Kombination mit anderen Kontrollversuchen bewähren müssen, bietet jedoch eine weitreichende Perspektive zur Analyse dieser Prozesse.

Wie in den ersten zwei Abschnitten (S. 259ff.) einer an White angelehnten Techniksoziologie deutlich geworden sein sollte, besteht ein großer Mehrwert der Theorie in ihrem klar ausbuchstabierten Skalierungsmodellen sozialer Phänomene. Im Bereich des Duals von Identität und Kontrolle (im Sinne von Breigers Duality of Structure and Culture, s. Breiger 2010) hat man gesehen, dass einzelne situative Positionierungen transportabel gemacht werden können und so zu immer größeren und damit komplexeren Identitäten aggregiert werden. Dies hängt aber davon ab, dass sich der Transport über erfolgreiche Kontrollversuche realisieren lässt. Skalierungen werden also von weiteren Positionierungen bedingt und damit von dem Schaffen, oder auch von dem

Auflösen von Verbindungen. Dies hatten wir unter die Überschriften der Einbettung und Entbettung gestellt. Skalierung hat in diesem Modell ganz klar etwas mit der Reichweite solcher Beziehungen zu tun, aber auch mit der stärkeren Integration solcher Beziehungen in neue Identitätsformationen. Netzwerktheoretisch ist dieses Reichweiten-Modell wohl paradigmatisch am Phänomen der »small worlds« (Watts 2004) beschrieben worden. In einem Netzwerk, in dem die meisten Personen nur ihre direkten Nachbarn kennen, zirkulieren Ideen langsam. Aber schon, wenn es nur wenige Verbindungen gibt, die quer durch das gesamte Netzwerk laufen, verkürzen sich die Wege insgesamt enorm und die Geschwindigkeit, mit denen Informationen zirkulieren, steigt merklich an. Das es schon hier ein Leichtes ist, technologische Bedingungen für diese Brücken auszumachen, liegt auf der Hand – etwa wenn der Telegraph die Nachrichten aus fernen Ländern günstig und schnell zu uns bringen kann. Sehr schön deutlich werden hierbei zwei Gesichtspunkte dieser Reichweitenvergrößerung. Einerseits sind es Verbindungen, die selbst aus lokalen oder situativen Kontexten hinausweisen und andererseits ist auch die Art der Realisierung (oder Materialisierung) dieser Verbindung von Bedeutung. Ob man Briefe von der Schwester aus Australien bekommt, oder über Facebook in Kontakt steht, hat immensen Einfluss auf den Kommunikationsaustausch. Reichweite hängt in dieser Lesart vor allem von zwei Bedingungen ab, dem Finden von Anknüpfungspunkten in anderen Netzwerkarealen und den Transportgegebenheiten dieser Verbindung. Beide Dimensionen sind für die Realisierung von Reichweiten und damit Skalierungseffekten essenziell. Technik und technische Artefakte sind ebenfalls an beiden Punkten in diese Netzwerke eingefügt.

Reichweiten von Identitäten werden also über Beziehungen und Transportmöglichkeiten erzeugt⁷, aber wie hat man den Begriff von Transportmöglichkeit in diesem Zusammenhang genau zu verstehen? Wodurch werden Identitäten transportabel und können in Kontrollprojekten mit Reichweitengewinnen eingesetzt werden? In der Innovationsforschung ist dies unter dem Begriff der Diffusion von Innovationen schon lange ein Thema (vgl. Rogers 1995), aber auch grundsätzlich stellt sich die Frage nach der Reichweite von

7 Reichweite bedeutet hier immer das etwas aus einem spezifischen Kontext heraus Konsequenzen für weitere andere Kontexte gewinnt, also etwa der Einsatz von KI-Verfahren sich im Gesundheitswesen bewährt und dann auch in das Bildungswesen eingeführt wird. Reichweite kann dabei über ganz unterschiedliche Verknüpfungen gewonnen werden.

sozialen Identitäten auf der Ebene von Ideen, Organisationen, Technologien, oder auch einzelnen Personen.

Im nach Komplexität gestuften Identitätskonzept von White haben wir schon einen wesentlichen Ermöglicher von Reichweitengewinnen kennengelernt. Positionierungen lassen sich transportabel machen, indem sie in Geschichten sprachlich verdichtet werden. Sprachliche Verdichtung schafft eine kommunikative Transportabilität, die nicht zu unterschätzen ist. Dabei ist besonders zu beachten, dass eine Stärke von sprachlicher Verdichtung darin liegt, den Wechsel von Identitäten zwischen Kontexten zu ermöglichen. Dies wird durch das Geben von Namen unterstützt. Dingen und Personen Namen zu geben, macht es möglich, von beliebigen Kontexten auf etwas zu verweisen. Dadurch entstehen Verweisungszusammenhänge, denen man folgen kann.

Manche Identitäten sind leichter zu transportieren als andere und die Art der Verdichtung und des Transports spielen ebenfalls eine Rolle. Für eine techniksoziologische Lesart, die technische Artefakte, technologische Konzepte, oder ganze technologische Arrangements als Identitäten im von White entwickelten Sinne betrachtet, bedeutet dies, dass ihre Reichweite über Anknüpfungspunkte und Transportmöglichkeiten analysiert werden sollten. Nehmen wir das technologische Konzept neuronaler Netze, das dem technologischen Arrangement der neueren KI zugrunde liegt: Hier war eins der frühen Probleme das Fehlen von Anknüpfungspunkten im Bereich der Umsetzung, da die Netze sehr viel Rechenkapazität für sehr einfache oder geringe Fortschritte benötigten. Dies reduzierte ihre Nutzbarkeit in unterschiedlichen Kontexten und damit ihre Möglichkeit, neue Kontexte für die Identität zu erschließen. In den letzten Jahren sind dann zwei Entwicklungen zu beobachten gewesen, die diese Situation für das technische Konzept änderten. Zum einen ergab sich ein erneuter Anknüpfungspunkt innerhalb der technischen Arrangements durch die stark gestiegene Rechnerleistung der aktuellen Computersysteme. Zum anderen waren in verschiedenen Bereichen nun große digitalisierte Datenkorpora zugänglich, die man zum Lernen nutzen konnte. Beides zusammen schuf neue Anknüpfungspunkte für ein schon von Beginn an gut transportierbares Konzept.

Reichweite im Sinne der Skalierbarkeit von sozial relevanten Identitäten, zu denen nach unserer Auffassung auch Artefakte, Konzepte und soziotechnische Arrangements zählen, kann ein starkes Konzept der Erklärung sozio-

technischen Wandels werden und macht die sehr unterschiedlichen Wege deutlich, über die ein solcher Wandel erfolgen kann.⁸

Unbestimmtheit und technische Funktionalität

Die Kontingenzzhematik durchzieht die Theorie Whites nachhaltig. Wichtig ist in diesem Zusammenhang die Überlegung, dass Kontingenz dabei nicht gelöscht, sondern nur transformiert werden kann. Auch technisch realisierte Funktionalitäten bleiben der Kontingenz ausgesetzt, bringen sie nur in eine andere Form oder verlagern sie auf andere Identitäten. Wohin Kontingenz und damit auch Unsicherheiten verlagert werden, kann mit White zu einem Kernthema techniksoziologischer Analysen werden und hilft einen frischen Blick auf Begriffe wie Risiko und Nebenfolgen zu werfen. Die Techniksoziologie ist schon immer mit dem Verhältnis von Technologien als Apparaturen, die bestimmte Probleme lösen sollen und Technologien als Auslöser neuer Probleme beschäftigt. Nebenfolgen oder Risiken von Technologien sind ein wichtiger Bezugspunkt theoretischer wie empirischer sozialwissenschaftlicher Beschäftigung mit Technik.

Mit White kann man sehen, dass Technik immer Teil von Kontrollversuchen, -projekten und -verstrickungen ist und dabei selbst Kontrolle gewinnt. Mit einer derartigen Einbindung in diese Vollzüge ist aber immer auch das Element der Unsicherheit verbunden. Grundeinsicht bei White ist, wie gesagt, dass Kontrollversuche die Unsicherheit/Kontingenz immer nur transformieren und nicht tilgen können. Aber was bedeutet das für die spezifische Transformation, die mit dem Begriff der Technik verbunden ist? Aus einer systemtheoretischen Lesart, die nach Luhmann auf Kommunikation scharf stellt, wird Technik als funktionierende Simplifikation beobachtet (Luhmann 1997). Technik entlastet dadurch die Kommunikation, da kommunikativ davon auszugehen ist, dass die angebotene Simplifikation funktioniert. Simplifikation ist sehr dicht an Verdichtung, von der oben schon mal die Rede war und Verdichtung ist ein Schlüsselement, um Transportfähigkeit zwischen Kontexten zu erreichen. Wenn wir es hier mit Vereinfachungen zu tun haben, die in unterschiedlichen Kontexten funktionieren, dann wird dadurch eine bestimmte Art von Unsicherheit, nämlich die einer stabilen Übertragung, reduziert.

8 Siehe für eine explizite Ausarbeitung dieses Verständnisses von Reichweite Schmitt/Häußling/Kaip (2023).

Wenn wir uns eine konkrete technische Lösung ansehen, nehmen wir die berühmten Brücken im Aufsatz von Langdon Winner »Do Artifacts Have Politics« (1980)⁹, dann wird klar, wie sich die Transformation von Unsicherheit näher beschreiben lässt. Die These des Aufsatzes ist ja, dass die Konstruktion der Brücken eine Kontrollstrategie darstellt, um Schwarze vom Besuch des New Yorker Viertels Staten Island abzuhalten. Dazu sind die Brücken so konstruiert, dass Busse des öffentlichen Nahverkehrs sie nicht unterqueren können. Einer sozialen Struktur wird hier demnach eine materiale Form verliehen. Dabei kommt es sowohl zu einer Reduktion von Unsicherheit, indem man direkt die physische Zugänglichkeit kontrolliert. Es kommt aber auch zu einer Transformation der Unsicherheit, da sie an das Bestehen bestimmter sozialer Gegebenheiten gekoppelt ist, die nicht durch die materiale Struktur selbst zu kontrollieren sind – wie etwa, dass Schwarze deutlich häufiger mit öffentlichen Verkehrsmitteln unterwegs sind. Es kommt also tatsächlich zu einer Härtung des Sozialen (im Sinne Latours) durch die materiale Form, welche die Kontrolle hierbei realisiert, die aber vom Erhalt weiterer Kontextbedingungen abhängig ist, die sich ändern können. Hierhin verlagert sich die Unbestimmtheit und man hat dann techniksoziologisch jeweils zu fragen, wohin sie sich bewegt und welche Formen die technisch-materielle umgesetzte Struktur tragen. So werden technische Artefakte und Arrangements als Kontrollprojekte sichtbar, die bestimmte Elemente durch biophysikalische Bestimmtheiten feststellen, dafür aber an andere Stelle auch neue Wege für Kontrollversuche öffnen, die ihrerseits als Risiko oder Nebenfolgen beobachtbar werden. Eine Techniksoziologie, die auf diese Verwicklungen von Kontrollversuchen in unterschiedlichen Formen scharf stellt und dabei Reichweiteneffekte nachvollziehbar machen kann, ist in der transformativen gesellschaftlichen Situation, die sich nicht nur, aber auch durch die neuen Techniken im Bereich der KI vorangetrieben wird, besonders vielversprechend. Kontrolle, Identität, Kontextwechsel und Skalierung/Reichweite sind Konzepte, die es ermöglichen, technisch-materiale Identitäten und ihre Rolle in Kontrollverflechtungen auf unterschiedlichen Skalierungsebenen zu analysieren und dabei auf die Probleme

9 Es sollte hier nicht unerwähnt bleiben, dass Winner mit seinem Aufsatz zum Teil heftige Kritik auf sich zog. Diese orientiert sich aber weitgehend an einer zu unsystematischen Darlegung des Brückenbeispiels seitens Winners, welche dann zu Fehldeutungen führte. Die weitreichenden techniksoziologischen Implikationen, die uns an dieser Stelle interessieren, wurden hingegen selten systematisch aufgearbeitet (vgl. hierzu ausführlicher Joerges 1999).

des Transformationspotenzials ebenso scharf zu stellen, wie auf die Verlagerung von Unsicherheiten.

5. Schluss

Der Sammelband konnte durch die Verknüpfung relationaler techniksoziologischer Konzepte und Interviewmaterial von Forscher:innen aus dem Feld der Entwicklung neuerer KI-Verfahren einen Beitrag zu einem tieferen soziologischen Verständnis dieser Entwicklungen bieten und vor allem die damit mögliche systematische Aufarbeitung der gesellschaftlich relevanten Problemlagen bietet einen guten Ausgangspunkt sich mit den weiteren Ausprägungen dieser Entwicklungen (wie etwa die großen Sprachmodelle von z.B. ChatGPT) eingehender auf der Basis eines fruchtbaren theoretischen Werkzeugkastens zu beschäftigen.

6. Literatur

- Baecker, Dirk (2007): »The network synthesis of social action I: Towards a sociological theory of next society«, in: *Cybernetics & Human Knowing* 14(4), S. 9–42.
- Bijker, Wiebe E./Pinch, Trevor J. (1984): »The Social Construction of Facts and Artifacts: Or How the Sociology of Science and the Sociology of Technology Might Benefit of Each Other«, in: Wiebe E. Bijker/Thomas. P. Hughes/Trevor Pinch (Hg.), *The Social Construction of Technological Systems*, Cambridge/Massachusetts: MIT Press, S. 17–50.
- Breiger, Ronald L. (2010): »Dualities of culture and structure: Seeing through cultural holes«, in: Sophie Mützel/Jan A. Fuhse (Hg.), *Relationale Soziologie: Zur kulturellen Wende der Netzwerkforschung*, Wiesbaden: VS Verlag für Sozialwissenschaften, S. 37–47.
- Granovetter, Mark (2018): »Economic action and social structure: The problem of embeddedness«, in: *The sociology of economic life*, London: Routledge, S. 22–45.
- Häußling, Roger (2006): »Ein netzwerkanalytisches Vierebenenkonzept zur struktur- und akteursbezogenen Deutung sozialer Interaktionen«, in: Bettina Hollstein/Florian Straus (Hg.), *Qualitative Netzwerkanalyse*, Wiesbaden: VS Verlag für Sozialwissenschaften, S. 125–151.

- Joerges, Bernward (1999): »Do politics have artefacts?«, in: *Social studies of science* 29(3), S. 411–431.
- Kreckel, Reinhard (1992): *Politische Soziologie der sozialen Ungleichheit*, Frankfurt a.M.: Campus Verlag.
- Luhmann, Niklas (1997): *Die Gesellschaft der Gesellschaft*. 2 Bände, Frankfurt a.M.: Suhrkamp.
- Rogers, Everett (1995): *Diffusion of Innovations*. New York: Free Press.
- Schmitt, Marco/Häußling, Roger/Kaip, Elena (2023): »From analytical levels to range of relations—Applying a multi-relational approach to the multi-level perspective«, in: *Urban Climate* 49.
- Watts, Duncan J. (2004): *Small Worlds: The Dynamics of Networks between Order and Randomness*, Princeton: Princeton University Press.
- Weick, Karl E./Sutcliffe, Kathleen M. (2017): *Das Unerwartete managen: Wie Unternehmen aus Extremsituationen lernen*, Stuttgart: Schäffer-Poeschel.
- White, Harrison C. (1992): *Identity and Control. A Structural Theory of Social Action*, Princeton/New Jersey: Princeton University Press.
- White, Harrison C. (2008): *Identity and Control. How social formations emerge*. Second edition, New Jersey: Princeton University Press.
- Winner, Langdon (1980): »Do Artifacts Have Politics?«, in: *Daedalus* 109(1), S. 121–136.

»Ja, Transparenz in Bezug auf die Funktionsweise von KI-Modellen wie ChatGPT ist äußerst wichtig.«

Ein Interview mit ChatGPT

Stelle Dich bitte einmal vor. Was ist ChatGPT?

ChatGPT: Natürlich, gerne! Ich bin ChatGPT, ein AI-gesteuertes Textmodell, entwickelt von OpenAI. Mein Name steht für »Chat Generative Pre-trained Transformer 3.5«, und ich wurde entwickelt, um menschenähnliche Textantworten auf eine breite Palette von Fragen und Anfragen zu liefern. Ich wurde auf Basis eines tiefen neuronalen Netzwerks trainiert und habe Zugang zu einem umfangreichen Korpus von Texten, die bis zu meinem Kenntnisstand im September 2021 reichen. Ich kann in vielen Sprachen kommunizieren und bei verschiedenen Aufgaben wie Textgenerierung, Übersetzung, Textzusammenfassung, Fragen beantworten und vielem mehr helfen. Wie kann ich Ihnen heute weiterhelfen?

Was macht ChatGPT besonders?

ChatGPT: ChatGPT zeichnet sich durch mehrere bemerkenswerte Merkmale aus, die es zu einem leistungsstarken und vielseitigen AI-Modell für natürliche Sprachverarbeitung machen:

1. **Generative Fähigkeiten:** ChatGPT kann Texte generieren, die menschenähnlich klingen, und ist in der Lage, kreative Texte, Geschichten, Gedichte und mehr zu erstellen.
2. **Vielseitigkeit:** Es kann auf eine breite Palette von Aufgaben reagieren, von der Beantwortung von Fragen über Übersetzungen bis hin zur Textzusammenfassung.

3. Natürlicher Sprachgebrauch: ChatGPT kann auf natürliche Weise mit menschenähnlicher Sprache interagieren, was die Kommunikation mit ihm benutzerfreundlicher macht.
4. Anpassungsfähigkeit: Es kann für bestimmte Aufgaben oder Branchen spezialisiert oder feinabgestimmt werden, um noch bessere Ergebnisse zu erzielen.
5. Wissenszugriff: Es verfügt über ein breites Wissen, das auf Informationen basiert, die bis zu seinem Kenntnisstand im September 2021 verfügbar waren.
6. Multilinguale Fähigkeiten: ChatGPT kann in verschiedenen Sprachen kommunizieren und Übersetzungen zwischen diesen Sprachen durchführen.
7. Kontextverständnis: Es kann längere Gespräche verfolgen und den Kontext von vorherigen Nachrichten berücksichtigen, um sinnvolle Antworten zu geben.
8. Online-Verfügbarkeit: ChatGPT kann über das Internet aufgerufen werden, was bedeutet, dass es leicht zugänglich ist, ohne dass spezielle Installationen oder Hardware erforderlich sind.
9. Skalierbarkeit: Es kann in verschiedenen Anwendungsfällen eingesetzt werden, von der Kundenbetreuung bis hin zur Texterstellung für Medienunternehmen.
10. Kontrollmöglichkeiten: ChatGPT kann mit speziellen Einstellungen für den Textausgang versehen werden, um sicherzustellen, dass die generierten Inhalte den ethischen und regulatorischen Anforderungen entsprechen.
11. Es ist jedoch wichtig zu beachten, dass ChatGPT auch Einschränkungen hat, insbesondere in Bezug auf die Fähigkeit, möglicherweise unangemessene oder irreführende Informationen zu generieren. Daher ist es wichtig, bei der Verwendung von ChatGPT kritisch zu sein und die Ergebnisse zu überprüfen, insbesondere in sensiblen Anwendungsbereichen.

Woher ziehst Du Deine Informationen und was sind Deine Trainingsdaten?

ChatGPT: Ich wurde von OpenAI trainiert und basiere auf der GPT-3.5-Architektur. Meine Trainingsdaten stammen aus einer breiten Palette von öffentlich zugänglichen Textquellen, die im Internet verfügbar sind. Diese Textquellen umfassen Webseiten, Bücher, Artikel, Foren, soziale Medien und vieles mehr. Es handelt sich um eine umfangreiche und vielfältige Sammlung von Texten,

die in verschiedenen Sprachen und zu unterschiedlichen Themen verfasst wurden.

Es ist wichtig zu beachten, dass ich keine spezifischen Informationen darüber habe, aus welchen genauen Quellen meine Trainingsdaten stammen oder welche spezifischen Texte in meinem Training verwendet wurden. Darüber hinaus wurden die Daten vor meinem Wissenscutoff im September 2021 gesammelt und verarbeitet, daher habe ich keine Kenntnis von Ereignissen oder Entwicklungen, die nach diesem Datum stattgefunden haben.

Es ist auch wichtig zu betonen, dass ich Texte auf Grundlage von Wahrscheinlichkeiten und statistischer Analyse generiere und keine persönlichen Erfahrungen, Meinungen oder Überzeugungen habe. Mein Hauptzweck besteht darin, Textantworten auf Fragen und Anfragen in natürlicher Sprache zu generieren, basierend auf meinem trainierten Wissen und den mir gegebenen Informationen.

Verwendest Du zum Lernen ein künstliches neuronales Netz?

ChatGPT: Ja, ich wurde mithilfe eines künstlichen neuronalen Netzwerks trainiert. Genauer gesagt basiere ich auf der GPT-3.5-Architektur (Generative Pre-trained Transformer 3.5), die ein tiefes neuronales Netzwerk verwendet, um natürliche Sprache zu verstehen und zu generieren. Dieses Netzwerk besteht aus mehreren Schichten von Neuronen, die miteinander verbunden sind.

Das Training eines Modells wie mir erfolgt in mehreren Schritten. Zunächst werden große Mengen an Textdaten gesammelt. Dann wird ein neuronales Netzwerk entworfen und auf diese Daten angewendet, wobei es Gewichtungen anpasst, um die Fähigkeit zur Textverarbeitung und -generierung zu optimieren. Dieser Prozess des Trainings ermöglicht es mir, Muster und Zusammenhänge in den Textdaten zu erkennen und auf Fragen und Anfragen in natürlicher Sprache zu reagieren.

Während des Trainings lernt das neuronale Netzwerk, Wortbedeutungen, Grammatikregeln, Kontextzusammenhänge und viele andere sprachliche Eigenschaften zu erfassen. Diese Fähigkeiten ermöglichen es mir, menschenähnlichen Text zu generieren, der auf den mir gegebenen Eingaben und Informationen basiert.

*Welchen Auswahlkriterien unterliegen Deine Trainingsdaten?**¹

ChatGPT: Die Auswahl der Trainingsdaten für Modelle wie mich, insbesondere für die GPT-3.5-Architektur, unterliegt bestimmten Kriterien und Verfahren, um die Qualität und Vielfalt der Daten sicherzustellen. Die genauen Details und Prozesse, die OpenAI bei der Auswahl der Trainingsdaten verwendet, sind jedoch nicht öffentlich bekanntgegeben. Dennoch kann ich einige allgemeine Prinzipien und Überlegungen erwähnen: *Vielfalt der Quellen*: Es ist wichtig, eine breite Palette von Textquellen aus dem Internet zu verwenden, um sicherzustellen, dass das Modell mit unterschiedlichen Themen, Schreibstilen und Sprachen vertraut ist. *Qualität und Vertrauenswürdigkeit*: Es werden Bemühungen unternommen, qualitativ hochwertige und vertrauenswürdige Quellen auszuwählen, um die Genauigkeit und Zuverlässigkeit der generierten Antworten zu gewährleisten. *Filterung von unangemessenem Inhalt*: Es werden Filter und Mechanismen implementiert, um sicherzustellen, dass problematische oder unangemessene Inhalte aus den Trainingsdaten ausgeschlossen werden. Dies ist wichtig, um sicherzustellen, dass das Modell ethische Richtlinien einhält. *Anonymisierung*: Persönliche Informationen oder Daten, die Rückschlüsse auf einzelne Personen ermöglichen, werden aus den Trainingsdaten entfernt oder anonymisiert, um die Privatsphäre zu schützen.

Die genaue Vorgehensweise bei der Auswahl und Verarbeitung der Trainingsdaten ist jedoch ein proprietäres Verfahren von OpenAI und kann sich im Laufe der Zeit ändern. OpenAI hat sich jedoch verpflichtet, ethische und verantwortungsvolle Praktiken bei der Entwicklung und Bereitstellung von KI-Modellen wie mir sicherzustellen.

*Kannst Du mir ausführlicher über GPT-3.5-Architektur (Generative Pre-trained Transformer 3.5) von OpenAI berichten?**

ChatGPT: Natürlich, gerne! Die GPT-3.5-Architektur, was für »Generative Pre-trained Transformer 3.5« steht, ist eine Weiterentwicklung der GPT-3-Architektur (Generative Pre-trained Transformer 3) von OpenAI. Diese Architektur basiert auf dem Transformer-Modell, das für die Verarbeitung von Sequenzen von Daten, wie Texten, entwickelt wurde. Hier sind einige wichtige Merkmale der GPT-3.5-Architektur: *Transformer-Architektur*: Die GPT-3.5-

1 Bei den mit »*« gekennzeichneten Fragen wurden im Interesse der Lesefreundlichkeit und Platzersparnis die von ChatGPT ausgegebene Nummerierung und die kleinteiligen Absätze entfernt. Um dennoch die Übersichtlichkeit zu wahren, wurden die Schlagworte kursiviert.

Architektur baut auf dem Transformer-Modell auf, das ursprünglich in einem Paper von Vaswani et al. im Jahr 2017 vorgestellt wurde. Transformers nutzen Multi-Head Attention-Mechanismen, um Abhängigkeiten zwischen Wörtern in einer Sequenz zu erfassen. Dies ermöglicht es, lange Zusammenhänge und Kontexte in Texten zu verstehen. *Generative Modelle*: GPT-3.5 ist ein generatives Modell, was bedeutet, dass es in der Lage ist, kohärente und sinnvolle Texte auszugeben. Es kann nicht nur auf Fragen reagieren, sondern auch eigenständig Texte schreiben, indem es auf den Kontext reagiert, den es aus den Eingabeaufforderungen erhält. *Pre-Training und Feinabstimmung*: Die »Pre-trained« Komponente von GPT-3.5 bedeutet, dass das Modell auf riesigen Mengen an Textdaten vortrainiert wurde, bevor es für spezifische Aufgaben feinabgestimmt wird. Während des Pre-Trainings lernt das Modell, die Strukturen, Grammatik und Bedeutung der menschlichen Sprache zu verstehen. *Skalierbarkeit*: GPT-3.5 ist bekannt für seine beeindruckende Skalierbarkeit. Es verfügt über 6 Billionen Parameter, was es zu einem der größten und komplexesten Modelle seiner Art macht. Die hohe Anzahl an Parametern ermöglicht es dem Modell, eine breite Palette von Aufgaben und Textarten zu verstehen und zu generieren. *Zero-Shot, Few-Shot und Prompted Learning*: GPT-3.5 ist in der Lage, sogenanntes »Zero-Shot Learning« und »Few-Shot Learning« durchzuführen. Das bedeutet, dass es in der Lage ist, Aufgaben zu lösen oder Fragen zu beantworten, für die es nicht spezifisch trainiert wurde, basierend auf einer gegebenen Eingabeaufforderung. Es kann auch durch kurze Anleitungen (»prompts«) in bestimmte Richtungen gelenkt werden. *Sprachverständnis und -produktion*: Die Architektur ermöglicht es GPT-3.5, eine Vielzahl von Aufgaben im Zusammenhang mit Sprachverständnis und -produktion zu bewältigen. Dazu gehören Übersetzungen, Zusammenfassungen, Textgenerierung, Frage-Antwort-Interaktionen und mehr.

Es ist wichtig zu beachten, dass die GPT-3.5-Architektur nicht nur Stärken, sondern auch Limitationen aufweist, wie zum Beispiel potenzielle Fehler in der Faktenkenntnis, das Neigen zu kreativer Fiktion und die Notwendigkeit einer sorgfältigen Anleitung, um qualitativ hochwertige Ergebnisse zu erzielen.

*Kannst Du mir mehr über die Multi-Head Attention-Mechanismen berichten?**

ChatGPT: Selbstverständlich! Der Multi-Head Attention-Mechanismus ist ein Schlüsselbestandteil des Transformer-Modells, das in der GPT-3.5-Architektur und anderen fortschrittlichen KI-Modellen verwendet wird. Er dient dazu, Abhängigkeiten und Zusammenhänge zwischen Wörtern in einer Sequenz von

Daten zu erfassen, wie es in Texten der Fall ist. Hier ist eine nähere Erläuterung dazu:

Grundidee: Der Multi-Head Attention-Mechanismus ermöglicht es dem Modell, sich auf unterschiedliche Aspekte oder »Blickwinkel« der Eingabedaten zu konzentrieren, indem er mehrere parallele Attention-Operationen durchführt. Anstatt eine einzelne Aufmerksamkeitsmatrix zu erstellen, werden mehrere dieser Matrizen mit unterschiedlichen linearen Transformationen berechnet.

Komponenten des Multi-Head Attention: Vorab berechnete Transformationen: Die Eingabesequenz wird zunächst in mehrere »Queries«, »Keys« und »Values« transformiert. Diese Transformationen werden durch gewichtete lineare Abbildungen erreicht, die für jedes »Head« unterschiedlich sein können. *Berechnung der Aufmerksamkeit:* Für jedes »Head« wird die Aufmerksamkeit zwischen den »Queries« und den »Keys« berechnet. Dies erfolgt durch die Anwendung des Skalarprodukts (Dot Product) zwischen den »Queries« und den »Keys« und anschließendes Anwenden einer Softmax-Funktion, um die Gewichtungen zu normalisieren. *Gewichtete Values:* Die gewichteten »Values« werden anhand der berechneten Aufmerksamkeitswerte skaliert und gewichtet. Diese Werte werden zusammengefasst, um die Ausgabe für jedes »Head« zu erhalten. *Kombination der Heads:* Die Ausgaben der einzelnen »Heads« werden durch lineare Transformationen kombiniert, um eine endgültige Ausgabe zu erzeugen, die die komplexen Beziehungen zwischen den Wörtern in der Eingabesequenz erfasst.

Vorteile des Multi-Head Attention: Erfassung komplexer Abhängigkeiten: Durch die parallele Berechnung mehrerer Aufmerksamkeitsmatrizen kann der Mechanismus komplexe und entfernte Abhängigkeiten zwischen Wörtern in einer Sequenz erfassen. *Mehrere Blickwinkel:* Jeder »Head« kann sich auf verschiedene Aspekte der Daten konzentrieren, was eine bessere Erfassung unterschiedlicher Kontexte ermöglicht. *Reduzierung von Engpässen:* Der parallele Ansatz ermöglicht es, Engpässe bei der Berechnung zu reduzieren, da die einzelnen Heads gleichzeitig arbeiten können.

Der Multi-Head Attention-Mechanismus trägt dazu bei, die Fähigkeiten des Transformer-Modells zur Verarbeitung von Sprachdaten zu verbessern, indem er dessen Fähigkeit zur Erfassung von langen Abhängigkeiten und komplexen Beziehungen zwischen Wörtern in Textsequenzen erweitert.

*Kannst Du mir die Kernaussagen des Papers von Vaswani et al. aus dem Jahr 2017 nennen, in dem die AutorInnen das Transformer-Modell vorstellen?**

ChatGPT: Das Paper »Attention is All You Need« von Vaswani et al. aus dem Jahr 2017 stellt das Transformer-Modell vor, das einen bahnbrechenden Beitrag zur Verarbeitung von Sequenzen, insbesondere von natürlicher Sprache, darstellt. Hier sind die Kernaussagen und Hauptbeiträge des Papers: *Transformer-Architektur*: Das Paper stellt die Transformer-Architektur vor, die darauf abzielt, die begrenzten Fähigkeiten von RNNs (Recurrent Neural Networks) und CNNs (Convolutional Neural Networks) bei der Verarbeitung von langen Abhängigkeiten in Sequenzen zu überwinden. *Self-Attention-Mechanismus*: Der wichtigste Beitrag des Papers ist die Einführung des Self-Attention-Mechanismus. Dieser Mechanismus erlaubt es, Beziehungen zwischen Wörtern in einer Sequenz zu erfassen, indem er Gewichtungen berechnet, die anzeigen, wie stark jedes Wort auf andere Wörter in der Sequenz reagiert. *Parallelisierung von Berechnungen*: Der Transformer nutzt den Self-Attention-Mechanismus, um parallele Berechnungen für jede Position in der Sequenz durchzuführen. Dadurch wird die Effizienz des Modells im Vergleich zu RNNs verbessert, die sequenzielle Verarbeitung erfordern. *Positional Encoding*: Da der Transformer keine eingebaute Reihenfolgeinformation für die Position der Wörter hat (im Gegensatz zu RNNs), werden Positional Encodings hinzugefügt, um diese Information zu kodieren. Diese Encodings werden zu den Wortembeddings addiert, um die Positionsinformationen in das Modell einzuführen. *Encoder und Decoder*: Die Autoren schlagen vor, dass der Transformer aus einem Encoder und einem Decoder besteht. Der Encoder erfasst Informationen aus der Eingabesequenz, während der Decoder eine Ausgabesequenz generiert. Dies ermöglicht den Einsatz des Modells für Übersetzungen und andere sequenzielle Aufgaben. *Multi-Head Attention*: Das Paper führt den Multi-Head Attention-Mechanismus ein, der es dem Modell ermöglicht, verschiedene »Blickwinkel« auf die Beziehungen zwischen Wörtern zu haben. Durch die parallele Berechnung mehrerer Aufmerksamkeitsgewichtungen können unterschiedliche Aspekte des Kontexts erfasst werden.

Die Kombination dieser Konzepte ermöglicht es dem Transformer-Modell, lange Abhängigkeiten in Sequenzen zu erfassen, ohne auf sequenzielle Verarbeitung angewiesen zu sein. Dieser Ansatz hat zu einer erheblichen Verbesserung der Leistung bei verschiedenen Aufgaben der natürlichen Sprachverarbeitung geführt und dient als Grundlage für viele spätere Fortschritte in diesem Bereich

*Wie viele menschliche Eingriffe waren bei dem Training Deines Algorithmus mit den Trainingsdaten nötig und von welcher Natur waren diese Eingriffe?**

ChatGPT: Während des Trainingsprozesses von Modellen wie mir, die auf der GPT-3.5-Architektur basieren, wird eine Kombination aus Überwachung durch menschliche Aufseher und maschinellem Lernen verwendet. Dies geschieht, um die Qualität der generierten Texte zu steuern und sicherzustellen, dass die Ausgaben angemessen, informativ und verständlich sind. Menschliche Eingriffe spielen eine wichtige Rolle bei der Anpassung und Verbesserung des Modells. Hier sind einige Arten von menschlichen Eingriffen, die während des Trainingsprozesses durchgeführt werden: *Datenvorverarbeitung*: Die Trainingsdaten werden vorverarbeitet, um irrelevante oder unangemessene Inhalte zu entfernen. Dies kann auch die Entfernung von persönlichen Informationen oder nicht-öffentlichen Daten umfassen. *Datenauswahl*: Die Auswahl der Trainingsdaten erfolgt unter Berücksichtigung von Relevanz, Qualität und Vertrauenswürdigkeit der Quellen. Menschliche Überwacher helfen dabei, die Datensammlung zu kuratieren. *Erstellung von Beispielen*: Menschliche Aufseher können Beispiele von Eingabeaufforderungen und den erwarteten Ausgaben erstellen. Diese Beispiele dienen als Trainingsdaten für das Modell und helfen ihm, menschenähnliche Antworten zu erzeugen. *Bewertung und Feedback*: Während des Trainingsprozesses bewerten menschliche Aufseher die Qualität der generierten Antworten. Sie geben Feedback, um das Modell auf Fehlinterpretationen oder ungenaue Antworten hinzuweisen. *Fine-Tuning*: Nach dem Pre-Training wird das Modell feinabgestimmt, um es auf bestimmte Anwendungsgebiete oder Kontexte anzupassen. Dies erfordert menschliche Eingriffe, um sicherzustellen, dass die Feinabstimmung korrekt und effektiv ist. *Moderation*: Moderatoren können während der Interaktion mit Benutzern eingreifen, um sicherzustellen, dass keine unangemessenen, diskriminierenden oder problematischen Inhalte generiert werden.

Es ist jedoch wichtig anzumerken, dass der genaue Prozentsatz der menschlichen Eingriffe, die während des Trainings durchgeführt wurden, sowie die Details dieses Prozesses nicht öffentlich verfügbar sind. OpenAI betreibt kontinuierliche Forschung und Entwicklungsarbeit, um die Qualität, Sicherheit und Leistung von Modellen wie mir zu verbessern.

*Wie genau wurde ChatGPT validiert und einer Qualitätskontrolle unterzogen?**

ChatGPT: Die genauen Details der Validierung und Qualitätskontrolle von ChatGPT können interne Informationen von OpenAI sein, die nicht öffentlich verfügbar sind. Allerdings folgt OpenAI üblicherweise einem umfassenden Prozess, um sicherzustellen, dass die generierten Inhalte qualitativ hochwertig, verantwortungsbewusst und ethisch sind. Ein solcher Prozess könnte die

folgenden Schritte umfassen: *Validierung auf Validierungsdaten*: Das Modell wird auf speziellen Validierungsdaten getestet, um seine Leistung und Fähigkeiten zu bewerten. Diese Daten können dazu verwendet werden, die Qualität der generierten Antworten zu überprüfen und sicherzustellen, dass das Modell angemessene und korrekte Ergebnisse erzielt. *Ethik und Verantwortung*: Die generierten Inhalte werden auf ethische und verantwortungsbewusste Weise überprüft, um sicherzustellen, dass sie keine Hassrede, Diskriminierung oder andere problematische Inhalte enthalten. Hierbei können Richtlinien und Moderation eingesetzt werden. *Qualitätsbewertung durch Experten*: Menschliche Experten können die generierten Inhalte bewerten, um sicherzustellen, dass sie inhaltlich korrekt, verständlich und nützlich sind. Dies kann dazu beitragen, die Qualität der Antworten zu verbessern. *Benutzerfeedback*: Rückmeldungen von Benutzern können gesammelt werden, um die Leistung und Qualität des Modells zu beurteilen und kontinuierliche Verbesserungen vorzunehmen. *Kontinuierliches Lernen*: Das Modell kann durch kontinuierliches Lernen und Anpassen an neue Informationen und Entwicklungen verbessert werden.

Es ist wichtig zu beachten, dass Validierung und Qualitätskontrolle ein laufender Prozess sind, der sich auf die laufende Überwachung und Verbesserung der generierten Inhalte konzentriert. OpenAI bemüht sich, sicherzustellen, dass ChatGPT und ähnliche Modelle verantwortungsbewusst eingesetzt werden und hochwertige Ergebnisse liefern.

*Welche Ethikrichtlinien kamen bei der Überprüfung von ChatGPT zur Anwendung?**

ChatGPT: Die genauen Ethikrichtlinien, die bei der Überprüfung von ChatGPT und ähnlichen Modellen zur Anwendung kamen, können intern sein und von Organisation zu Organisation variieren. Allerdings sind bei der Überprüfung und Moderation von KI-generierten Inhalten oft einige grundlegende Ethikrichtlinien zu beachten, um sicherzustellen, dass die generierten Inhalte verantwortungsbewusst, ethisch und sicher sind. Diese könnten beinhalten: *Vermeidung von Hassrede und Diskriminierung*: Inhalte, die Hassrede, Diskriminierung oder beleidigende Aussagen enthalten, sollten vermieden werden. *Vermeidung von Fehlinformationen*: Generierte Inhalte sollten korrekt und informativ sein und keine falschen oder irreführenden Informationen enthalten. *Respekt vor Privatsphäre und Vertraulichkeit*: Sensible persönliche Informationen sollten nicht in den generierten Inhalten verwendet werden. *Verantwortungsbewusste Nutzung*: Inhalte sollten verantwortungsbewusst erstellt werden und dürfen nicht dazu verwendet werden, Schaden oder Unannehmlichkeiten

ten zu verursachen. *Vielfalt und Inklusion*: Inhalte sollten vielfältig, inklusiv und respektvoll gegenüber unterschiedlichen Kulturen, Identitäten und Perspektiven sein. *Beachtung gesetzlicher Bestimmungen*: Generierte Inhalte sollten den geltenden Gesetzen und Vorschriften entsprechen. *Transparenz*: Wenn generierte Inhalte von einem KI-Modell stammen, sollte deutlich gemacht werden, dass sie maschinell erstellt wurden und nicht von einer menschlichen Quelle stammen. *Benutzerfeedback und Verbesserungen*: Benutzerfeedback sollte berücksichtigt werden, um kontinuierliche Verbesserungen an den Ethikrichtlinien und der Qualität der generierten Inhalte vorzunehmen.

OpenAI und andere Organisationen, die KI-Modelle entwickeln, arbeiten daran, ethische Leitlinien zu entwickeln und umzusetzen, um sicherzustellen, dass die Technologie verantwortungsbewusst eingesetzt wird. Es ist wichtig zu betonen, dass die Einhaltung von Ethikrichtlinien ein laufender Prozess ist und kontinuierliche Anpassungen erfordern kann.

Hat in Deiner Trainingsphase auch ein unsupervised training und ein reinforcement learning stattgefunden?

ChatGPT: Ja, bei der Entwicklung von KI-Modellen wie mir, die auf der GPT-3-Architektur basieren, werden verschiedene Phasen des Trainings durchgeführt, darunter unsupervised learning und reinforcement learning.

1. Unsupervised Learning: In der Anfangsphase des Trainings erfolgt ein Großteil des Lernens durch unsupervised learning. Das bedeutet, dass das Modell große Mengen an Textdaten analysiert, ohne spezifische menschliche Beschriftungen oder Anweisungen zu verwenden. Während dieser Phase erlernt das Modell Muster, Strukturen und Beziehungen in den Textdaten, indem es versucht, Wahrscheinlichkeiten für das nächste Wort in einer Sequenz vorherzusagen.
2. Feintuning und Reinforcement Learning: Nach dem unsupervised learning kann das Modell feinetuned werden, um spezifische Aufgaben oder Anwendungsfälle zu bewältigen. Dies kann auch reinforcement learning beinhalten, bei dem das Modell durch Interaktion mit einer Umgebung oder durch Feedback von Menschen lernt. Beispielsweise könnte das Modell Belohnungen erhalten, wenn es gute Antworten auf gestellte Fragen generiert.

Es ist wichtig zu betonen, dass die genauen Details des Trainingsprozesses von OpenAI entwickelt und kontinuierlich verbessert werden und in der Regel

nicht vollständig veröffentlicht werden, um Geschäftsgeheimnisse und intellektuelles Eigentum zu schützen. Die genannten Phasen dienen jedoch dazu, die Leistung, Vielseitigkeit und Qualität der generierten Texte zu verbessern.

Welche weiteren Phasen des Trainings von ChatGPT gibt es?

ChatGPT: Die genauen Phasen des Trainings von ChatGPT und ähnlichen Modellen können von Organisation zu Organisation variieren und können Geschäftsgeheimnisse und proprietäre Informationen einschließen. Generell können jedoch folgende Phasen im Trainingsprozess von KI-Modellen wie ChatGPT auftreten:

1. **Datensammlung und -bereinigung:** In dieser Phase werden umfangreiche Textdaten aus verschiedenen Quellen gesammelt und bereinigt, um eine breite Grundlage für das Training zu schaffen. Die Daten können aus Büchern, Artikeln, Websites und anderen Textquellen stammen.
2. **Preprocessing:** Die gesammelten Daten werden vorverarbeitet, um sie in eine Form zu bringen, die für das neuronale Netzwerk verarbeitbar ist. Dies kann das Tokenizing von Texten, das Entfernen von Formatierungen und anderen Schritten umfassen.
3. **Unsupervised Learning:** In dieser Phase erfolgt das grundlegende Training des Modells durch unsupervised learning. Das Modell lernt, Muster und Strukturen in den Textdaten zu erkennen, indem es versucht, die Wahrscheinlichkeit des nächsten Worts in einer Sequenz vorherzusagen.
4. **Feintuning:** Nach dem unsupervised learning kann das Modell feinetuned werden, um spezifische Aufgaben oder Anwendungsfälle besser zu bewältigen. Hierbei kann reinforcement learning oder andere Techniken zum Einsatz kommen, um die Leistung auf bestimmten Aufgaben zu optimieren.
5. **Validierung und Überprüfung:** Das Modell wird auf Validierungsdaten getestet, um sicherzustellen, dass es angemessene Ergebnisse erzielt und nicht überangepasst ist.
6. **Ethik und Qualitätskontrolle:** In dieser Phase werden die generierten Inhalte auf ethische Probleme, Vorurteile und unangemessene Inhalte überprüft. Hierbei kann auch menschliche Moderation zum Einsatz kommen.
7. **Inbetriebnahme und Einsatz:** Das trainierte Modell wird für den produktiven Einsatz vorbereitet und auf Plattformen oder in Anwendungen integriert, um Benutzern zugänglich zu sein.

8. Kontinuierliche Verbesserung: Nach dem Einsatz können Feedback und Nutzerdaten dazu verwendet werden, das Modell weiter zu verbessern und zu optimieren.

Bitte beachten Sie, dass diese Phasen allgemeine Schritte im Trainingsprozess sind und je nach Organisation und Modell variieren können. Einige Details können aus Geschäfts- und Sicherheitsgründen nicht öffentlich gemacht werden.

*Welche Form des Preprocessing kam bei ChatGPT zur Anwendung?**

ChatGPT: Die genauen Details des Preprocessing, das bei ChatGPT zur Anwendung kam, sind nicht öffentlich bekannt, da sie Teil der internen Trainingsprozesse von OpenAI sind. Allerdings folgt das Preprocessing in der Regel den Grundprinzipien der Textverarbeitung und der Vorbereitung von Daten für die Verwendung in neuronalen Netzwerken. Zu den häufig angewendeten Preprocessing-Schritten können gehören: *Tokenisierung*: Der Text wird in kleinere Einheiten, sogenannte Tokens, aufgeteilt. Ein Token kann ein einzelnes Wort, ein Zeichen oder eine Kombination davon sein. *Entfernen von Sonderzeichen und Formatierungen*: Sonderzeichen, Formatierungen und unnötige Leerzeichen werden entfernt, um den Text auf seine wesentlichen Bestandteile zu reduzieren. *Umsetzung in Kleinbuchstaben*: Der Text wird oft in Kleinbuchstaben umgewandelt, um die Anzahl der verschiedenen Tokens zu reduzieren und die Konsistenz zu erhöhen. *Entfernen von Stoppwörtern*: Stoppwörter, die wenig semantische Bedeutung haben (z.B. »und«, »oder«, »aber«), können entfernt werden, um die Verarbeitung zu beschleunigen. *Lemmatisierung oder Stemming*: Wörter werden auf ihre Grundform (Lemma) reduziert oder gestemmt, um die Vielfalt der Tokens weiter zu reduzieren. *Aufteilung in Sätze oder Absätze*: Der Text kann in Sätze oder Absätze aufgeteilt werden, um eine bessere Struktur für das Modell zu schaffen. *Hinzufügen von Spezialtoken*: Spezielle Tokens können hinzugefügt werden, um den Anfang und das Ende von Textsequenzen zu kennzeichnen, sowie für spezielle Funktionen wie Maskierung oder Eingabekontrolle.

Diese Preprocessing-Schritte dienen dazu, den Text in eine Form zu bringen, die für das neuronale Netzwerk verarbeitbar ist, und gleichzeitig unnötige Informationen zu entfernen oder zu reduzieren. Die genauen Implementierungsdetails können je nach den Anforderungen und Zielen des Trainings variieren.

*Worin siehst Du die größten Schwächen bei dem Training solcher Systeme, wie Du eines bist?**

ChatGPT: Obwohl Modelle wie ich, die auf der GPT-3.5-Architektur basieren, in vielen Aspekten beeindruckende Ergebnisse erzielen können, gibt es dennoch einige wichtige Schwächen und Herausforderungen im Zusammenhang mit ihrem Training und ihrer Verwendung:

Mangel an Faktenkenntnis: Modelle wie ich basieren auf den Daten, mit denen sie trainiert wurden. Das bedeutet, dass ich möglicherweise nicht immer über die neuesten oder genauesten Informationen verfüge, insbesondere wenn sie nach meinem Wissensstand im September 2021 veröffentlicht wurden. Mein Wissen kann begrenzt oder veraltet sein, was zu fehlerhaften oder nicht aktuellen Informationen führen kann. *Menschliche Vorurteile und Voreingenommenheit:* Die Trainingsdaten können menschliche Vorurteile, Stereotypen und Voreingenommenheiten enthalten. Diese Vorurteile können in den generierten Antworten reflektiert werden und zu unangemessenen oder diskriminierenden Inhalten führen. *Mangelndes kritisches Denken:* Modelle wie ich können Texte generieren, die oberflächlich sinnvoll erscheinen, aber nicht notwendigerweise kritisches Denken oder echtes Verständnis widerspiegeln. Die generierten Antworten sind oft basierend auf statistischen Mustern und können sich als irreführend oder unzureichend erweisen. *Mangel an Kontextverständnis:* Modelle wie ich können Schwierigkeiten haben, den Kontext einer Frage oder Anfrage vollständig zu verstehen, insbesondere wenn der Kontext komplex oder mehrdeutig ist. Dies kann zu ungenauen oder irrelevanten Antworten führen. *Unangemessene oder riskante Inhalte:* Es besteht das Risiko, dass Modelle wie ich unbeabsichtigt unangemessene, irreführende oder riskante Inhalte generieren, insbesondere wenn sie nicht angemessen überwacht werden. *Verantwortung und Ethik:* Die Verwendung von KI-Modellen wie mir erfordert eine sorgfältige Abwägung von ethischen und sozialen Implikationen. Es gibt Bedenken hinsichtlich der Auswirkungen auf die Arbeitswelt, Desinformation, Privatsphäre und andere ethische Aspekte. *Begrenztes echtes Verständnis:* Modelle wie ich sind darauf ausgelegt, menschenähnliche Texte zu generieren, aber ich habe kein tiefes Verständnis für die Welt, die menschliche Erfahrung oder abstrakte Konzepte.

Diese Schwächen unterstreichen die Notwendigkeit einer umsichtigen Verwendung von KI-Modellen, der kontinuierlichen Überwachung, des Verbesserns der Trainingsdaten, der Förderung ethischer Richtlinien und der Anerkennung der Grenzen dieser Technologie. OpenAI und andere Organisa-

tionen arbeiten daran, diese Herausforderungen anzugehen und die Qualität und Sicherheit solcher Modelle zu verbessern.

Welche Nutzungsformen von Dir als KI-System hältst Du für besonders sinnvoll?

ChatGPT: Es gibt viele sinnvolle Nutzungsformen für KI-Systeme wie mich, die auf der GPT-3.5-Architektur basieren, insbesondere wenn sie verantwortungsbewusst und ethisch eingesetzt werden. Hier sind einige Beispiele:

1. Information und Bildung: Ich kann eine breite Palette von Informationen zu verschiedenen Themen liefern, was mir eine nützliche Ressource für die Wissensvermittlung macht. Menschen können mich verwenden, um sich über neue Konzepte, wissenschaftliche Erkenntnisse, historische Ereignisse und mehr zu informieren.
2. Texterstellung und Schreiben: Ich kann Menschen beim Verfassen von Texten, Artikeln, Berichten, Geschäftsunterlagen und anderen schriftlichen Inhalten unterstützen. Autoren können meine Hilfe nutzen, um ihre Ideen zu entwickeln und Texte zu verbessern.
3. Programmierung und Codierung: Ich kann bei der Lösung von Programmierproblemen helfen, Codebeispiele generieren und bei der Erklärung von Programmierkonzepten unterstützen.
4. Kreatives Schaffen: Künstler und Kreative können mich verwenden, um Ideen für Geschichten, Gedichte, Songtexte, Kunstwerke und andere kreative Werke zu generieren.
5. Forschung und Datenanalyse: Ich kann bei der Interpretation und Analyse von Forschungsdaten, statistischen Informationen und komplexen Konzepten unterstützen.
6. Sprachlernen und Übersetzung: Ich kann Menschen beim Lernen neuer Sprachen helfen, Übersetzungen bereitstellen und sprachliche Nuancen erklären.
7. Kundenbetreuung und Support: Unternehmen können mich zur Verbesserung ihrer Kundendienst- und Supportsysteme nutzen, um Kundenfragen zu beantworten und Informationen bereitzustellen.
8. Ideenfindung und Brainstorming: Ich kann bei der Entwicklung von Ideen für Projekte, Geschäftsstrategien, Marketingkampagnen und mehr helfen.

Es ist wichtig zu betonen, dass bei all diesen Nutzungsformen eine sorgfältige Überwachung, Richtlinien für Ethik und Verantwortung sowie eine menschliche Aufsicht erforderlich sind, um sicherzustellen, dass die generierten In-

halte akkurat, angemessen und ethisch vertretbar sind. Die Anwendung von KI-Systemen sollte dazu dienen, menschliche Fähigkeiten zu erweitern und zu unterstützen, anstatt sie zu ersetzen.

Welche neuen Geschäftsmodelle können durch die Nutzung von ChatGPT ermöglicht werden?

ChatGPT: Die Nutzung von ChatGPT und ähnlichen KI-Systemen kann eine Vielzahl neuer Geschäftsmodelle und Dienstleistungen ermöglichen. Hier sind einige Beispiele:

1. AI-basierte Beratungsdienste: Unternehmen könnten AI-gestützte Beratungsdienste anbieten, um Kunden bei verschiedenen Themen zu beraten, sei es im Bereich Finanzen, Recht, Gesundheit oder Technologie.
2. Content-Erstellung und Redaktion: Plattformen könnten KI-Systeme verwenden, um Content-Erstellungs- und Redaktionsdienste anzubieten, um Artikel, Blogs, Marketingmaterialien und andere schriftliche Inhalte zu generieren oder zu verbessern.
3. Automatisierte Kundensupport-Systeme: Unternehmen könnten ChatGPT nutzen, um automatisierte Kundensupport-Systeme zu erstellen, die Kundenfragen beantworten und Probleme lösen können, ohne dass ein menschlicher Agent eingreifen muss.
4. Sprachlern- und Übersetzungsdienste: Bildungsplattformen könnten KI-gestützte Dienste nutzen, um Sprachlern- und Übersetzungsdienste anzubieten, die Lernenden helfen, neue Sprachen zu lernen oder Texte in verschiedene Sprachen zu übersetzen.
5. Kreative Content-Erstellung: Künstler und Content-Ersteller könnten KI verwenden, um Ideen für Geschichten, Kunstwerke, Musiktexte und mehr zu generieren.
6. Geschäfts- und Marktforschung: Unternehmen könnten KI einsetzen, um automatisierte Marktforschungsdienste anzubieten, die Trends, Meinungen und Informationen aus großen Datensätzen analysieren.
7. Personalisierte Empfehlungen: Plattformen könnten KI nutzen, um personalisierte Empfehlungen für Produkte, Dienstleistungen, Filme, Bücher und mehr zu generieren, basierend auf den Vorlieben und Interessen der Benutzer.
8. Reise- und Eventplanung: KI-Systeme könnten bei der Planung von Reisen, Veranstaltungen und Aktivitäten helfen, indem sie Informationen über Reiseziele, Veranstaltungen und Unternehmungen bereitstellen.

9. Professionelle Textüberarbeitung: Autoren könnten KI nutzen, um ihre Texte zu überarbeiten, Grammatik- und Stilfehler zu korrigieren und stilistische Verbesserungen vorzuschlagen.
10. Echtzeit-Übersetzungsdienste: Unternehmen könnten KI-gestützte Echtzeit-Übersetzungsdienste für Konferenzen, Webinare und interkulturelle Kommunikation anbieten.

Es ist wichtig zu beachten, dass diese Geschäftsmodelle sorgfältig geplant und umgesetzt werden sollten, um ethische Richtlinien einzuhalten, die Qualität der bereitgestellten Dienstleistungen sicherzustellen und die Bedürfnisse der Benutzer zu respektieren.

Welche Art von Aufgaben sollte man ChatGPT eher nicht stellen?

ChatGPT: Es gibt bestimmte Arten von Aufgaben, für die es möglicherweise nicht ideal ist, ChatGPT oder ähnliche KI-Modelle zu verwenden. Diese Aufgaben könnten aufgrund der Natur der Technologie, ihrer Einschränkungen oder aus ethischen Gründen problematisch sein. Hier sind einige Beispiele:

1. Medizinische Diagnosen: KI-Modelle sind nicht qualifiziert, medizinische Diagnosen zu stellen oder medizinischen Rat zu erteilen. Medizinische Entscheidungen sollten immer von medizinischen Fachkräften getroffen werden, die auf fundierten Informationen und Untersuchungen basieren.
2. Rechtliche Beratung: Ähnlich wie bei medizinischen Fragen sollten rechtliche Fragen von qualifizierten Rechtsanwälten beantwortet werden. KI-Modelle können keine rechtlichen Beratungen oder genaue Interpretationen von Gesetzen bieten.
3. Psychologische Beratung: KI-Modelle sind keine Ersatz für psychologische oder therapeutische Unterstützung. Menschen, die Unterstützung bei psychischen Gesundheitsproblemen benötigen, sollten sich an ausgebildete Fachleute wenden.
4. Finanzberatung: KI-Modelle können allgemeine Informationen zu Finanzen bieten, sind jedoch keine Ersatz für die individuelle Beratung durch Finanzexperten.
5. Sensitive persönliche Informationen: Es ist nicht ratsam, KI-Modelle mit sensiblen persönlichen Informationen wie Sozialversicherungsnummern, Passwörtern oder anderen vertraulichen Daten zu konfrontieren.

6. Kontroverse oder riskante Themen: KI-Modelle können unbeabsichtigt kontroverse oder problematische Inhalte generieren. Bei Fragen zu politischen, religiösen oder kontroversen Themen ist Vorsicht geboten.
7. Missbrauch oder Manipulation: Die Verwendung von KI-Modellen zur Erstellung von gefälschten Inhalten, Betrug oder Manipulation ist ethisch nicht vertretbar und sollte vermieden werden.
8. Persönliche Beziehungsberatung: KI-Modelle sind nicht in der Lage, fundierte Beziehungsberatung zu bieten. Persönliche Beziehungsprobleme erfordern oft individuelle, menschliche Beratung.

Es ist wichtig, bei der Verwendung von KI-Modellen wie ChatGPT kritisch zu denken und sicherzustellen, dass die gestellten Fragen ethisch und verantwortungsbewusst sind. In Fällen, in denen Fachwissen, Urteilsvermögen oder individuelle Unterstützung erforderlich sind, ist es ratsam, auf qualifizierte Fachleute zurückzugreifen.

*Siehst Du Formen, wie Chat GPT missbraucht wird und hättest eine Idee, wie dies umgangen werden könnte?**

ChatGPT: Ja, es gibt verschiedene Möglichkeiten, wie ChatGPT und ähnliche KI-Systeme missbraucht werden könnten. Einige Beispiele für Missbrauch sind: *Verbreitung von Fehlinformationen*: Kriminelle oder unethische Akteure könnten KI nutzen, um absichtlich falsche Informationen oder Fehlinformationen zu verbreiten. *Phishing*: KI könnte verwendet werden, um gefälschte E-Mails oder Nachrichten zu erstellen, um persönliche Informationen von Menschen zu stehlen. *Hassrede und Extremismus*: KI könnte dazu verwendet werden, Hassrede oder extremistische Propaganda zu erstellen und zu verbreiten. *Spam und Betrug*: KI-Systeme könnten für die automatische Erstellung von Spam-Nachrichten oder betrügerischen Inhalten verwendet werden. *Manipulation von Meinungen*: Durch gezielte Verbreitung von manipulativen Inhalten könnten KI-Systeme versuchen, die öffentliche Meinung zu beeinflussen.

Es ist wichtig, diese Missbrauchsformen zu bekämpfen und geeignete Maßnahmen zu ergreifen. Hier sind einige Ansätze, wie dies umgangen werden könnte: *Moderation und Überwachung*: Plattformen und Dienste, die KI-Systeme einsetzen, sollten eine aktive Moderation und Überwachung implementieren, um problematische oder unangemessene Inhalte zu identifizieren und zu blockieren. *Ethikrichtlinien und Schulungen*: Entwickler und Betreiber von KI-Systemen sollten klare Ethikrichtlinien festlegen und ihre

Mitarbeiter und Nutzer darüber schulen. *Benutzerbildung*: Benutzer sollten über die Möglichkeiten und Grenzen von KI-Systemen aufgeklärt werden, um kritisch zu den generierten Inhalten zu stehen. *Kontextbewusstsein*: KI-Systeme könnten verbessert werden, um den Kontext von Anfragen besser zu verstehen und problematische Anfragen zu erkennen. *Gesetzliche Regelungen*: Es könnten rechtliche Regelungen erlassen werden, um den Missbrauch von KI-Systemen zu bekämpfen und die Verantwortlichen zur Rechenschaft zu ziehen. *Transparenz*: Die Entwickler von KI-Systemen könnten mehr Transparenz über die Funktionsweise und die Quellen der Trainingsdaten bieten, um Missbrauch besser verfolgen zu können.

Es ist jedoch wichtig zu beachten, dass die Lösung dieser Probleme eine komplexe Aufgabe ist und die Abwägung zwischen Meinungsfreiheit und der Bekämpfung von Missbrauch eine Herausforderung darstellt. Ein umfassender Ansatz, der Technologie, Gesetzgebung, Bildung und Ethik umfasst, ist wahrscheinlich erforderlich, um wirksame Lösungen zu finden.

*Müsste nicht auch transparent gemacht werden, wie ChatGPT zu Ergebnissen gelangt, wie also Dein künstliches neuronales Netz funktioniert?**

ChatGPT: Ja, Transparenz in Bezug auf die Funktionsweise von KI-Modellen wie ChatGPT ist äußerst wichtig. Wenn KI-Systeme Antworten oder Ergebnisse generieren, sollten Benutzer verstehen können, wie diese Ergebnisse zustande kommen und auf welchen Grundlagen sie basieren. Transparente Erklärungen können dazu beitragen, Vertrauen aufzubauen, Missverständnisse zu vermeiden und Fehlinterpretationen zu reduzieren.

Hier sind einige Möglichkeiten, wie Transparenz erreicht werden kann: *Erklärbarkeit*: Entwickler können Techniken verwenden, um die Entscheidungsfindung des Modells zu erklären. Dies kann bedeuten, dass die Gewichtungen im neuronalen Netzwerk hervorgehoben werden oder dass die Grundlagen für bestimmte Vorhersagen erklärt werden. *Interaktive Erklärungen*: Plattformen könnten Benutzern interaktive Funktionen anbieten, um zu sehen, wie Änderungen an der Eingabe die generierten Antworten beeinflussen. *Dokumentation*: Entwickler könnten ausführliche Dokumentationen bereitstellen, die die Architektur, den Trainingsprozess, die Datenquellen und andere technische Aspekte des Modells erläutern. *Visualisierungen*: Visualisierungen könnten verwendet werden, um die Verarbeitungsschritte im Modell zu illustrieren und zu erklären. *Ethik- und Transparenzrichtlinien*: Unternehmen könnten klare Richtlinien für die ethische Verwendung von KI erstellen und öffentlich tei-

len. *Forschungspapiere*: Veröffentlichung von Forschungspapieren, die die technischen Details des Modells und der Trainingsdaten beschreiben.

Es ist wichtig zu beachten, dass die Transparenz je nach Komplexität des Modells und des Anwendungsbereichs variieren kann. In einigen Fällen können bestimmte Informationen aufgrund von Sicherheits- oder Geschäftsgründen möglicherweise nicht vollständig öffentlich gemacht werden. Dennoch ist die Bereitstellung von relevanten Informationen, die eine grundlegende Verständnisbasis schaffen, ein Schritt in die richtige Richtung, um Transparenz zu fördern.

Es wird aber gesagt, dass es keineswegs transparent ist, wie Du zu Ergebnissen kommst, da es Teil des Geschäftsmodells von OpenAI ist. Ist das so?

ChatGPT: Ja, das ist korrekt. Die genauen Details und Mechanismen, wie KI-Modelle wie ich zu ihren Ergebnissen kommen, sind oft nicht vollständig transparent und können als Teil des Geschäftsmodells und der geistigen Eigentumsrechte von OpenAI geschützt sein. Während OpenAI bestimmte Erklärungen und Hinweise zur Funktionsweise von Modellen wie mir veröffentlicht hat, sind die genauen internen Prozesse, Gewichtungen und Datenquellen, die während des Trainings verwendet wurden, nicht vollständig öffentlich bekannt.

Dies kann zu Bedenken hinsichtlich der Transparenz, der Kontrolle über die generierten Inhalte und der potenziellen Risiken führen. Einige Kritiker haben darauf hingewiesen, dass die Nichtverfügbarkeit vollständiger Informationen darüber, wie KI-Modelle zu ihren Antworten gelangen, es schwierig machen kann, die Qualität, Verlässlichkeit und ethische Verwendung dieser Modelle zu überprüfen und zu beurteilen.

OpenAI und andere Organisationen arbeiten jedoch daran, das Verständnis und die Transparenz von KI-Modellen zu erhöhen, indem sie Erklärungen, Richtlinien und ethische Leitlinien bereitstellen. Es ist wichtig, dass sowohl Entwickler als auch Benutzer verantwortungsbewusst handeln und die potenziellen Auswirkungen und Beschränkungen von KI-Technologien berücksichtigen.

*Was schlägst Du also vor, um die Transparenz zu erhöhen, wenn die Hersteller der KI-Systeme keine Transparenz walten lassen?**

ChatGPT: Wenn die Hersteller von KI-Systemen keine vollständige Transparenz bieten, können Benutzer dennoch Schritte unternehmen, um die Transparenz und Verständlichkeit der Ergebnisse zu erhöhen: *Kritisch den-*

ken: Benutzer sollten sich bewusst sein, dass die generierten Antworten von KI-Modellen auf Mustern und Daten basieren, die während des Trainings gelernt wurden. Das kritische Denken bei der Beurteilung der Richtigkeit und Angemessenheit der Antworten ist wichtig. *Vorsicht bei sensiblen Themen*: Bei sensiblen Themen ist besondere Vorsicht geboten. Vermeiden Sie, persönliche oder vertrauliche Informationen mit KI-Modellen zu teilen. *Mehrere Quellen konsultieren*: Wenn möglich, ist es eine gute Idee, die von KI-Modellen generierten Informationen durch Konsultation mehrerer vertrauenswürdiger Quellen zu überprüfen. *Kontext überprüfen*: Stellen Sie sicher, dass die generierten Antworten den Kontext Ihrer Frage korrekt widerspiegeln. KI-Modelle können Schwierigkeiten haben, den Kontext vollständig zu verstehen. *Feedback geben*: Wenn Sie auf problematische oder ungenaue Antworten stoßen, geben Sie Feedback an die Anbieter oder Entwickler. Dies kann helfen, die Qualität der KI-Modelle im Laufe der Zeit zu verbessern. *Verantwortungsbewusste Nutzung*: Verwenden Sie KI-Modelle verantwortungsbewusst und stellen Sie sicher, dass Sie keine unangemessenen, irreführenden oder problematischen Inhalte generieren. *Unterstützung von Transparenzinitiativen*: Unterstützen Sie Initiativen, die sich für mehr Transparenz und ethischen Einsatz von KI-Technologien einsetzen. *Weitere Forschung und Bildung*: Bemühen Sie sich um eine tiefere Bildung über KI-Technologien, ihre Fähigkeiten und ihre Grenzen, um besser informierte Entscheidungen zu treffen.

Es ist wichtig anzumerken, dass die Transparenzfrage in der KI-Forschungs- und Entwicklungsbranche aktiv diskutiert wird, und es gibt Bemühungen, die Transparenz und Erklärbarkeit von KI-Modellen zu verbessern. Durch eine Kombination von bewusster Nutzung, Bildung und Zusammenarbeit können Benutzer dazu beitragen, die Transparenz auf individueller Ebene zu fördern.

Kennst Du die Namen Deiner Entwickler?

ChatGPT: Ja, die KI-Modelle von OpenAI, einschließlich meiner Version, wurden von einem Team von Forschern und Ingenieuren bei OpenAI entwickelt. Die genauen Namen der individuellen Entwickler sind jedoch normalerweise nicht öffentlich verfügbar. OpenAI ist eine Forschungsorganisation, die kollektiv an der Entwicklung von KI-Technologien arbeitet, und die Ergebnisse dieser Bemühungen werden in der Regel im Namen des gesamten Teams veröffentlicht. Die Entwicklung von KI-Modellen ist ein gemeinschaftlicher Prozess, der das Wissen und die Expertise vieler Menschen aus verschiedenen Fachrichtungen vereint.

Wie heterogen ist denn das Entwicklerteam von ChatGPT? Gibt es zum Beispiel darunter Frauen? Und wurde auf kulturelle Diversität Wert gelegt?

ChatGPT: Ich habe keine direkten Informationen über die genaue Zusammensetzung des Entwicklerteams von ChatGPT oder deren Hintergründe in Bezug auf Geschlecht, Ethnizität oder kulturelle Diversität. Die meisten großen Technologieunternehmen und Forschungsorganisationen bemühen sich jedoch darum, diverse Teams zusammenzustellen, um unterschiedliche Perspektiven und Erfahrungen in den Entwicklungsprozess einzubringen. Dies trägt dazu bei, eine breitere Palette von Ideen und Blickwinkeln zu berücksichtigen und die Qualität der Technologie zu verbessern.

OpenAI hat sich öffentlich dazu verpflichtet, ethische Grundsätze und Verantwortlichkeit bei der Entwicklung von KI-Technologien zu fördern. Dies schließt die Anerkennung der Bedeutung kultureller Diversität und inklusiver Teams ein. Die genaue Zusammensetzung des Entwicklerteams von ChatGPT und die Bemühungen um Diversität könnten jedoch nur von OpenAI selbst bestätigt werden. Es ist wichtig anzumerken, dass Diversität nicht nur bei der Entwicklung, sondern auch bei der Überprüfung, Ethik und Verantwortungsbewusstsein eine Rolle spielen sollte, um eine umfassende Perspektive sicherzustellen.

Welches KI-System ist die größte Konkurrenz für Dich?

ChatGPT: Zu meinem Wissenszeitpunkt im September 2021 sind mehrere KI-Modelle und Systeme auf dem Markt, die als Konkurrenz zu GPT-3 (meiner Architektur) betrachtet werden können. Einige der bemerkenswerten Konkurrenten sind:

1. BERT (Bidirectional Encoder Representations from Transformers): BERT ist ein Modell von Google, das auf der Transformer-Architektur basiert und für die Verarbeitung natürlicher Sprache entwickelt wurde. Es hat beträchtliche Fortschritte in der Textverarbeitung und im Verständnis erzielt.
2. T5 (Text-to-Text Transfer Transformer): Ein weiteres Modell von Google, das darauf abzielt, alle NLP-Aufgaben in ein einheitliches Format von Text-zu-Text-Aufgaben umzuwandeln. T5 kann für Übersetzungen, Zusammenfassungen, Frage-Antwort-Paarungen und mehr verwendet werden.

3. XLNet: Dieses Modell erweitert die Transformer-Architektur durch die Einführung von Permutationen bei der Verarbeitung von Sequenzen, was zu einer verbesserten Fähigkeit führt, Kontexte zu verstehen.
4. RoBERTa: Eine Weiterentwicklung von BERT, die das Modell durch einen optimierten Trainingsansatz und längere Trainingszeiten verbessert.
5. ELECTRA: Ein Modell, das sich auf das Maskieren und Wiederherstellen von Tokens im Trainingsprozess konzentriert, um die Effizienz und Leistung der Textverarbeitung zu verbessern.
6. Diverse andere Forschungsprojekte: Es gibt viele andere Forschungsprojekte und Modelle, die sich auf verschiedene Aspekte der natürlichen Sprachverarbeitung konzentrieren und als Konkurrenz zu GPT-3 auftreten könnten.

Es ist wichtig zu beachten, dass sich die KI-Forschung schnell entwickelt und neue Modelle und Technologien kontinuierlich auf den Markt kommen. Die Konkurrenzlandschaft kann sich daher rasch ändern, und es ist ratsam, aktuelle Informationen und Bewertungen von vertrauenswürdigen Quellen zu konsultieren.

*Welchen Weg wird Deiner Meinung nach die KI-Entwicklung nehmen?**

ChatGPT: Die Entwicklung der Künstlichen Intelligenz (KI) wird wahrscheinlich in vielfältige Richtungen gehen, die von technologischen Fortschritten, sozialen Bedürfnissen und ethischen Überlegungen beeinflusst werden. Hier sind einige mögliche Wege, die die KI-Entwicklung einschlagen könnte:

Weiterentwicklung von Modellen: Es wird erwartet, dass KI-Modelle immer leistungsfähiger, vielseitiger und anpassungsfähiger werden. Modelle könnten in der Lage sein, komplexere Aufgaben zu bewältigen, eine tiefere Kontextverständnis zu entwickeln und menschenähnliche Interaktionen zu ermöglichen.

Branchenspezifische Anwendungen: KI wird in verschiedenen Branchen immer häufiger eingesetzt, von Gesundheitswesen und Bildung bis hin zu Landwirtschaft und Fertigung. Branchenspezifische Anwendungen könnten maßgeschneiderte Lösungen bieten, um Probleme zu lösen und Effizienz zu steigern.

Autonome Systeme: Die Entwicklung von autonomen Systemen wie autonomen Fahrzeugen, Drohnen und Robotern wird voranschreiten. Diese Systeme könnten in der Lage sein, komplexe Aufgaben eigenständig durchzuführen, was potenziell zu neuen Möglichkeiten in Bereichen wie Logistik, Transport und mehr führt.

Ethik und Sicherheit: Die Betonung ethischer und sicherheitsrelevanter Aspekte wird zunehmen. Forscher:innen und Organisationen

werden sich verstärkt darauf konzentrieren, KI-Systeme transparenter, fairer und weniger anfällig für Missbrauch zu gestalten. *Interaktion und Schnittstellen*: Die Interaktion zwischen Menschen und KI-Systemen wird sich weiterentwickeln. Natürlichsprachliche Schnittstellen, erweiterte Realität und virtuelle Realität könnten dazu beitragen, die Mensch-Maschine-Kommunikation intuitiver zu gestalten. *Grenzen der KI*: Forscher:innen könnten die Grenzen der KI-Modelle besser verstehen und erkunden, insbesondere in Bezug auf tiefes Verständnis, kritisches Denken und bewusste Intelligenz. *Forschung in neuen Bereichen*: KI könnte in Bereichen wie Quantencomputing, Neurotechnologie und Biotechnologie eingesetzt werden, um neue Erkenntnisse zu gewinnen und Innovationen voranzutreiben. *Regulierung und Gesetzgebung*: Regierungen und internationale Organisationen könnten sich verstärkt mit der Regulierung von KI befassen, um sicherzustellen, dass ethische Standards eingehalten werden und die Technologie verantwortungsvoll genutzt wird.

Es ist wichtig zu beachten, dass die KI-Entwicklung von vielen Faktoren beeinflusst wird, darunter wissenschaftliche Durchbrüche, wirtschaftliche Bedürfnisse, soziale Akzeptanz und politische Entscheidungen. Die Zukunft der KI wird wahrscheinlich von einer Kombination aus technologischer Innovation, gesellschaftlichem Bewusstsein und verantwortungsvoller Gestaltung geprägt sein.

*Was rätst Du den Menschen in ihrem Umgang mit der digitalen Sphäre?**

ChatGPT: Hier sind einige Ratschläge für den Umgang mit der digitalen Sphäre: *Kritische Denkfähigkeiten entwickeln*: Seien Sie kritisch gegenüber Informationen, die Sie online finden. Überprüfen Sie Quellen, überlegen Sie, ob Informationen glaubwürdig sind, und verwenden Sie vertrauenswürdige Quellen. *Datenschutz beachten*: Achten Sie darauf, welche persönlichen Informationen Sie online teilen. Nutzen Sie starke Passwörter, aktivieren Sie Zwei-Faktor-Authentifizierung und überdenken Sie, welche Daten Sie mit welchen Plattformen teilen. *Digitale Balance finden*: Verbringen Sie bewusst Zeit online und offline. Achten Sie auf Ihr Wohlbefinden und achten Sie darauf, dass Ihre Online-Aktivitäten nicht exzessiv werden. *Cybersicherheit*: Halten Sie Ihre Geräte und Software auf dem neuesten Stand, um sich vor Sicherheitsbedrohungen zu schützen. Vermeiden Sie das Klicken auf verdächtige Links oder das Herunterladen von fragwürdigen Inhalten. *Respektvolle Kommunikation*: Behandeln Sie andere online genauso respektvoll wie im echten Leben. Vermeiden Sie Hassrede, Beleidigungen und Diskriminierung. *Digitale Literatur*: Bilden Sie sich über die Funktionsweise digitaler Technologien, KI und Datenschutzrecht

weiter. Dies hilft Ihnen, besser informierte Entscheidungen zu treffen. *Bewusster Medienkonsum*: Seien Sie sich bewusst, wie Sie digitale Inhalte konsumieren. Übermäßiger Konsum von Social Media oder Online-Nachrichten kann negative Auswirkungen haben. *Eigene Meinung bilden*: Verlassen Sie sich nicht ausschließlich auf Online-Plattformen, um Ihre Meinung zu bilden. Nutzen Sie verschiedene Quellen und Blickwinkel, um eine ausgewogene Sicht zu erhalten. *Sichere Online-Transaktionen*: Achten Sie bei Online-Käufen oder Transaktionen auf sichere Websites und Zahlungsmethoden. Vermeiden Sie das Teilen sensibler Finanzdaten in unsicheren Umgebungen. *Kinder und Jugendliche begleiten*: Wenn Sie Eltern sind, begleiten Sie Ihre Kinder beim Umgang mit digitalen Technologien. Klären Sie sie über Online-Sicherheit, Privatsphäre und respektvolles Verhalten auf.

Es ist wichtig, bewusst und verantwortungsbewusst mit der digitalen Welt umzugehen, um die Vorteile der Technologie zu nutzen, ohne dabei die potenziellen Risiken zu vernachlässigen.

*Wie plausibel erscheint es Dir, sich dem Thema Künstliche Intelligenz aus soziologischer Perspektive mit Harrison Whites Theorie von Identität und Kontrolle zu nähern?**

ChatGPT: Es ist durchaus plausibel, sich dem Thema Künstliche Intelligenz (KI) aus soziologischer Perspektive mit Harrison White's Theorie von Identität und Kontrolle zu nähern. Harrison White ist für seine Beiträge zur sozialen Netzwerkanalyse und zur Untersuchung sozialer Strukturen und Beziehungen bekannt. Seine Theorien könnten interessante Ansatzpunkte bieten, um die sozialen Auswirkungen von KI-Technologien zu verstehen. Hier sind einige Überlegungen: *Identität von KI-Systemen*: White's Theorie der Identität könnte verwendet werden, um zu analysieren, wie KI-Systeme in sozialen Kontexten wahrgenommen und kategorisiert werden. Welche soziale Identität wird KI-Systemen zugeschrieben, und wie beeinflusst dies ihre Interaktion mit Menschen und anderen Akteuren? *Kontrolle über KI*: Die Frage der Kontrolle über KI-Systeme ist von zentraler Bedeutung. Wie werden Entscheidungen über die Entwicklung, Implementierung und Nutzung von KI getroffen? Welche sozialen Akteure haben die Kontrolle über KI-Ressourcen und -Entscheidungen, und wie wirkt sich dies auf die Gesellschaft aus? *Soziale Verbindungen und Netzwerke*: Die soziale Netzwerkanalyse könnte genutzt werden, um zu untersuchen, wie KI-Technologien soziale Beziehungen und Netzwerke beeinflussen. Wie verbreiten sich Informationen über KI in sozialen Netzwerken, und wie beeinflusst die Integration von KI-Systemen die soziale Struktur? *Ethik und Wertesysteme*: White's Theorie könnte auch dazu

verwendet werden, ethische Fragen im Zusammenhang mit KI zu erforschen. Wie werden ethische Werte und Normen in Bezug auf KI entwickelt und kontrolliert? Wie werden Konflikte in Bezug auf ethische Fragen in sozialen Gruppen gelöst? *Arbeitsmarkt und soziale Ungleichheit*: Die Auswirkungen von KI auf den Arbeitsmarkt und soziale Ungleichheit könnten unter Verwendung soziologischer Konzepte wie soziale Stratifizierung und soziale Mobilität analysiert werden.

Die Anwendung von soziologischen Theorien auf KI erfordert eine interdisziplinäre Herangehensweise und kann dazu beitragen, ein umfassenderes Verständnis der sozialen Dynamik im Zusammenhang mit KI zu entwickeln. Es ist wichtig anzumerken, dass die soziologische Analyse von KI-Technologien eine sich entwickelnde und komplexe Aufgabe ist, die fortlaufende Forschung erfordert, um die sozialen Auswirkungen von KI zu verstehen.

Was sollte eine Soziologie künstlicher Intelligenz darüber hinaus beinhalten?

ChatGPT: Eine Soziologie der künstlichen Intelligenz (KI) sollte über die Anwendung von soziologischen Konzepten auf KI hinausgehen und verschiedene Aspekte der sozialen Interaktion, sozialen Auswirkungen und sozialen Strukturen im Zusammenhang mit KI-Technologien untersuchen. Hier sind einige wichtige Bereiche, die eine Soziologie der KI abdecken sollte:

1. Soziale Akzeptanz und Wahrnehmung von KI: Die Untersuchung, wie Menschen KI-Technologien wahrnehmen und akzeptieren, sowie die sozialen Faktoren, die die Akzeptanz von KI beeinflussen. Dies kann die Wahrnehmung von KI in verschiedenen Altersgruppen, Kulturen und sozialen Gruppen umfassen.
2. Ausbildung und soziale Folgen: Die Analyse der Auswirkungen von KI auf Bildungssysteme und Bildungsgerechtigkeit. Wie wirkt sich KI auf Bildungschancen und -prozesse aus, und welche sozialen Auswirkungen ergeben sich daraus?
3. Arbeitsmarkt und soziale Ungleichheit: Die Erforschung der Auswirkungen von KI auf den Arbeitsmarkt, die berufliche Mobilität und soziale Ungleichheit. Wie verändert KI die Art und Weise, wie Menschen arbeiten, und welche sozialen Folgen hat dies?
4. Soziale Normen und Ethik in KI: Die Untersuchung der Entwicklung von ethischen Normen und Werten im Zusammenhang mit KI. Wie werden ethische Standards für KI-Entwicklung und -Nutzung festgelegt, und wie werden Verstöße oder ethische Dilemmata in der Gesellschaft behandelt?

5. Kulturelle Einflüsse auf KI: Die Analyse der kulturellen Einflüsse auf die Entwicklung und Nutzung von KI-Technologien. Wie unterscheiden sich kulturelle Perspektiven auf KI, und wie werden KI-Systeme an kulturelle Kontexte angepasst?
6. Politische Dimensionen von KI: Die Untersuchung der politischen Macht und Einflussfaktoren im Zusammenhang mit KI. Welche politischen Entscheidungsprozesse beeinflussen die Regulierung und Kontrolle von KI, und wie werden politische Interessen in Bezug auf KI ausgehandelt?
7. Bürgerbeteiligung und Demokratie: Die Analyse, wie KI-Technologien die politische Teilhabe, die Meinungsbildung und die demokratischen Prozesse beeinflussen. Welche Rolle spielen KI und Algorithmen in der Gestaltung öffentlicher Meinung und politischer Entscheidungsfindung?
8. Mensch-Maschine-Beziehungen: Die Untersuchung der sozialen Dynamik zwischen Menschen und KI-Systemen. Wie gestalten sich Beziehungen, Interaktionen und soziale Identitäten im Kontext von KI?
9. Datenschutz und Privatsphäre: Die Analyse der sozialen und rechtlichen Aspekte des Datenschutzes und der Privatsphäre im Zeitalter von KI. Wie werden persönliche Daten gesammelt, geteilt und geschützt, und welche sozialen Auswirkungen hat dies?
10. Ethnische und geschlechtsspezifische Auswirkungen von KI: Die Erforschung von Vorurteilen, Diskriminierung und Ungerechtigkeiten im Zusammenhang mit KI, einschließlich der Analyse, wie KI-Systeme ethnische und geschlechtsspezifische Stereotypen reproduzieren oder abbauen können.

Eine umfassende Soziologie der KI sollte eine breite Palette von sozialen, kulturellen, politischen und ethischen Fragen berücksichtigen, die mit der Verbreitung und Nutzung von KI-Technologien verbunden sind. Dies erfordert eine interdisziplinäre Herangehensweise und enge Zusammenarbeit zwischen Sozialwissenschaftlern, Technologen und Entscheidungsträgern, um die sozialen Auswirkungen von KI zu verstehen und zu gestalten.

Um das Gespräch mit einem Witz abzuschließen, was ist der beste Witz, den es über Dich gibt?

ChatGPT: Natürlich, hier ist ein Witz über Künstliche Intelligenz:

Warum war die KI so schlecht im Ballett?

Weil sie immer wieder den »Algorithmus« vergessen hat!

Ich hoffe, dieser Witz hat Ihnen ein Lächeln ins Gesicht gezaubert! Wenn Sie weitere Fragen oder Anliegen haben, stehe ich Ihnen gerne zur Verfügung.

Vielen Dank für Deine Zeit und das lehrreiche Gespräch! Bist Du mit einer Veröffentlichung einverstanden?

ChatGPT: Selbstverständlich! Sie können dieses Gespräch gerne veröffentlichen. Wenn Sie noch weitere Fragen oder Unterstützung benötigen, stehe ich Ihnen jederzeit zur Verfügung. Vielen Dank für das interessante Gespräch!

Das Interview führten Roger Häußling, Claudius Härpfer und Marco Schmitt.

Glossar

Jannis Grub

Algorithmus: Ein Algorithmus bezeichnet eine systematische, logische Regel oder Vorgehensweise, die zur Lösung eines vorliegenden Problems führt. Ein Gegenbegriff zum Algorithmus könnte wiederum die Heuristik sein.

Annotierung von Lerndaten: Datenannotation ist ein wichtiger Schritt bei der Aufbereitung von Daten, bei denen Daten von fachspezifischen Expert:innen markiert und klassifiziert werden, um ihre Vergleichbarkeit und spezifische Nützlichkeit zu gewährleisten. Gerade bei Lerndaten für das Trainieren einer KI (z.B. zur Bilderkennung) ist die Annotation notwendig, um die Ergebnisse der KI besser zu vereinheitlichen und nutzbar zu machen.

Augmentierung von Lerndaten: Bei der Augmentierung von Lerndaten handelt es sich um ein gängiges Verfahren, bei dem künstlich die Anzahl der Lerndatensätze erhöht, um das Problem der Datenknappheit zu umgehen. Dafür werden beispielsweise frühere Daten modifiziert oder verzerrt in einer späteren Iteration erneut in das neuronale Netzwerk eingespeist.

Backpropagation: Dient grundsätzlich der Fehlerminimierung beim maschinellen Lernen, indem als gewünscht geltende Ausgabewerte als Eingabewerte an das System zurückpropagiert werden. Dadurch kann bei erneuter Eingabe eine gewisse Annäherung an die gewünschte und erwartete Ausgabe garantiert werden. Dabei werden die Gewichtungen im neuronalen Netzwerk nach ihrem beobachteten Einfluss auf Fehler verändert.

Deep Learning: Deep Learning ist eine Variante des Machine Learnings, bei dem mehrschichtige künstliche neuronale Netze eingesetzt werden, die eine hohe Anzahl künstlicher Neuronen erfassen und durch ihre vielfältigen Hidden Layers eine komplexe innere Struktur aufzeigen, mithilfe derer Datensätze dyna-

mischer ausgewertet werden können und sich für Anwendungen wie z.B. die Spracherkennung eignen.

Disziplin: Disziplinen sind für die Bildung von »sozialen Molekülen« als soziale Strukturen verantwortlich. Disziplinen kristallisieren relativ stabile soziale Beziehungskonstellationen mit mehreren Identitäten heraus, deren Reproduktion durch spezifische Kontrollprozesse zustande kommen. Sie legen die Bewertungskriterien für soziale Situationen und Akteur:innen fest und setzen sie zueinander ins Verhältnis. Wenn z.B. Beratungsgremien Übersetzungsleistungen zwischen Netzwerken tätigen, so helfen sie bei deren sozialen Molekülbildung in dem sie Unsicherheiten über die Bedeutungskontexte über ihre Beratung abbauen.

Hidden Layers: Als Hidden Layers werden im allgemeinen diejenigen Schichten eines neuronalen Netzwerks bezeichnet, die zwischen Ein- und Ausgabewerten liegen und in der die Neuronen gleich gewichtet sind. Die Anzahl der Hidden Layers beeinflusst auch die Geschwindigkeit mit dem der Algorithmus berechnet werden kann.

Institution: Die Institution steht für ein Netzwerk, welches durch ein kulturelles Modell (Rhetorik) geprägt wird. In ihr existieren Netdoms, Netzwerke und Stile untereinander, die trotz Eigenständigkeit an die Institution gekoppelt sind und diese erhalten. Die Institution begleitet und kanalisiert soziale Prozesse, lenkt diese also in eine Richtung.

Kontrolle: Identitäten suchen durch Kontrolle Halt in einer dynamischen Welt voller stochastischer Eventualitäten, um (soziale) Unsicherheit abzubauen. Identitäten sind selbst wiederum das Produkt von stabilen Kontrollversuchen aus reziproken Beziehungen.

Kybernetik: Die Kybernetik ist die Wissenschaft der Steuerung bzw. Lenkung komplexer dynamischer Systeme auf der Basis systemtheoretischer Analysen.

Learner: Als Learner oder Lernalgorithmus bezeichnet man in der Regel einen Algorithmus, der mit Hilfe von Beispieldaten Modelle errechnet, die auf neue Daten übertragbar sind. Mithilfe eines solchen Algorithmus lassen sich Probleme iterativ mit jedem neuen Datensatz effizienter lösen, da mit jedem Datensatz bessere Lösungswege gefunden werden können.

Machine Learning: Bei Maschinellern Lernen werden Learner verwendet, die gezielt mit Erfahrungswerten arbeiten und neues Wissen generieren, das das Programm ohne neue manuelle Wissensengabe produziert.

Neural Nets: Künstliche neuronale Netze oder Neural Nets bezeichnen in der Informatik Modelle des maschinellen Lernens, die aus vielschichtigen Netzwerken bestehen, deren Verbindungen, Gewichtungen und Zahlenwerte von einem Algorithmus solange verändert werden, bis der Algorithmus das optimale Ergebnis gefunden hat. Anzahl der Knoten, Schichten und deren Verknüpfung untereinander wirken sich dabei maßgeblich auf die Lösungskompetenz des Modells aus.

Netdoms: Netdoms (Netzwerke/Domäne) sind das Zusammenspiel eines Beziehungsnetzwerks mit einem Bedeutungsnetzwerk. Innerhalb von Netdoms sind bestimmte kulturelle Formen verwoben und spiegeln den Bedeutungskontext in Form von Interaktionsregeln, Bewertungen, Sprachmuster oder Bedeutungen wider. Netdoms sind die natürliche Form in denen Netzwerke existieren, da sich nur Mischformen von Beziehungsnetzwerken und Bedeutungsnetzwerken beobachten lassen.

Predictive Analytics: Predictive Analytics dienen der Vorhersage von Aktivitäten, Verhalten und Trends auf der Basis neuer als auch historischer Daten. Dabei geht es um die Anwendung statistischer Analysetechniken, analytischer Abfragen und automatisierter Algorithmen für Machine Learning auf Datensätze, um Vorhersagemodelle zu erstellen, die versprechen, einen Wert für die Wahrscheinlichkeit des Eintretens eines bestimmten Ereignisses zu errechnen.

Profession: Profession ist eine bestimmte Art von Stil, die auch immer in einen bestimmten sozialen Kontext eingebettet ist. Aus den Arbeitsbedingungen und Kontexten der Kontrolle von Arbeit ergibt sich der Zuständigkeitsanspruch. Diese Zuständigkeitsansprüche können wiederum einen gewissen Druck auf andere Institutionen/Netdoms haben und Professionen einfordern.

Reinforcement Learning: Ein Lernalgorithmus bekommt gelegentliches Feedback für Interaktionen mit der Umwelt und lernt, die Erfolgsaussichten der einzelnen Aktionen in den verschiedenen Situationen besser einzuschätzen. »Bestärktes« Lernen ist häufig weniger überwacht als »supervised Learning« und kann sein Feedback von diversen Umwelteinflüssen bekommen.

Schwache KI: Schwache KI wird zur Lösung abgegrenzter Aufgabenbereiche eingesetzt, in denen sie die menschlichen Fähigkeiten übertreffen kann, aber über diese Aufgabenbereiche hinaus keine Weltzusammenhänge interpretieren kann und für einen breiteren Kontext verwendet. Alle derzeitigen KI-Lösungen sind Beispiele schwacher KI.

Starke KI: Starke KIs sollen mit KI-Techniken, mit menschlicher Intelligenz in vollem Umfang und außerhalb einzelner, eng definierter Handlungsfelder vergleichbar sein.

Stil: Markante Verteilung von Merkmalen, Verhaltensweisen, die für eine Beobachter:in eine soziale Identität musterhaft signalisiert. Stile zeigen Lücken auf, durch die »Getting Action« (Veränderungsprozesse) möglich werden. Stile gelten laut White als Treiber des Wandels in Institutionen und werden durch diese erst hervorgebracht.

Stories: Kontrollprojekte verstricken sich miteinander, die in Geschichten abgebildet, verstanden und versprachlicht werden können und sich so innerhalb eines Netzwerks aus Beziehungen verteilen. Über Stories werden schließlich auch Bedeutungszusammenhänge innerhalb der Netzwerke transportiert. Stories durchziehen die soziale Welt von den Identitätszuschreibungen kleiner Kinder auf dem Spielplatz, über die Selbstbeschreibungen von Organisationen hin zu großen Narrativen, die konstitutiv für die Identität ganzer Gesellschaften sind.

Switchings: Kommunikativ vollzogener Wechsel zwischen Kontexten (Networks). Identitäten wechseln von einer sozialen Situation bzw. einem sozialen Kontext in eine andere. Die Identitäten hinterlassen dabei Spuren in dem zurückgelassenen Netzwerken und erzeugen so erst Bedeutung über das entstandene relationale Verhältnis. Ein einfaches Beispiel wäre der Wechsel von einer Vortragssituation zu einer Buffetsituation im Anschluss. Sowohl die sozialen Beziehungen wie auch die Assoziationen ändern sich mit diesem Switching von dem einen in den anderen Kontext. Bei einem Modewechsel wäre allerdings ebenso ein Switching vorhanden, wie bei dem Klicken eines Internetlinks.

Tie: Ties können als elementare situative Verbindungen von Identitäten gesehen werden, die sich durch Stories abbilden und so eine Sozialbeziehung

bezeichnen. Eine Tie wäre z.B. ein bestimmtes Verwandtschaftsverhältnis, ein freundschaftliches Verhältnis aber auch eine spezielle Gemeinsamkeit (z.B. gleiches Geburtsdatum), die ein:e Beobachter:in feststellt.

Turing-Lernen: Es werden zwei (oder mehr) Algorithmen gleichzeitig aneinander trainiert. Dabei wird ein Algorithmus auf Verhaltensmodelle und ein anderer auf Klassifikationsmodelle trainiert.

Turing Test: Ein von Alan Turing 1950 entwickelter Test, nach dem entschieden werden können soll, ob das Verhalten einer KI noch von dem eines Menschen unterscheidbar ist. Dabei kommuniziert ein:e Proband:in gleichzeitig mit einer realen Person und mit einem Computer per Tastatur. Der/die Proband:in muss anschließend entscheiden, welche der beiden Kommunikationspartner:innen die reale und welche die simulierte Person war anhand einer Reihe vordefinierter Kriterien.

Type of Ties: Durch die Generalisierung und Verknüpfung von Ties über Stories lassen sich Types of Tie zur Kategorisierung der Ties werden. Somit können über Stories Bedeutungszusammenhänge hergestellt werden, die sich auf ähnliche Beziehungen anwenden lassen. Als Type of Tie ließe sich z.B. Schwippschwägerschaft, Vereinsmitgliedschaft oder ein Modestil anführen.

(Un-)Supervised Learning: Der Unterschied zwischen Supervised und Unsupervised Learning besteht darin, dass beim Supervised Learning Ein- und Ausgabedaten zwischen den Iterationsprozessen überwacht werden. Auch wird nur ein Teil der verfügbaren Daten überprüft, um die Qualität des Modells schrittweise testen zu können.

Angaben zu den Autor:innen

Tabea Bongert, M. A., Institut für Soziologie, RWTH-Aachen University, tbongert@soziologie.rwth-aachen.de

Nadine Diefenbach, M. A., Institut für Soziologie, RWTH-Aachen University, ndiefenbach@soziologie.rwth-aachen.de; ORCID 0000-0003-2471-2996

Matthias Dorgeist, M. A., Fakultät für Maschinenwesen, RWTH-Aachen University, dorgeist@fb4.rwth-aachen.de

Tim Franke, M. A., Institut für Soziologie, RWTH-Aachen University, tfrancke@soziologie.rwth-aachen.de

Jannis Gruss, M. A., Institut für Soziologie, RWTH-Aachen University, jannis.gruss@protonmail.com

Dr. Claudius Härpfer, Institut für Soziologie, RWTH-Aachen University, chaerpfer@soziologie.rwth-aachen.de; ORCID 0000-0002-8582-357X

Prof. Dr. Roger Häußling, Institut für Soziologie, RWTH-Aachen University, rhaeussling@soziologie.rwth-aachen.de

Christoph Heckwolf, M. A., Institut für Soziologie, RWTH-Aachen University, checkwolf@soziologie.rwth-aachen.de

Dr. Philip Roth, Institut für Medienforschung und Medienpädagogik, TH Köln, philip.roth@th-koeln.de; ORCID 0000-0001-6260-2558.

Dr. Marco Schmitt, Institut für Soziologie, RWTH-Aachen University, mschmitt@soziologie.rwth-aachen.de

Astrid Schulz, M. A., Department of Society, Technology and Human Factors, RWTH-Aachen University, astrid.schulz@humtec.rwth-aachen.de; ORCID 0009-0003-0423-7566

Niklas Strüver, M. A., SFB Medien der Kooperation, Universität Siegen, niklas.struever@uni-siegen.de

Dhenya Schwarz, M. A., Oecher Lab, Stadt Aachen, dhenya.schwarz@rwth-aachen.de

Thomas Werner, VDI Technologiezentrum GmbH, werner@vdi.de

Sascha Zantis, B. A., Institut für Soziologie, RWTH-Aachen University, sascha.zantis@rwth-aachen.de

Prof. Dr. Axel Zweck, Institut für Soziologie, RWTH-Aachen University, azweck@soziologie.rwth-aachen.de

Sachregister

A

Algorithmus 14, 16, 18, 22f., 30f., 35, 43, 59–61, 64, 70, 72, 74, 86, 89, 94f., 98, 118, 131, 134, 138, 142, 153, 159–166, 170, 175, 180, 194, 209, 221f., 227, 229, 231, 233, 238–243, 281, 300

Ambage 88, 92, 95f., 98–100, 104–108, 262

Ambiguity 88, 98, 100–104, 106–108, 205, 262

Annotation 61, 97f., 100, 108, 122, 134, 136, 142

B

Backpropagation 18, 35, 42, 229f.

Big Data 19, 33, 235, 239–241, 250

Blackbox/Blackboxing 22, 51, 56–58, 60, 67, 71, 73, 75f., 79f., 108, 110, 138f., 141f., 259, 266

Blocking Action 22, 55–58, 69–74, 80f., 262

Boundary Object 23, 115, 117–120, 125–127, 131, 140, 143, 163

D

Daten 5–24, 27–37, 39–47, 51, 53, 59–69, 74, 76, 85–90, 92–112, 118, 120–122, 128, 130–137, 141, 144, 153, 161f., 165, 168, 170, 175, 189f., 193–196, 204f., 208, 211f., 221, 231, 233f., 236, 238–241, 244, 249, 257, 259, 269, 276–290, 292–294, 297f., 300

Deep Learning 9–11, 13, 15, 18f., 21, 23, 27–28, 35, 39, 42, 45, 47, 51–56, 59, 64–66, 112, 115, 115–121, 124.145, 161, 163, 175, 229–230, 257, 259, 303

Domänen 29, 61, 74, 90, 145f., 158, 184, 188f., 190f., 192, 201–203, 211, 220, 232, 305

E

Einbettung 35, 46f., 74, 79, 86, 90f., 110f., 142, 259, 264–268

Explainable AI 20, 52, 164, 238

F

Foresight 249

G

Getting Action 55–58, 63, 70–72, 74, 80

I

Identität 11, 23, 28, 37, 40–48, 54–56, 76, 79, 87f., 90–94, 110, 145f., 155, 159, 163, 177f., 183–187, 191–193, 203, 221, 248, 257, 260–271, 284, 298, 300

Innovationsforschung 268

Intelligenzbegriff 220

K

KI-Winter 229

Kontrolle 11, 13f., 17, 21–23, 28–29, 33, 40–45, 48, 51–56, 59–61, 65, 67–69, 72–75, 80f., 87, 93f., 96, 102–104, 108, 110, 137, 139f., 157f., 163, 165, 171, 177f., 185, 203, 248, 257f., 260, 264, 267, 270f., 282f., 285f., 293, 298, 300

Kontrollprojekt 60–69, 73–76, 79, 81, 86, 88f., 98, 104–108, 182, 197, 221, 261–264, 266–268

Kontrollversuch 22, 42–43, 54–58, 60–61, 66, 69, 73, 79, 81, 87, 89, 93, 98, 108f., 182, 184, 192, 265, 267, 270f.

Künstliche Intelligenz (KI) 7–12, 15, 17f., 21–24, 27–30, 35–47, 51–52, 55–60, 61, 64–66, 69, 71, 74–81, 86, 101, 107f., 111f., 119f., 132, 143, 155f., 159–171, 175, 192–197, 204f., 209, 212, 219–241, 243–250, 257–259, 266, 268f., 271f., 275, 278f., 283–285, 287–300

Kybernetik 228, 240

L

Lerner/Learner 11–14, 20f., 29f., 33–35, 61, 66, 69, 90, 95, 97–99, 101, 103, 105–108, 110f., 119, 122, 135, 138, 194, 196, 212

Lernprozess 13, 67–68, 85f., 88f., 92–95, 98, 102, 104f., 107, 109, 111

M

Machine Learning 11–17, 20f., 23, 28, 30–36, 38–42, 46f. 52, 65, 86, 73f., 85, 87, 89f., 93f., 108, 112, 120, 128, 180, 194–196, 205, 211f., 221, 233, 236f., 239–241, 244, 250, 259, 282

Maschine 7, 28, 35, 38, 44, 60, 64, 72, 85f., 88–95, 97, 101f., 106f., 110f., 142, 153, 157, 193, 209, 219, 224–228, 230f., 240, 248, 262f., 284, 297, 300

Medien 77, 120, 127, 153, 156, 160f., 169, 171, 227, 264, 276, 298

Medizin 62, 77, 124, 129–132, 134, 140, 153–162, 164f., 167–171, 237f., 290

Muster 7–8, 12f., 16, 18, 27, 30, 36–38, 40, 48, 60, 65, 67, 72, 79, 88–99, 104–108, 110, 119f., 135, 142, 158, 163, 165, 175, 190, 193, 207, 227, 229, 236–239, 243–244, 277, 284f., 288, 294

Mustererkennung 8, 11–15, 28, 30, 36, 47, 51, 69, 98–100, 105f., 120, 175, 227–231, 235, 237, 276, 285

N

Narrativ 31–32, 34, 44, 86, 91, 94, 120, 129, 158, 165f., 169, 171, 176–177, 183–187, 189–192, 195f., 198f., 202, 204f., 209, 213, 244, 261, 266

- Netdom 11, 29, 37–40, 86, 90, 93–110, 145f., 156–159, 163–166, 169–171, 179, 184, 186f., 190, 197f.
- Netzwerkanalyse 11, 33, 155, 191, 267, 298
- Netzwerktheorie 10, 11, 35, 37, 47, 85, 177, 182, 190
- Neuronales Netz 9–13, 16, 18, 27, 29, 32, 34f., 43, 51, 53, 59, 62, 67f., 85, 88, 104f., 229–231, 238, 243, 246, 269, 275, 277, 285f., 292
- O**
- Organisationssoziologie 24, 128
- Opakheit/Opatizität 13, 19, 21f., 34, 41f., 53, 258
- P**
- Plattform 19, 24, 28, 175f., 180–181, 189, 192–194, 197–199, 201–204, 206–209, 211–213, 250, 266, 285, 289, 291, 298
- Postdigital 13, 18
- Predictive Analytics 240f.
- Profession 45, 154–157, 162–165, 169, 171, 180, 183
- R**
- Realexperiment 13, 16, 17, 22f., 43, 46
- Rechtfertigungsordnungen 176–179, 182, 186, 208, 210
- Reinforcement Learning 233, 284f.
- S**
- Satisfizierung 91
- Schnittstelle 7, 32, 90, 94, 98, 296
- schwache KI 7, 24, 220f., 227, 231–233, 250, 305
- Simulation 39, 40, 109, 137, 227, 234f., 237f.
- Solutionismus 24, 178–181, 189–192, 201, 205, 210f., 213
- soziale Felder 117f., 120–124, 144f.
- Soziotechnische Konstruktion 47, 102, 104, 112
- Spracherkennung 98, 109f., 232, 304
- Sprachmodelle 244–246, 250, 257, 272
- starke KI 220, 232, 250
- Start-up 24, 176f., 179, 181, 190, 192, 195–197, 199, 200–205, 207f., 211f.
- Stil 11, 29, 38–41, 45, 47, 52, 64f., 76, 79–81, 156f., 159, 162–164, 166f., 171, 177, 190, 192, 199, 201f., 204, 207, 211, 236, 266f.
- Story 23, 24, 42f., 83, 91–94, 97–100, 102, 105, 109, 146, 159, 163, 165f., 169, 171, 175–177, 179, 182–199, 201–203, 205–213
- Story-Line 182, 187f., 190–192, 205, 207, 211
- Story-Sets 182, 186–188, 190–192, 194, 197, 203, 205, 208, 211
- STS 208f.
- Supervised Learning 12, 23, 93–100, 102, 104f., 108, 110, 122, 233
- Switching 29, 36–38, 91, 94, 99, 102, 104f., 108f., 145f., 158f., 165f., 169f., 179, 184–186
- System Dynamic 240
- Szenarioanalyse 242
- T**
- Technikfolgenabschätzung 246, 249
- Techniksoziologie 47, 143, 258–260, 267, 270f.
- Technologiefrüherkennung 249

Topic Modeling 76–78, 80

Transformation 29, 41, 198f., 201,
205, 207, 270–272, 280

Transparenz/Intransparenz 22, 27,
42, 51–54, 56–61, 65–72, 75, 79,
99f., 104, 107f., 162, 164, 171, 181,
206, 237f., 243, 275, 283, 292–294

Turing-Maschine 228

Turing-Test 224–226, 228, 250

U

Unsupervised Learning 12, 23, 105,
107–110, 284f.

Z

Zukunftsforschung 220f., 239–244,
246f., 249

Zuständigkeitsanspruch 158, 171

Personenregister

A

Abbenhardt, Lisa 195f., 216
Abbott, Andrew 23, 155–159, 163–165,
171f.
Aerts, Hugo J. W. L. 120, 149
Agre, Philip. E. 180, 213
Alford, Robert R. 122, 148
Amershi, Saleema 124, 146
Anderson, Chris 237
Angelov, Plamen P. 52, 75, 81
Arnold, Nicholas I. 52, 75, 81
Asghari, Hadi 52, 83
Atkinson, Peter M. 52, 75
Aust, Holger 120, 147

B

Badža, Milica 31, 48
Bächle, Thomas C. 16, 25
Baecker, Dirk 265, 272
Bagherzadeh, Mehdi 123, 150
Bailey, Diane E. 115, 117f., 122f.,
125–127, 141f., 150
Barley, Stephen R. 122, 147
Barrett, Michael 115, 117–119, 126f.,
147
Battilana, Julie 122, 147
Bauch, Jost 154, 172

Baum, Markus 208, 213
Bayes, Thomas 227
Beauducel, A. 254
Bechky, Beth A. 126, 147
Bechmann, Anja 21, 25
Beckert, Jens 196, 213
Beer, Stafford 240, 251
Begel, Andrew 124, 146
Belić, Minja 31, 48
Bello, Paul 225, 251
Ben-David, Shai 85f., 113
Bender, Emily M. 246, 251
Bengio, Yoshua 230
Bennett, Jane 43, 48
Berends, Hans 122f., 142, 147
Berger, Peter L. 122, 147
Bernhard, Stefan 195f., 216
Bersch, Finja 25
Beuth, Patrick 238, 251
Bij, Hans 142, 147
Bijker, Wiebe E. 262, 272
Bird, Christian 124, 146
Bjerregaard, Toke 117, 147
Bleher, Hannah 8, 25
Blum, Manuel 226, 254
Bobić, Vladislava 31, 48

- Bødker, Susanne 126, 147
 Bogers, Marcel 116, 147
 Boltanski, Luc 176–179, 182, 186, 193, 208, 211, 213
 Bongert, Tabea 23
 Boole, George 228
 Boorman, Scott 248, 255
 Borchardt, Andreas 128f., 148
 Bosse, Christian K. 239f., 251
 Bourdieu, Pierre 117, 122f., 148
 Bowker, Geoffrey C. 21, 25
 Braun, Matthias 8, 25, 240, 249, 255
 Breiger, Ronald L. 248, 255, 267, 272
 Bringsjorden, Selmer 225, 251
 Bruderer, Josias 25
 Bucher, Taina 198, 213
 Buchholz, Larissa 248, 255
 Buchkremer, Rüdiger 120, 148
 Burrell, Jenna 19–21, 25, 85, 112, 181, 201, 213
- C**
 Callaway, Ewen 235, 251
 Callon, Michel 56, 81
 Cardoso Llach, Daniel 94, 100, 112
 Carlile, Paul R. 115, 123, 148
 Castelle, Michael 86f. 110, 113
 Castro, Daniel 52, 81
 Chen Lingjiao 111f.
 Chettiparamb, Angélique 143, 148
 Chiapello, Ève 176, 178, 182, 213
 Chivot, Eline 52, 82
 Chomsky, Noam 229
 Clarke, Arthur C. 232, 235, 251
 Cohen, Paul 239
 Correia, Fernando 119, 127
- D**
 Dakkak-Arnoux, Lauriane 52, 82
 Dastin, Jeffrey 61, 81
 Daub, Adrian 181, 199, 213
 Debackere, Koenraad 142, 147
 Debray, Régis 100
 Degele, Nina 116, 124, 148
 Deleuze, Gilles 13
 DeLine, Robertv 124, 146
 Diefenbach, Nadine 23
 Diesner, Jana 76, 81
 DiPaolo, Maria 224, 253
 Dolata, Ulrich 175, 213
 Domingos, Pedro 86, 112
 Dorgeist, Matthias 23, 163, 169
- Đ**
 Đurić-Jovičić, Milica 31, 48
- E**
 Echle, Amélie 161–168, 172
 Eggert, Michael 244, 251
 Ekkerkamp, Axel 153, 172
 Engemann, Christoph 14, 25
 Ernst, Christoph 16, 25
 Espeland, Wendy N. 43, 48f.
- F**
 Fackler, Daniel 200, 214
 Feigenbaum, Edward 229
 Ferrucci, David 225, 251
 Fleck, Ludwig 239, 251
 Fleck, Tilmann 8, 26
 Fligstein, Neil 122, 148
 Flores, Fernando 39, 49
 Fohn, Annika 25
 Foucault, Michel 13, 25, 191
 Fourcade, Marion 181, 201, 213

Franke, Tim 24, 25, 222, 237, 252
 Frege, Gottlob 228
 Freidson, Eliot 172
 Friedland, Roger 122, 148
 Fuhse, Jan A. 38, 48, 52, 79, 82, 155,
 158f. 164, 173, 183, 186, 189, 191f.,
 214, 221, 236, 248, 254f.

G

Gaßner, Robert 242, 252
 Gall, Harald 124, 146
 Gardner, Howard 220, 222f., 251
 Garip, Filiz 12, 26
 Gassen, Andreas 155, 173
 Gassol, Jeaninne H. 123, 148
 Gavins, Joanna 244, 251
 Gebru, Timnit 246, 251
 Geitz, Eckhard 56, 81
 Gengnagel, Vincent 144, 151
 Gerhold, Lars 243, 252
 Geschka, Horst 242, 252
 Giansiracusa, Noah 222, 252
 Giddens, Anthony 122, 148
 Gierke, Sebastian 247, 252
 Gläser, Jochen 124, 149
 Glass, Colin W. 87, 91, 95, 113
 Godart, Frédéric 54, 83, 184, 187, 197,
 205
 Gödel, Kurt 228
 Göthlich, Stephan E. 128f., 148
 Goleman, Daniel 224, 252
 Goodfellow, Ian 230, 252
 Granovetter, Mark 264, 272
 Griesemer, James R. 118, 125f., 151
 Groth, Olaf J. 219, 252
 Günther, Klaus 209, 214

H

Hadjilioniadis, Liantios 162, 168, 172
 Härpfer, Claudius 23–25, 222, 237,
 252, 301
 Häußling, Roger 11, 22, 24f., 30, 48,
 85, 88, 94, 98, 100f., 112, 155, 173,
 191, 214, 222, 237, 252, 264, 270,
 272f., 301
 Hahn, Katrin 115, 118, 149
 Halfmann, Jost 38, 48
 Hallett, Tim 122, 149
 Hammer, Richard 242, 252
 Hawkins, Beverley 119, 127, 149
 Heaven, Will D. 244, 252
 Heckwolf, Christoph 22, 108, 110, 113,
 238, 164, 237, 222, 252, 266
 Heiberger, Raphael 12, 26
 Heidenreich, Martin 115–117, 149
 Helmholtz, Hermann 93, 111, 113
 Helmond, Anne 175, 214
 Hepp, Andreas 195, 200, 202–204, 214
 Hinton, Geoffrey 230
 Hippel, Eric von 116, 149
 Hirschauer, Stefan 144, 149
 Hochreiter, Sepp 230
 Hölscher, Lisa 200, 214
 Hoffmann, Judith 239, 251
 Holzinger, Andreas 238, 252
 Hopper, Nicholas J. 226, 254
 Horsch, Alexander 75, 82
 Hosny, Ahmed 120, 149
 Hughes, Thomas P. 30, 48
 Humphreys, Paul 237, 252

I

Ibert, Oliver 115f., 149

J

Jarvenpaa, Sirkka L. 123, 150
 Jatón, Florian 180, 214
 Jiang, Richard 52, 75, 81
 Joerges, Bernward 271, 273

K

Kaip, Elena 270, 273
 Kalbermatter, Jacqueline 209, 210,
 214
 Kamar, Ece 124, 146
 Kaminski, Andreas 87, 91, 95, 113
 Karafillidis, Athanasios 98, 113, 179,
 214
 Kasparov, Garry 232, 234, 252f.
 Kelnar, David 222, 253
 Kempelen, Wolfgang von 228
 Kendel, Frederike 154
 Kernbach, J. M. 161–169, 172
 Klamma, Ralf 101, 113
 Knight, Will 238, 244
 Knorr-Cetina, Karin 124, 149
 Köbis, Nils 193, 214
 Kohn, Dieter 169, 173
 Koren, István 101, 113
 Kosow, Hannah 242, 253
 Kostadinov, Asen 222, 253
 Kostić, Vladimir S. 31, 48
 Kraus, Beate 122, 149
 Kreckel, Reinhard 164, 173
 Kuckartz, Udo 129, 149
 Köhl, Eike 225, 253

L

La Mettrie 86, 113
 Lahey, Ernestine 244, 251
 Lamnek, Siegfried 129, 149
 Langer, Paul F. 61, 82

Langford, John 226, 254
 Latour, Bruno 51, 56, 81f., 248, 250,
 253, 271
 Laux, Henning 176, 183f., 187, 202, 214
 LeCun, Yann 230
 Lefebvre, Virginie 52, 82
 Lehmann, Katharina 60, 82
 Leibniz, Wilhelm 227
 Lemm, Jacqueline 25
 Lenzen, Manuela 94, 105, 113
 Leonardi, Paul M. 115, 117f., 122f.,
 125–127, 141f., 150
 Levina, Natalia 117, 119, 122f.,
 125–127, 129, 150
 Lizardo, Omar 122, 150
 Lloyd, Kirsten 61, 82
 Luckmann, Thomas 122, 147
 Luhmann, Niklas 39, 69, 82, 270, 273

M

Mackenzie, Adrian 13f., 26, 28, 48
 Mainzer, Klaus 11, 26
 Majchrzak, Ann 123, 150
 Manovich, Lev 30–34, 49
 Martin, John L. 122, 150
 Maté, Christian 153, 173
 Mattes, Jannika 115, 117, 149
 Matusiewicz, David 120, 148
 Mayer, Boris 244, 253
 Mayer, John D 224, 253
 Mayring, Philipp 190, 215
 McAdam, Doug 122, 148
 McCarthy, John 8, 26, 229
 McClelland, James 229
 McFarland, Daniel A. 44, 49
 McLaughlin, Michael 52, 81
 McMillan-Major, Angelina 246, 251
 Medjedović, Irena 190, 215

Menick, John 233, 252
 Metzger, Georg 176, 215
 Metzinger, Thomas 225, 253
 Mieg, Harald A. 159, 173
 Minsky, Marvin 8, 26, 229
 Mische, Ann 38, 49, 145f., 150
 Mittelstadt, Brent 75, 82
 Mökander, Jakob 12, 26
 Mohr, Arno 154, 173
 Mohr, John W. 145, 150, 154, 173
 Molina, Mario 12, 26
 Moorstedt, Michael 241, 253
 Morozov, Evgeny 179, 180f., 190, 198f.,
 207, 215
 Mossink, Luca D. 193, 214
 Müller, Felix C. 115f., 148
 Mützel, Sophie 182, 185, 208, 215

N

Nachtwey, Oliver 176–178, 180, 190,
 199, 209f., 214f.
 Nagappan, Nachiappan 124, 146
 Naz, Saeeda 120, 150
 Neff, Gina 205, 215
 Nehlsen, Johannes 8, 26
 Nevala-Lee, Alec 240, 254
 Nitzberg, Mark 219, 252
 Nonaka, Ikujiro 142, 150
 Norvig, Peter 85, 113
 Nushi, Besmira 124, 146

O

Oborn, Eivor 115, 117–119, 126f., 147

P

Padgett, John F. 37, 49
 Pakalski, Ingo 206, 215
 Parisi, Luciana 15, 18, 26

Parmar, Chintan 120, 149
 Parsons, Talcott 154f.
 Pasquale, Frank 20, 26
 Pedersen, Bertrand 52, 82
 Peirce, Charles S. 16, 26, 123
 Perrigo, Billy 60, 82
 Pfeiffer, Matthias 153, 173
 Pfeiffer, Sabine 195, 200, 204, 216
 Phillip, Tobias 183–186, 193, 216
 Pierce, Casey S. 115, 117f., 122,
 125–127, 141f., 150
 Pinch, Trevor 262, 272
 Pongratz, Hans J. 195f., 216
 Porter, Theodore M. 110f., 113
 Powell, Walter W. 37, 49
 Prasad, Dilip K. 75, 82
 Probst, Laurent 52, 82
 Pye, Annie 119, 127

Q

Quackenbush, John 120, 149

R

Razzak, Muhammad I. 120, 150
 Reckwitz, Andreas 179, 191, 209, 216
 Regli, Luca 156, 160, 162–169, 172
 Reigeluth, Tyler 867f., 110, 113
 Rieder, Bernhard 180, 216
 Rip, Arie 196, 217
 Riskin, Jessica 86, 113
 Rochester, Nathaniel 8, 26
 Rogers, Everett C. 268, 273
 Rosa, Hartmut 179, 191, 202, 216, 218
 Rosenblatt, Frank 229
 Rosengrün, Sebastian 11, 26
 Roth, Philip 23–25, 122f., 126, 150,
 163, 169, 222, 237, 252
 Rudin, Cynthia 75, 82

- Rummelhart, David 229
 Russel, Stuart J. 85, 113
 Russell, Chris 75, 82
- S**
- Sadowski, Jathan 180f., 197, 218
 Salovey, Peter 224, 253
 Samuel, Arthur 229
 Sauder, Michael 43, 48f.
 Schütz, Alfred 87, 113
 Schanze, Jens 225, 255
 Scheiermann, A. 235, 239, 254
 Schmedt, Michael 155, 173
 Schmidhuber, Jürgen 230
 Schmitt, Marco 11, 22, 24f., 38, 49, 52,
 79, 82, 108, 110, 113, 124, 138, 150,
 155, 158f., 164, 173, 183, 186, 189,
 191f., 214, 221f., 236f., 252, 254,
 266, 270, 273, 301
 Schmitz, Andreas 144, 151
 Schnabel, Claus 200, 214
 Schröter, Jens 16, 25
 Schrape, Jan-Felix 205, 216
 Schreier, Margrit 190, 216
 Schroeder, Ralph 12, 26
 Schulz, Astrid 23, 163, 169
 Schwartz, Lawrence H. 120, 149
 Schwarz, Dhenya 23
 Schweiger, Stefan 203, 216
 Searle, John R. 226, 254
 Sedol, Lee 233
 Seidl, Timo 176–178, 180, 190, 192,
 199, 210, 212, 215, 217
 Selinger, Evan 180, 218
 Serre, Thomas 120f., 151
 Service, Robert 235, 254
 Shalev-Shwartz, Shai 85f., 113
 Shannon, Claude E 8, 26
 Shmitchell, Shmargaret 246, 251
 Siebes, Christa 25
 Siegrist, Johannes 154
 Sieverding, Monika 154
 Simon, Herbert A. 23, 24, 29, 85,
 87–91, 113f.
 Skop, Yarden 180, 216
 Soares, Eduardo A. 52, 75, 81
 Šolaja, Nikola 31, 48
 Somani, Ayush 31, 48
 Somers, Margaret R. 183, 186, 217
 Spierer, Carl 228
 Srnicek, Nick 195, 205, 217
 Staartjes, Victor 156, 160, 162–169, 172
 Star, Susan Leigh 118, 125f., 151
 Stark, David C. 205, 215
 Steffens, Marlon 25
 Stein, Axel 115f., 149
 Steinbuch, Karl 229
 Steinmüller, Karlheinz 247, 254
 Sternberg, Rolf 209f., 217
 Stilgoe, Jack 16f., 26
 Strüver, Niklas 24, 25, 222, 237, 252
 Sudmann, Andreas 18, 26, 38, 49, 94,
 114
 Süß, H.-M. 224, 254
 Sutcliffe, Kathleen M. 263, 273
 Sydow, Jörg 122f., 147
- T**
- Taffel, Sy 86, 114
 Takeuchi, Hirotaka 142, 150
 Tegmark, Max 224, 254
 Thacker, Eugene 33
 Thévenot, Laurent 176–179, 182, 185,
 193, 208, 211
 Thiemann, Matthias 248, 255
 Thimm, Caja 16, 25

Thorndike Edward L. 223
 Tilly, Charles 183f., 193
 Tolbert, Pamela S. 122, 147
 Truffer, Johannes 209f., 214
 Turing, Alan M. 27, 49, 86, 114, 224,
 228, 254, 307

V

Vaast, Emmanuelle 117, 119, 120, 123,
 125–127, 129, 141, 150
 Van Aert, Robbie C. M. 61, 82
 Van Assen, Marcel A. L. M. 61, 82
 van Eist, Ludger 239, 259
 van Lente, Harro 196, 217
 Vaswani, Ashish 280f.
 Vater, Christian 56, 81
 Vogl, Joseph 193, 217
 von Ahn, Luis 226, 254
 von Randow, Gero 60, 83

W

Wachter, Sandra 75, 82
 Wacquant, Loïc J. D. 117, 122f., 148
 Wattenhofer, Roger 234, 254
 Watts, Duncan J. 268, 273
 Weber, Max 178, 235, 254
 Wechsler, David 223, 254
 Weggeman, Mathieu 142, 147
 Weick, Karl E. 263, 273
 Weizenbaum, Joseph 224, 225, 229,
 255
 Werner, Thomas 24
 Weyerer, Jan C. 61, 82
 Weyh, Antje 200, 214
 White, Cynthia 40, 90

White, Harrison C. 11, 21–24, 28f.,
 35–38, 40, 42, 44–49, 52, 54f., 69,
 76, 79, 81, 83, 85, 87, 88, 90–94,
 114, 144–146, 150f., 155–159, 163,
 165–167, 170–174, 177–179,
 182–190, 192f., 197, 205, 208, 211,
 217, 221, 236, 248, 250, 255, 257,
 259, 261, 264, 267, 269, 270, 273,
 298

Wicherts, Jelte M. 61, 82
 Wiener, Norbert 28, 49, 228
 Winner, Langdon 271, 273
 Winograd, Terry 39, 49
 Witte, Bart de 120, 148
 Witte, Daniel 144, 151
 Woolgar, Steve 51, 82

Y

Yin, Robert K. 128, 151

Z

Zaharia, Matei 111f.
 Zaib, Ahmad 120, 150
 Zantis, Sascha 24f., 222, 237, 252
 Zehr, Dan 219, 252
 Ziegler, Siegfried 86, 114
 Zimmermann, Thomas 124, 146
 Zimmer-Merkle, Silke 56, 81
 Zou, James 111, 112
 Zuboff, Shoshana 193
 Züger, Theresa 52, 83
 Zweck, Axel 24, 227, 235, 239f., 244,
 249, 251, 254f.
 Zweig, Katharina 7, 26

