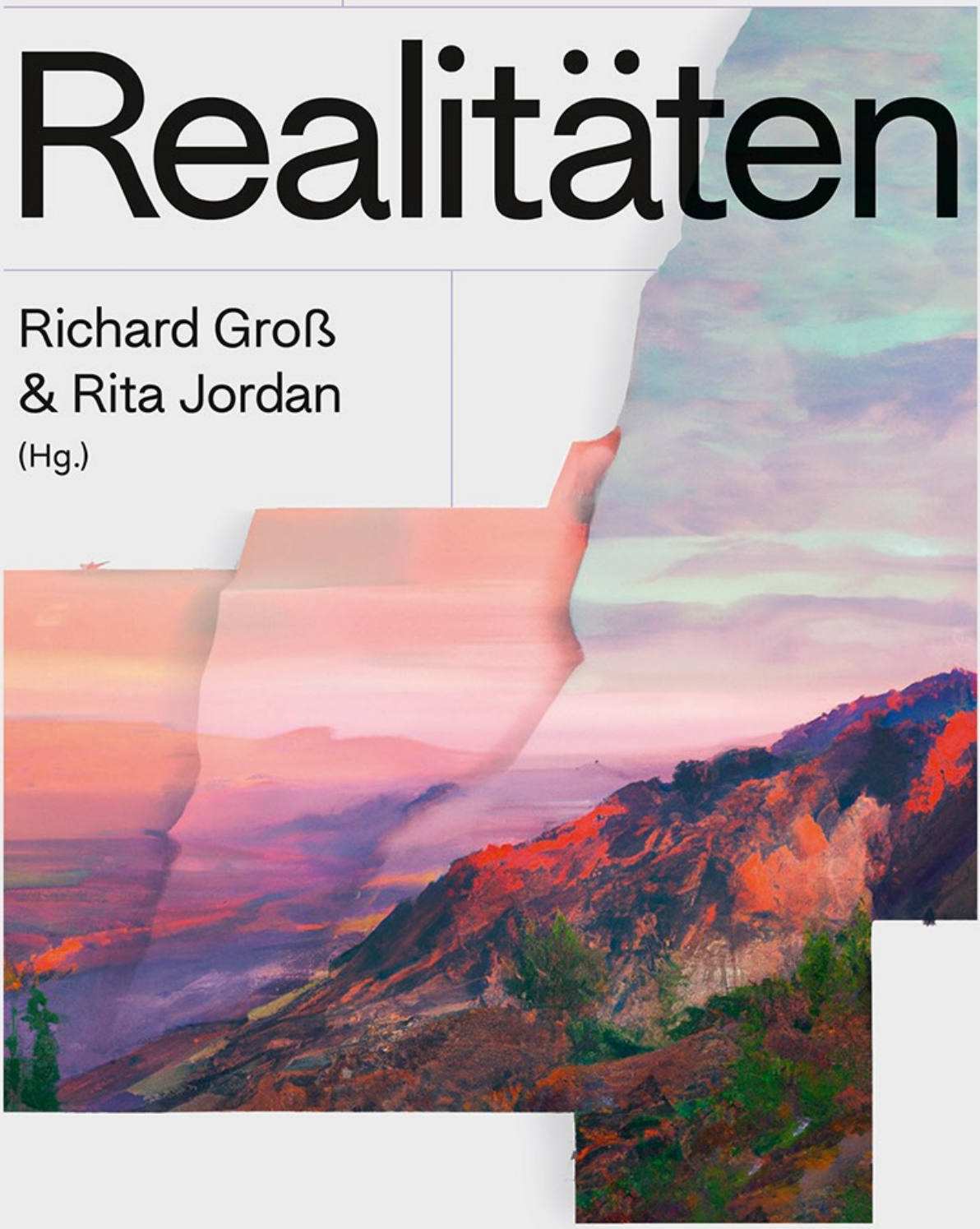
KI-Realitäten

Modelle, Praktiken und Topologien maschinellen Lernens

Richard Groß
& Rita Jordan
(Hg.)



[transcript] KI-Kritik | AI Critique

Richard Groß, Rita Jordan (Hg.)
KI-Realitäten

KI-Kritik / AI Critique | Band 5

Editorial

Kritik heißt zum einen seit Kant das Unternehmen, die Dinge in ihrer Funktionsweise und auf die Bedingungen ihrer Möglichkeit hin zu befragen, sowie zum anderen nach Foucault das Bemühen um Wege, »nicht dermaßen regiert zu werden«. **KI-Kritik / AI Critique** veröffentlicht kultur-, medien- und sozialwissenschaftliche Analysen zur (historischen) Entwicklung maschinellen Lernens und künstlicher Intelligenzen als maßgeblichen Aktanten unserer heutigen technischen Welt.
Die Reihe wird herausgegeben von Anna Tuschling, Andreas Sudmann und Bernhard J. Dotzler.

Richard Groß verfolgt als Promotionsstipendiat des Schaufler Lab@TU Dresden ein ethnografisches Dissertationsprojekt zu Anwendungen maschinellen Lernens in Wissenschaft und Kunst und ist zudem Projektkoordinator der Arnold Gehlen-Gesamtausgabe. Er studierte Soziologie, Kunstgeschichte und Musikwissenschaft in Dresden und New York. Zu seinen Forschungsschwerpunkten zählen Technik- und Medientheorie, Systemtheorie, Philosophische Anthropologie sowie Zeitsoziologie.
Rita Jordan ist Vorstandsreferentin bei der Technologiestiftung Berlin. Zuvor war sie wissenschaftliche Mitarbeiterin am ScaDS.AI Dresden/Leipzig sowie an der Professur für Rechts- und Verfassungstheorie mit interdisziplinären Bezügen der TU Dresden und assoziiertes Mitglied des Schaufler Lab@TU Dresden. Sie erforscht die Schnittstellen von Recht, Politischer Theorie und Technologie. Sie hat Rechts- und Politikwissenschaften in Amsterdam, Berlin, London, Dresden und Wien studiert.

Richard Groß, Rita Jordan (Hg.)

KI-Realitäten

Modelle, Praktiken und Topologien maschinellen Lernens

[transcript]

Dieser Band wurde gefördert durch das gemeinsam von The Schaufler Foundation und der Technischen Universität Dresden finanzierte Schaufler Lab@TU Dresden.

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Erschienen 2023 im transcript Verlag, Bielefeld

Umschlaggestaltung: Bureau Neue, Leipzig
Umschlagabbildung: Arne Winter (Bureau Neue)
Lektorat: Margaret May, Steffen Schröter (text plus form, Dresden)
Druck: Majuskel Medienproduktion GmbH, Wetzlar
https://doi.org/10.14361/9783839466605
Print-ISBN 978-3-8376-6660-1
PDF-ISBN 978-3-8394-6660-5
Buchreihen-ISSN: 2698-7546
Buchreihen-eISSN: 2703-0555

Gedruckt auf alterungsbeständigem Papier mit chlorfrei gebleichtem Zellstoff.
Besuchen Sie uns im Internet: *https://www.transcript-verlag.de*
Unsere aktuelle Vorschau finden Sie unter *https://www.transcript-verlag.de/vorschau-download*

Inhalt

KI-Realitäten/AI Realities
Richard Groß & Rita Jordan 9

Danksagung/Acknowledgments 35

Embedded Models of Meaning

Parrots All the Way Down
Controversies within AI's Conquest of Language
Jonathan Roberge & Tom Lebrun 39

Testimonial Injustice in Governmental AI Systems
Catriona Gray 67

Grade Prediction Is Not Grading
On the Limits of the e-rater
Jan Georg Schneider & Katharina A. Zweig 93

Intelligente Ökologien

Von allopoietischen zu autopoietischen algorithmischen Systemen
Überlegungen zur Eigenlogik der operationalen Schließung algorithmischer Systeme
Jan Tobias Fuhrmann 115

Robotermaterial und ›Künstliche Intelligenz‹
Posthumanistische Potenziale der Robotik
Hannah Link 143

Menschmaschinen und Maschinenmenschen
Überlegungen zur relationalen Ontogenese von Identität
Jonathan Harth & Maximilian Locher 169

Wie die Bildung pragmatischer Handlungsmuster die Mensch-Maschine-Kommunikation gestaltet
Yaoli Du & Nadine Schumann 193

Generative Praktiken

If I Say the Word Out Loud, It Will Be More Real
Jakob Claus & Yannick Schütte 211

do nOt FOrGET AnaRCHist
Syntaktische Sprachexperimente im künstlich neuronalen Wortraum
Christian Heck 235

KI-Kunst als Skulptur
Fabian Offert 273

Repräsentation, Kritik und Anlass
Eine Trichotomie der künstlerischen Nutzungsaspekte von KI
Michael Klipphahn-Karge 287

Composing AI
Formen, Ideen, Visionen
Miriam Akkermann 315

Computational Topologies

Can Artificial Neural Networks Be Normative Models of Reason?
Limits and Promises of Topological Accounts of Orientation in Thinking
Lukáš Likavčan & Carl Christian Olsson 333

Pointless Topology
Figuring Space in Computation and Cognition
AA Cavia & Patricia Reed 351

Autor:innenverzeichnis 365

Contributors 369

KI-Realitäten/AI Realities

Richard Groß & Rita Jordan

KI-Realitäten

Der vorliegende Band präsentiert verschiedene Perspektiven auf Aspekte der realweltlichen Einbettung von maschinellem Lernen (ML). Die hier versammelten Beiträge gehen auf Vorträge zurück, die auf einer vom Schaufler Lab@TU Dresden ausgerichteten interdisziplinären Online-Tagung im Dezember 2021 gehalten wurden. Der Titel des Bandes, *KI-Realitäten*, verweist einerseits auf das Ausmaß, in dem ML-basierte Technik als schon seit mehr als einem Jahrzehnt dominanter Ansatz sogenannter Künstlicher Intelligenz (KI)[1] die Wirklichkeit zu prägen begonnen hat. Andererseits spielt er auf die reflexiven Herausforderungen an, die sich theoretischen Bemühungen um ein adäquates sozial- und geisteswissenschaftliches sowie philosophisches Verständnis von KI stellen. ML stellt in Form praktischer Anwendungen wie auch als Gegenstand spekulativer Imagination einen wichtigen Faktor und zugleich ein Produkt der Realitäten dar, in die es als Technik und Vorstellung eingebettet ist. Die folgenreiche Integration von auf ML-Verfahren beruhender Technik in bestehende Realitäten macht es erforderlich, die einschneidenden und teils widersprüchlichen Effekte ihrer Verbreitung in den Blick zu nehmen. Das als Titel gewählte Kompositum soll daher nicht allein als Ausdruck der gegenwärtigen Größenordnung realweltlicher Auswirkungen von ML-Anwendungen verstanden werden, sondern vor allem auf die in den Auswirkungen sichtbar werdende Vielzahl unterschiedlicher Aspekte von ML als Forschungsgegestand hinweisen.

1 Den Begriff Künstliche Intelligenz nutzen wir im Folgenden, sofern nicht anders angegeben, synonym mit maschinellem Lernen, da ML-Ansätze maßgeblich verantwortlich sind für die wesentlichen Fortschritte jüngeren Datums in der KI-Forschung wie auch in der Anwendung von KI.

Aufgrund dieser Vielzahl solcher, sich mitunter zudem wechselseitig bedingender Aspekte scheint uns eine behutsame Annäherung an die theoretische Auseinandersetzung mit ML ratsam zu sein. So birgt die Auseinandersetzung mit einer – noch dazu von lautstark vernehmbaren Kontroversen begleiteten – vermeintlichen ›Schlüsseltechnologie‹ die Gefahr, einem Gegenstand, der als ergiebiges Spekulationsobjekt immerzu von interessengeleiteten Projektionen überzogen und dabei womöglich mit vielem verwechselt wird, vorschnell bestimmte Eigenschaften zuzuschreiben. Eine Einbettung in historisch gewachsene gesellschaftliche Realitäten bedeutet für jedes noch so technisch fixiert anmutende Phänomen, dass es in dynamischen Beziehungen wechselseitiger Abstimmung und Beeinflussung mit anderen sozialen Entitäten steht. Diese Annahme legt für geistes- und sozialwissenschaftliche Forschung nahe, von der praktischen Situiertheit maschinellen Lernens auszugehen, um deren Bedeutung innerhalb gelebter Realitäten nachzuvollziehen und damit technikwissenschaftlich gewonnene Gegenstandsbestimmungen kritisch reflektierend zu ergänzen. An solche Bestimmungen angelehnt beruht ML – schematisch formuliert – auf dem ›Training‹ von Algorithmen durch die Verarbeitung von (zumeist sehr umfangreichen) Datensätzen. Dieser Vorgang bezeichnet die stochastische Modellierung von Daten, die in der Folge auf Basis des erstellten Modells weitere Daten(-analysen) produzieren können. Die Bedeutung der so produzierten Daten ist oft schwer zu erschließen. Dies gilt selbst und gerade auch dann, wenn die Eigenschaften des Datensatzes im Sinne formaler Verteilungen und Regelmäßigkeiten als bekannt und verstanden gelten (Dourish 2016: 7). Die mitunter überraschenden und unheimlich anmutenden Outputs der Modelle (Bucher 2017) sind meist nicht ohne Weiteres erklärbar und lassen sich hinsichtlich ihrer Implikationen aufgrund der Opazität der rechnerischen Modellierung oft auch im Nachhinein nicht erschließen (Burrell 2016).

Eine wesentliche Herausforderung für sozial- und geisteswissenschaftliche wie philosophische Forschung besteht daher in der adäquaten Interpretation von mit ML-Verfahren generierten Outputs wie etwa Texten und Bildern, besonders im Hinblick auf ihre Genese. Als sinnhafte Beiträge zu gesellschaftlichen Realitäten werden diese häufig zum Gegenstand von Kontroversen. Kaum nachvollziehbare Resultate von Mustererkennungen und -generierungen verweisen auf die nicht intuitiv erschließbare operative Logik der Verfahren. Die auf dieser Logik beruhende nichttriviale Transformation digitaler Inputdaten in sinnförmig rezipierte Outputs vollzieht sich auf eine dem menschlichen Denken fremde Weise. Dies stellt eine Hürde für die geis-

tes- und sozialwissenschaftliche Annäherung an KI dar, die sich in mancherlei Hinsicht als »alien subject« (Parisi 2019) erweist.

Daran anschließend besteht eine weitere Herausforderung für die Forschung zu ML im weiteren Sinne in der Bestimmung der Beziehung zwischen gesellschaftlichen Realitäten und technischen Arrangements, die als KI bezeichnet werden. In angewandter Form ist ML-basierte Technik innerhalb der Realitäten, die sie hervorgebracht haben, allgegenwärtig geworden und trägt nun ihrerseits zur Veränderung wie Reproduktion dieser Realitäten bei. Dies heißt für Theorien maschinellen Lernens, dass sie zugleich Theorien über die Wirklichkeit voraussetzen und im Umkehrschluss auch selbst als solche lesbar sind. Gegenwärtige ML-Anwendungen sind dazu in der Lage, datenförmig verfügbare Text- und Bildwerke zu verarbeiten wie auch selbst zu generieren. Sie vermitteln Sinn und mithin Wirklichkeit, was zur Schaffung neuer Welten und zu neuen Formen des Umgangs mit diesen beitragen kann. Diesem Umstand haben sozial- und geisteswissenschaftliche Theorien maschinellen Lernens Rechnung zu tragen.

KI-Pathologien

Innerhalb dieser Gemengelage gibt es zahlreiche Stimmen, die zu Recht auf die ideologischen Funktionen hinweisen, die mystifizierende Deklarierungen von KI als für Menschen unergründlich und nicht verstehbar erfüllen. Eine solche Rhetorik, so etwa Galloway (2021), überhöhe die Komplexität und Leistungsfähigkeit der technischen Verfahren und lenke von der Tatsache ab, dass ML letztlich nicht viel mehr sei als ein »fancy way to calculate an average« (ebd.). Der anhaltende Hype stelle im Grunde einen »total scam« (ebd.) dar, der die tatsächlichen sozialen Probleme, die sich mit der Verbreitung von ML-Anwendungen ergeben, in den Hintergrund treten lasse.

›KI‹ ist das Resultat kontingenter und umkämpfter historischer Entwicklungen und kann als Ausdruck derselben verstanden werden. Die realen Auswirkungen ihrer Anwendung werden unter anderem von ökonomischen Interessen bestimmt, die aus den Strukturen und Dynamiken einer bestimmten Gesellschaftsordnung erwachsen. Da Outputs aus Modellen generiert werden, die auf durch menschliche Arbeit produzierten Datensätzen beruhen, meint ML praktisch die Extraktion bestehenden Wissens (vgl. Joler/Pasquinelli 2020), nicht zuletzt auch in der Extrapolierung darauf beruhender Prädiktionen. Es setzt sich in soziologischer Betrachtung aus einer Reihe von materiellen Praktiken zusammen, deren Organisation in mancher Hinsicht

dem für den Industriekapitalismus typischen Fließbandprinzip entspricht (vgl. ebd.: 2). ML-Produkte verweisen auf die mühsame Arbeit, die mit der Erhebung und Aufbereitung von Daten einhergeht. Eine solche Sichtweise unterstreicht zudem den Ausbeutungscharakter des soziotechnischen Arrangements, das für das Training und die Anwendung eines Modells vorausgesetzt werden muss. In diesem Sinne operiert ML akkumulativ und stellt keine radikale Transformation dar (vgl. Mackenzie 2017). Seine soziale Einbettung impliziert die Abhängigkeit von und die Integration in gesellschaftliche Strukturen, innerhalb derer es Wirksamkeit entfalten kann.

ML-Anwendungen, so lässt sich beobachten, wirken in bestehenden Realitäten dabei häufig derart, dass sie deren pathologische ideologische Tendenzen reproduzieren, wenn nicht sogar verstärken. Zu diesen Tendenzen gehören die Diskriminierung marginalisierter sozialer Gruppen (vgl. Noble 2018; Apprich et al. 2018; Chun 2021), die Konzentration von Kapital und politischer Macht bei Plattformen, über die Oligarchen walten (vgl. Whittaker 2021), die Zunahme vernetzter datenbasierter Formen sozialer Kontrolle, die sich mit dem Bild des distribuierten »Polyopticon« (Sherman 2022) beschreiben lassen, sowie die Missachtung von Nachhaltigkeitsfragen, die sich in den von ML-Technik verursachten Umweltbelastungen äußert (vgl. Strubell et al. 2019; Dhar 2020; Bender et al. 2020). Diese Probleme – wenn auch nicht auf ML beschränkt, sondern gleichermaßen für andere digitale Technologien charakteristisch – sind zentral für die realen Auswirkungen von ML-Anwendungen auf soziale Welten und deren biologische, geologische wie klimatische Umwelten auf dem Planeten Erde. »Your computer« – so ›smart‹ er auch erscheinen mag – »is on fire« (Mullaney et al. 2021), und dies nicht erst seit gestern.

Maschinenlogik und ihre Folgen

Unser Fokus in diesem Band liegt auf theoretisch herausfordernden und mitunter bislang unterbelichteten Aspekten maschinellen Lernens. Uns interessieren insbesondere jene Merkmale des Phänomens, die es als eigenständigen Forschungsgegenstand auszeichnen; dies im Gegensatz zu Facetten von ML-Anwendungen, die wesentlich strukturell prädeterminiert sind und sich etwa primär aus den wirtschaftlichen und politischen Dynamiken erklären, die den ›KI-Hype‹ der letzten zehn Jahre angetrieben haben. Von derartigen Auseinandersetzungen versprechen wir uns Perspektiven, die ein erweitertes

Verständnis und neue Potenziale für den Umgang mit ML-basierter Technik aufscheinen lassen. Dieser Band stellt eine Reihe solcher Perspektiven vor.

Einige der Beiträge des Bandes befassen sich mit Fragen nach einem angemessenen Verständnis der sich in ML-Verfahren manifestierenden »computational reason« (Cavia 2022) sowie den praktischen Auswirkungen ihrer realen Umsetzung. Über theoretische Bemühungen um eine begriffliche Abgrenzung dieses Konzepts von gängigen, in der Regel anthropozentrischen Verständnissen von Vernunft hinaus finden sich ebenso Analysen der praktischen Rolle und Bedeutung konkreter materieller Manifestationen ›komputationeller Vernunft‹, insbesondere im Hinblick auf ihre nichtintendierten Nebenfolgen. Effekte, die von den Absichten und Erwartungen der Anwendungsentwickler:innen und -nutzer:innen abweichen, sind angesichts der Vielzahl und Größenordnung von Kontexten, in denen ML-Anwendungen stattfinden, keine Seltenheit, sondern ein wesentliches Charakteristikum (vgl. Broussard 2018). ML-Anwendungen haben ihren Anteil an der Emergenz neuartiger kultureller Dynamiken, die nicht zuletzt auf dem »epistemischen Schock« beruhen, den die Eigenlogiken maschinellen Lernens im Zuge ihrer Verbreitung gerade deshalb auslösen, weil sie zur Vermittlung und Produktion symbolvermittelter Sinngehalte bzw. von Bedeutung eingesetzt werden (vgl. Roberge/Castelle 2021: 2, 7).

Unser Interesse gilt insbesondere Entwicklungen im Zusammenhang mit ML-Anwendungen, die verändernd auf Handlungs-, Organisations- oder Denkweisen, mithin auf die Möglichkeitsbedingungen menschlicher Vergesellschaftung einwirken. In seiner Keynote im Rahmen jener Tagung des Schaufler Lab@TU Dresden Ende 2021, auf die dieser Band zurückgeht, schlug Matteo Pasquinelli vor, maschinelles Lernen als Wissensmodell im Sinne einer politischen Epistemologie zu verstehen. Seinen Ausführungen zufolge trägt ML zur Errichtung eines »epistemischen Gerüsts« bei, das die ideologische Form, die logische Form, die technische Form und die soziale Form in einen integrierten Zusammenhang bringe. Pasquinellis Vortrag unterstrich die Multidimensionalität maschinellen Lernens als ein gesellschaftliches Phänomen, das sich aus dem dynamischen Zusammenspiel von Mythologie, kollektiver Imagination, statistischem und mechanischem Denken, Rechenprozessen und der Automatisierung von Arbeit sowie der Überwachung und Kontrolle von sozialem Verhalten speise. Als realweltlicher Zusammenhang stelle es uns vor teils grundlegende Veränderungen in der Art und Weise, wie Wissen produziert, vorgestellt, navigiert und verteilt wird.

Diese Veränderungen haben einschneidende praktische Konsequenzen. Wie Louise Amoore (2020: 40f.) ausführt, sind ML-Anwendungen dazu in der Lage, subtile, manchmal gänzlich unsichtbare Veränderungen in der menschlichen Weltwahrnehmung zu bewirken. Ihr zufolge verschieben algorithmenbasierte experimentelle Datenverarbeitungsverfahren die Schwellen des für menschliche Beobachtung Zugänglichen. ML-basierte »Computer Vision« könne zwar nicht selbst im eigentlichen Sinne ›sehen‹, doch spiele sie eine wichtige Rolle dabei, was für Menschen sichtbar ist. Die Weltwahrnehmung hänge daher zunehmend davon ab, was von einem ML-vermittelten Sichtbarkeitsregime für menschliche Beobachtung verfügbar gemacht wird. Diese Überlegung zielt weniger auf den Generalverdacht, dass mit Computer Vision in erster Linie ein Verschleierungsinstrument am Werk sei, vielmehr geht es Amoore um die Feststellung, dass technisch vermitteltes Sehen (und mithin auch Erkenntnis im weiteren Sinne) das Ergebnis selektiver Prozesse ist, die zu kontingenten und mitunter schwer verstehbaren Resultaten führen.

Maschinelles Lernen als theoretisches Problem

Unser Interesse an ML ergibt sich ebenso aus dem Umstand, dass die realweltlichen Effekte von dessen Anwendung die Ausrichtung einiger Grundbegriffe der Geistes- und Sozialwissenschaften infrage zu stellen vermögen. Betroffen sind davon solch zentrale Konzepte wie Subjektivität, Interaktion, Kognition oder Kommunikation. Wie Luciana Parisi (2019) feststellt, handle es sich bei KI um ein »alien subject«, das zugleich unsere Vorstellung davon, was unter menschlichen Subjekten zu verstehen sei, auf die Probe stelle. Die zunehmende Automatisierung kognitiver Arbeit, bei der ML-Verfahren eine zentrale Rolle spielen, problematisiere das menschliche Selbstverständnis. Seinem Privileg der alleinigen Entscheidungsgewalt in vielen sozialen Zusammenhängen beraubt, werde das menschliche Subjekt Zeuge der Ausdehnung eines »alien space of reasoning« (ebd.: 30), die in einer »crisis of conscious cognition« (ebd.: 28) resultiere. Und die mit »machine thinking« einhergehenden Unwägbarkeiten vervielfachen sich gewissermaßen, wenn vernetzte Maschinen in einem ihnen eigenen »space of communication opaque to human vision« (ebd.: 31) aktiv werden. Fraglich wird vor diesem Hintergrund nicht nur, was ›Denken‹ impliziert, welche Rolle es in sozialen Prozessen spielt und wessen Privileg es ist, zu denken, sondern auch geläufige Vorstellungen von Kommunikation geraten unter den Verdacht, irreführenderweise zu anthropomorph konzipiert zu sein.

Elena Esposito (2022) versucht mit ihrer systemtheoretischen Konzeption »künstlicher Kommunikation« den merkwürdigen Charakter der Begegnung von Menschen und ML-basierten sozialen Agenten analytisch adäquat zu beschreiben. In ihrer Keynote auf unserer Tagung im Dezember 2021 führte sie die wesentlichen Grundannahmen dieses Ansatzes aus. Es sei, so Esposito, eine fehlgeleitete Vorstellung, dass ML die Eigenschaften menschlichen Verhaltens, Denkens oder Kommunizierens reproduziere. Stattdessen seien dessen Anwendungen gerade dann am erfolgreichsten, wenn sie für Aufgaben eingesetzt werden, in denen es nicht um die Imitation menschlichen Bewusstseins oder Denkens geht. ML-Technik könne sich an Kommunikation beteiligen, indem sie Kontingenz »virtuell« verarbeite. Wenn Nutzer:innen mit ›ausgereiften‹ Bots interagierten, sähen sie sich mit einer Kontingenz konfrontiert, die sie sich nicht selbst zurechnen können; andererseits lasse sich das kommunikative Geschehen nicht hinreichend als alleinige Projektion der User:innen erklären. Zugleich, so Esposito, stünden ML-Beiträge ebenso wenig für die Kontingenz der Maschine, da diese allein auf fremde Referenzen verweise. Ein ML-Modell wäre nichts ohne einen (auf menschlicher Kommunikation beruhenden) Satz von Trainingsdaten, deren Bedeutung und Ursprung oft unklar seien. Das Modell präsentiere den Nutzer:innen intransparent verarbeitete und daher auf kaum nachvollziehbare Weise reflektierte Perspektiven anderer Nutzer:innen, mithin deren Kontingenz, die vom Modell verarbeitet und damit virtualisiert werde. Virtuelle Kontingenz konstituiert Esposito zufolge daher einen Modus von Kommunikation, der insofern ohne distinkte Alterität operiert, als sich hinter ML-Beiträgen nicht ohne Weiteres eine kompakte Adresse ausmachen lässt, der etwa Autor:innenschaft und damit verbundene Absichten oder Verantwortung zugeschrieben werden könnten. Dies deutet auf eine Verschiebung der Grundbedingungen von Sozialität hin. »Artificial Communication« steht für einen sozialen Operationsmodus, der bisher als Privileg der Beziehungen zwischen Menschen galt: sprachbasierte Kommunikation, aber eben »künstliche«, wie Esposito es ausdrückt – ein Attribut, das hier in erster Linie für das Fehlen eines vertrauten Elements von Kommunikation steht, nämlich einer Adresse, der Kommuniziertes als Handlung zugeschrieben werden kann.

Es zeigt sich hier ein doppeltes Problem, denn die Verstehbarkeit von Kommunikation hängt allgemein davon ab, dass sie einer kompakten sozialen Adresse – etwa einem menschlichen Individuum – zugerechnet werden kann, was wiederum Voraussetzung für die Möglichkeit der Zuschreibung von Verantwortung ist. Adressierbarkeit – in technischer Hinsicht ein wesent-

liches Merkmal aller Computertechnik (vgl. Dhaliwal 2022) – ist angesichts der ML-spezifischen Opazität auf sozialer Ebene demzufolge keine Selbstverständlichkeit mehr (vgl. Burrell 2016). Es verwundert daher nicht, dass viele Bemühungen in der KI-Forschung dem Anliegen gewidmet sind, diese »responsibility gap« (Matthias 2004) zu schließen. So soll etwa das Problem mangelnder Transparenz durch die Entwicklung von »Explainable-AI«-Methoden und -Verfahren gelöst werden, die den Bedürfnissen der an den ML-Anwendungen beteiligten »stakeholders« nach einer besseren Interpretierbarkeit der Outputs gerecht zu werden versuchen (vgl. Zednik 2021). Bei allen Fortschritten in diesem Forschungsfeld ist bis heute bei vielen – wenn nicht den meisten – ML-Anwendungen selten klar, was ein Modell dazu veranlasst hat, einen bestimmten Output zu erzeugen. Diese womöglich konstitutive Lücke lässt Raum für Projektionen und Spekulationen. KI – weniger in der gegenwärtigen, deutlich häufiger in antizipierten, zukünftigen Formen – wird nicht selten ein sublimer Status zugeschrieben (vgl. Ames 2018); man denke nur an die verschiedenen Varianten des »Superintelligenz«-Diskurses (vgl. etwa Bostrom 2014). Andere Autor:innen wiederum amüsieren sich angesichts der Banalität des Scheiterns vieler ML-Anwendungen über Spekulationen nahender technischer Singularität; Anekdoten über »Artificial Unintelligence« (Broussard 2018) oder »Artificial Stupidity« (Steyerl 2017; vgl. auch Mackinnon 2017) gibt es zuhauf.

Wir sehen diese vielfältigen, sich teils widersprechenden Positionen mit Blick auf die in diesem Band versammelten Annäherungen an verschiedene Aspekte maschinellen Lernens als Ausdruck von dessen Vielschichtigkeit und Ambivalenz als Forschungsgegenstand. In diesem Sinne scheint es naheliegend, ML als eine Chimäre zu verstehen, deren Wesen jenseits konventionell geltender Dichotomien liegt, wie Ilan Manouach und Anna Engelhardt (2022: 9) einleitend über das Ungeheuer aus der griechischen Mythologie schreiben, das für die von ihnen vorgelegte »Inventory of Synthetic Cognition« titelgebend ist. Maschinelles Lernen kann gleichzeitig Verarbeitung von Daten, Produktion von Sinn, Verhandlung sozialer Normen und Machtausübung meinen. Es findet in so vielen Bereichen Anwendung, nimmt dabei so variable Gestalten an und erfüllt derart unterschiedliche Funktionen, dass es angesichts der Vielfalt all dieser Aspekte naheliegt, grundlegende Fragen seiner Bestimmung als Forschungsgegenstand zu reflektieren. Es ist uns mit diesem Band ein Anliegen, einen Schritt zurückzutreten, um ML in realweltlich eingebetteter Form im Hinblick auf die benannten Herausforderungen gegenstandsangemessen theoretisch begegnen zu können.

ML-basierte Technik hat sich in so vielen Bereichen verbreitet und ist ein solch integraler Bestandteil des gesellschaftlichen Lebens geworden, dass sie in ihrer weitgehend selbstverständlichen und zumeist als unproblematisch erlebten Allgegenwärtigkeit heutzutage »unsichtbar« zu werden scheint. Eine solche »Transparenz«, um eine Einsicht aus Susan Leigh Stars und Karen Ruhleders (2017) instruktiver Forschung zu »großangelegten Informationsräumen« aufzugreifen, ist für die soziale Funktionalität jeder technischen Infrastruktur charakteristisch – solange sie funktioniert. Der zunehmend infrastrukturelle Status von ML-Technik spricht für ihre weitgehende Akzeptanz, was angesichts der damit einhergehenden Normalisierungs- und Habitualisierungsprozesse die Frage aufwirft, wie ML als Gegenstand kritischer Beobachtung und Analyse – Voraussetzungen jedweder adäquaten Theoriebildung – verfügbar gemacht und gehalten werden kann.

Dimensionen maschinellen Lernens: Modelle – Praktiken – Topologien

Für einen angemessenen theoretischen Umgang mit den oben dargelegten Herausforderungen schlagen wir einen analytischen Rahmen vor, der drei verschiedene Dimensionen der geistes- und sozialwissenschaftlichen wie philosophischen Auseinandersetzungen mit ML identifiziert. Die Unterscheidung von Modellen, Praktiken und Topologien ermöglicht es uns, einige wesentliche Aspekte der realweltlichen Einbettung und Situierung von ML in den Blick zu nehmen. Wir nutzen dieses Differenzierungsschema als den vorliegenden Band strukturierende Heuristik, um Besonderheiten der in den Beiträgen entwickelten Perspektiven hervorzuheben. Gleichermaßen erlaubt uns dieses Schema, Gemeinsamkeiten der unterschiedlichen Ansätze zu unterstreichen, die womöglich auf verallgemeinerbare Charakteristika des Phänomens verweisen.

Die Betrachtung von *ML als Modell* – bezogen auf dessen empirische Beobachtung wie auch auf die Reflexion des sich darin vollziehenden Umgangs mit Wissen – lenkt den Blick auf epistemologische Implikationen. ML modelliert realweltliche Phänomene datenförmig und vermittelt damit die Realitäten, deren Teil es zugleich selbst ist. Um die zentrale Bedeutung dieses Aspekts auf eine pointierte Formel zu bringen: »The model is the message« (Bratton/Agüera y Arcas 2022). Nach den spezifischen Möglichkeitsbedingungen und Implikationen von ML-basierter Modellbildung zu fragen heißt, sich wortwörtlich dem

Charakter des ›Lernens‹ wie auch dessen ›maschineller‹ Logik theoretisch anzunähern.

Die Auswirkungen von ML-Anwendungen hängen andererseits von der konkreten praktischen Situierung der Modelle ab (vgl. Groß/Wagenknecht 2023). *ML als Praxis* erfordert eine eigenständige Betrachtungsebene, die zwar mit den Charakteristika der Modelle verknüpft ist, über diese jedoch gleichwohl hinausweist. Maschinelles Lernen als Praxis zu verstehen, erlaubt Einblicke in die Differenz zwischen Modellannahmen und ihrer praktischen Umsetzung. Es ist genau diese Unterscheidung, anhand derer sich beschreiben lässt, was Einbettung von ML in bestehende kulturelle, wirtschaftliche und politische Realitäten heißt und welche Dynamiken und Interdependenzen dabei zum Tragen kommen.

Im Bemühen um ein adäquates Verständnis der oben beschriebenen nichtmenschlichen und mithin »alien« (Parisi 2019) Gegenstandseigenschaften erachten wir schließlich *ML-Topologien* als eine dritte wichtige Analysedimension der Theoriebildung. Rechenoperationen in vieldimensionalen Vektorräumen – in denen sich Modelle als Ausdruck der Wechselbeziehungen zwischen einzelnen Datenpunkten herausbilden – haben ihre eigenen Dynamiken und folgen spezifischen Logiken. Ein angemessenes Verständnis ihrer Eigenschaften kann vermitteln, was das »Denken« oder die »Kognition« mutmaßlich intelligenter Maschinen ausmacht. Tatsächlich könnte der Vektorraum in rechnerischer Hinsicht als das Medium maschinellen Lernens verstanden werden, da seine Eigenschaften die Möglichkeitsbedingungen von ML-Technik bestimmen. Aus diesem Grund sehen wir in der Auseinandersetzung mit »computational topologies« einen wesentlichen Ansatzpunkt, um die Annahme, dass die operative Logik von ML-Verfahren prinzipiell unverstanden bleiben müssen, überwinden zu können. Wir sind zuversichtlich, dass sich ein angemessenes Verständnis der topologischen Implikationen maschinellen Lernens produktiv auf geistes- und sozialwissenschaftliche wie philosophische Auseinandersetzungen mit KI auswirken wird.

Praktische Relationalität, strukturelle Interdependenz, kooperative Generativität

Die Beiträge des Bandes durchziehen – über die drei Betrachtungsdimensionen Modell, Praxis und Topologie hinweg – mehrere Leitmotive zur Charakterisierung der realweltlichen Einbettung maschinellen Lernens.

Ein erstes wiederkehrendes Motiv ist die Betonung der Relationalität heterogener sozialer Entitäten in ML-Praktiken. Die wechselseitige Beeinflussung von Menschen, Maschinen, Dingen, Infrastrukturen und anderen für die jeweilige Praxis relevanten sozialen Entitäten erfolgt jeweils unter konkreten situativen Bedingungen, die spezifische Dynamiken mit sich bringen. **Hannah Link** (S. 143-168) stellt mit Blick auf den Umgang von Robotikforscher:innen mit ihren Prototypen fest, dass sich nicht nur die Leistungsfähigkeit von Robotern weiterentwickelt, sondern sich dabei auch eine Form von »posthumane[r] Interaktion zwischen Wissensobjekt und -subjekt« (S. 143) entsteht, die sich im Rückschluss auf die Vorstellungen von Menschlichkeit auswirkt. Der Frage nach einem adäquaten theoretischen Ansatz zur Beschreibung der Entwicklung von Kommunikationsfähigkeit zwischen Menschen und Maschinen begegnen **Yaoli Du und Nadine Schumann** in ihrem Beitrag (S. 193-207) mit dem Vorschlag, die interaktive Herausbildung gemeinsamer pragmatischer Handlungsmuster in einer geteilten Welt als zentral zu betrachten. Auch **Jonathan Harth und Maximilian Locher** nehmen mit dem Programm der Relationsmuster praktische Beziehungsaspekte von Mensch und Maschine in den Blick und beleuchten, wie diese auf die Identitäten der beteiligten Entitäten zurückwirken und sie kontinuierlich neu generieren (S. 169-191). Ein empirisches Beispiel für ein solches interaktives Verhältnis ist der »e-rater«, den **Jan Georg Schneider und Katharina Zweig** untersuchen (S. 93-111), um anhand dieses ML-basierten Bewertungssystems die Potenziale und Kontingenzen der automatisierten Beurteilung von Texten auszuloten. Die Beiträge begegnen sich im gemeinsamen Bezug auf postanthropozentrische Formen des Sozialen und explorieren Erweiterungen von Sozialtheorien angesichts neuer mit ML verbundener Problemlagen.

Die kritische Reflexion der Interdependenzverhältnisse innerhalb der wirtschaftlichen und politischen Zusammenhänge, in denen ML-basierte Technologien entwickelt und eingesetzt werden, stellt ein weiteres Leitmotiv dar, das wiederkehrend thematisiert wird. Modellbasierter ML-Technik – etwa automatisierten Sprachverarbeitungssystemen – sind auf folgenschwere Weise Herrschaftsverhältnisse eingeschrieben. Diese Beobachtung nimmt **Christian Heck** zum Anlass, *adversarial hacking* als subversive ästhetische Strategie für die partizipative Öffnung von hegemonial strukturierten KI-Sprachmodellen praktisch vorzustellen (S. 235-286). **Jan Fuhrmann** nähert sich dem Problem systemtheoretisch und findet eine Übersetzungslücke zwischen den für die Grammatiken der Diskriminierung blinden autopoietischen algorithmischen Systemen und den ihrer Umwelt entspringenden semanti-

schen Gehalten (S. 115-141). Mit einem kritischen Blick auf die Entwicklung sogenannter Large Language Models (LLMs) durch Google/Alphabet verhandeln **Jonathan Roberge und Tom Lebrun** die hermeneutischen Implikationen der jüngeren Entwicklungen im Feld automatisierter Textgenerierung sowie die Machtverhältnisse, innerhalb derer diese Entwicklungen stattfinden (S. 39-65). Die Autoren kommen zu dem Schluss, dass die ökonomischen Rahmenbedingungen kaum Anreize für eine hinreichende Sorgfalt bei der qualitativen Kuratierung von Datensätzen schaffen und stattdessen die schnellstmögliche Skalierung von Sprachmodellen befördern, was häufig eine Verstärkung von nichtintendierten Effekten – nicht selten solchen diskriminierender Art – zur Folge hat. Einen Versuch, die Schnittstellen von Ethik, Epistemologie und Politik zu bestimmen, unternimmt **Catriona Gray** am Beispiel von Empfehlungssystemen, die zur Unterstützung staatlicher Entscheidungspraxis genutzt werden (S. 67-92). Ihr Beitrag untersucht die Umstände und Konsequenzen des Einsatzes von ML-Anwendungen in der öffentlichen Verwaltung und erörtert, wie dadurch vorherrschende Gerechtigkeitskonzepte auf den Prüfstand gestellt werden.

Ein dritter Aspekt, den die Beiträge des Bandes aufgreifen, ist die praktische kooperative Generativität von Menschen und ML-Verfahren. Dieses Leitmotiv greift Diskussionen zur Autor:innenschaft in soziotechnischen Systemen auf und lenkt den Fokus auf spezifische Dynamiken, die kreative Prozesse unter Beteiligung von ML-Technik kennzeichnen. **Miriam Akkermann** nähert sich dem Zusammenhang anhand einer Untersuchung der musikalischen Freiräume, die die Einbindung unterschiedlicher Formen von KI in Kompositionsprozesse erlaubt (S. 315-329). Der Beitrag von **Jakob Claus und Yannick Schütte** (S. 211-233) widmet sich den Dynamiken in der literarischen Ko-Produktion von Autor:in und automatischem Textgenerierungsverfahren am Beispiel des Buches *Pharmako-AI*, das K Allado-McDowell (2021) gemeinsam mit dem Sprachmodell GPT-3 verfasst hat. **Michael Klipphahn-Karge** untersucht die Nuancen verschiedener maschineller Ästhetiken aus kunstwissenschaftlicher Perspektive und entwickelt ein analytisches Kategorienschema zur differenzierten Auseinandersetzung mit sogenannter KI-Kunst (S. 287-314). **Fabian Offert** schlägt mit Blick auf zeitgenössische ML-basierte Kunst vor, diese aufgrund des subtraktiven Charakters der bei ihrer Entstehung genutzten multimodalen Bildsynthese als Skulptur zu verstehen (S. 273-286).

Die letzten beiden Beiträge des Bandes setzen einer reduktionstischen Sicht auf »computational reason« Positionen entgegen, die Potenziale topologisch informierter Perspektiven auf das Verständnis der generativen Ei-

genlogik maschinellen Lernens erörtern. **AA Cavia und Patricia Reed** (S. 351-363) folgen ausgehend von Ansätzen der konstruktivistischen Mathematik der topologischen Annahme, dass unter der nichteuklidischen Voraussetzung der Möglichkeit räumlicher Pluralität jeder Raum eine in ihn eingebettete begleitende Struktur mit sich führt, ohne dieser vorausgehen zu müssen. Die Autor:innen loten sodann die Möglichkeiten eines begrifflichen Vokabulars aus, das die Domänen des Diskreten und des Kontinuierlichen – der Algebra und der Geometrie – zusammenführt, um eine Erweiterung des »inferential toolkit available to computational reason« (S. 361) zu ermöglichen. **Lukáš Likavčan und Carl Christian Olsson** (S. 333-349) setzen sich in ihrer topologischen Annäherung an ML kritisch mit Immanuel Kants Analogie zwischen der Orientierung im Denken und der geografischen Orientierung auseinander. Ihr Aufsatz nimmt die topologischen Merkmale von künstlichen neuronalen Netzen zum Anlass, unser Verständnis von Denken mit Blick auf die Möglichkeiten der sich in Deep Learning-Verfahren manifestierenden räumlichen Konstellationen zu reflektieren.

* * *

AI Realities

The title of this volume alludes to the impact that technologies based on machine learning (ML) have had across countless domains of social life in recent years. *AI Realities* also hints at the reflexive challenges in endeavors to theorize so-called artificial intelligence (AI) as a real-world phenomenon.[2] We chose this title as an expression of the ambivalent tensions inherent to ML both as a transformative technology and as a challenging research subject. ML's increasing practical integration into social life necessitates focusing on its varied and partially conflicting characteristics as both a factor and a product of the realities in which it is embedded. Real-world embeddedness implies relationships of mutual influence and attunement and hence asks us to consider how ML is shaped as a codependent entity in relation to other entities and structures. This volume is dedicated to exploring the theoretical challenges that stem from the difficulties in understanding ML in terms of its multifaceted ways of appearing in the world. Its chapters are based on talks given at an interdisciplinary conference organized by Schaufler Lab@TU Dresden that took place online in December 2021.

ML's various forms of real-world appearance include speculations as well as practical applications. Its impact is driven by its functional capacities as a technology just as much as by its potential for imagination. In this vein, we approach theorizing ML carefully in a way that avoids misattributing properties to it that actually explain a different phenomenon or the dynamics of the interplay between the two. This points to the challenge of determining the nature of the relationship between AI and the realities that produce it. In applied form, ML has become a ubiquitous part of the realities it has emerged into, now contributing to the reproduction and change of those realities. In its capacity to make meaning via language or image processing and generation, ML can contribute to the world in generative ways and allows for the making of new worlds and new forms of worlding. On the one hand, this disposition implies that theorizing ML presupposes theories of the worlds it has started to inhabit. On the other hand, it suggests theories of ML in the humanities should pay close attention to its practical situatedness to grasp its relevance and effects within lived experiences. Such an approach puts research on ML in the humanities in

2 We use the terms machine learning and artificial intelligence interchangeably in the following – unless indicated otherwise – to describe the technology in question, given ML's dominance in AI research and application in the past decade.

a position to critically engage with conceptualizations of the subject in disciplines such as data science and engineering as well as the adjacent technology industries.

Drawing on these conceptualizations, ML – schematically put – relies upon the 'training' of algorithms to build a model through the processing of a dataset compiled to recognize patterns contained in the data. ML-based pattern recognition essentially produces stochastic data analyses. While these analyses are often well understood in terms of the formal patterns and regularities of a given dataset, they are frequently hard to grasp, if not unknowable, in terms of the domain the data represents (Dourish 2016, 7). Hence, the often surprising and "uncanny" output (Bucher 2017) produced by such models is not easy to make sense of. This is not least due to the opacity of the computational modeling: "When a computer learns […], it does so without regard for human comprehension" (Burrell 2016: 10). From this follows another challenge that sits in the adequate interpretation of ML's contributions to the world. ML-generated outputs can be harmful in their real-world effects. Therefore, the models frequently become subject to controversies, in part because a model's 'reasoning' can not be fully accounted for. ML's hard-to-grasp non-human characteristics continue to complicate its understanding in the humanities and beyond.

Pathologies of AI

Many researchers in the humanities tend to dismiss the relevance of ML's alienness and highlight instead how emphasizing its complexity mystifies the ideological nature of such rhetoric. They point out, for instance, that it basically stands for "a fancy way of calculating an average" and is, in effect, a "total scam" (Galloway 2021) that distracts from the real challenges resulting from the proliferation of applications associated with "AI". ML has not spontaneously emerged into the world and its real-world impact is, among other factors, driven by economic interests expressing the structures of a particular social order. It performs "knowledge extractivism" (Joler/Pasquinelli 2020) by generating information from the models its productivity depends upon. The knowledge contained in these models is based on datasets produced by human labor. ML, as a set of material practices, takes the shape of an assembly line (p. 2), resembling a form of organizing labor typical of industrial capitalism precisely because it relies on that labor. ML often depends on laborious efforts that allow data to be gathered and compiled into a dataset from which a model is generated that can finally be applied for a dedicated purpose. In this sense, ML

"is an accumulation rather than a radical transformation" (Mackenzie 2017: 5) – its social embeddedness implies dependence on and integration into the social forms it appears in.

From this, it follows that ML leaves its mark on reality by often reproducing, if not amplifying, some of the latter's pathological political and ideological tendencies. Such tendencies include the discrimination against and exclusion of marginalized people (Noble 2018; Apprich et al. 2018; Chun 2021), the concentration of power in the hands of oligarch-owned platforms (Whittaker 2021), the rise of networked, data-centric forms of social control, constituting a distributed "polyopticon" (Sherman 2022), and the disregard of environmental issues (Strubell/Ganesh/McCallum 2019; Dhar 2020; Bender et al. 2020). These problems, though not exclusive to ML but characteristic of many computational technologies, are central to its impact on social worlds and the planet. "Your computer" – however intelligent it may seem – "is on fire" (Mullaney et al. 2021), and has been for a long time.

Machine logic and its consequences

Our specific focus in this volume is on the challenging and perhaps hitherto neglected aspects of ML that distinguish it as a genuine research subject in and of itself, as opposed to one whose characteristics are essentially predetermined by structures that precede it, e.g., the political economy that has given rise to the recent 'AI Hype' of the past ten years. We are interested in the theoretical challenges that ML confronts us with and how we can approach their resolutions – which includes identifying affordances implied in new perspectives. This volume presents a variety of such approaches and continues discussions that started at our conference in December 2021.

Some of the contributions explore issues related to aspects of "computational reason" (Cavia 2022) as manifested in ML and how they play out in practice. Beyond philosophical efforts to distinguish this notion from conventional conceptions of reasoning, this includes assessing the consequences of ML's practical deployment specifically with regard to unintended consequences in its use. To the extent that it is employed to generate and distribute information, and hence meaning, effects that differ from the intentions of the technology's makers and users abound (Broussard 2018). ML, in some respects, introduces new cultural dynamics based on the "epistemic shock" it brings about, based on its "claim to meaning itself" (Roberge/Castelle 2021: 2, 7).

Machine learning as a theoretical problem

Our interest is in these new ways in which ML as a computational technology feeds back to human ways of acting, organizing, classifying, and reasoning. In his keynote at our conference, Matteo Pasquinelli suggested conceiving of ML as a knowledge model in terms of a political epistemology. According to his conceptualization, there is an "epistemic scaffolding" at work that simultaneously connects ideological form, logical form, technical form, and social form. Pasquinelli's talk highlighted ML's multidimensionality as a real-world phenomenon. Having emerged from the dynamic interplay of mythology, collective imagery, statistical and mechanical thinking, computation, and the automation of labor, as well as surveillance and control of social behavior, ML presents us with a shift in how knowledge is produced, navigated, distributed, and modeled.

This shift has practical consequences. As Louise Amoore points out concerning "apparently coherent technologies of observation" (2020: 40), ML is capable of causing subtle and sometimes invisible changes down to the level of human perception and sensorium. She argues that "experimental algorithmic techniques" act "in and through data to modify the threshold of perceptibility itself" (p. 41). While ML-based "computer vision" cannot actually 'see', it does play a significant part in *what* becomes visible, and hence intelligible, to human observers. How humans perceive the world increasingly depends on what algorithmic systems make available for observation. This is not to argue per se that there is obfuscation at work. Rather, it reveals that technologically mediated vision results from a selective process that can lead to contingent, potentially confusing outcomes.

Our interest in ML is stimulated by how difficult it seems to grasp in theoretical terms and how its real-world impact, in turn, challenges some quite elementary concepts of the humanities, such as notions of cognition, subjectivity, and communication. As Luciana Parisi observes, AI is an "alien subject" (2019) that questions what we consider to be human subjects. To the extent that it plays a major part in the automation of cognitive tasks, ML problematizes our understanding of ourselves. Deprived of its privileged access to decision-making, the human subject experiences the "expansion of an alien space of reasoning" which corresponds to a "crisis of conscious cognition" (pp. 30, 28). And "machine thinking" gets multiplied, so to speak, when connected machines become active in their own "space of communication opaque to human vision"

(p. 31). Therefore, it is not only thinking but also communication that becomes questionable in its human-centered theoretical form.

Elena Esposito introduces the concept of "artificial communication" (2022) to theorize the strange encounters between humans and ML-based agents. In her keynote at our conference, she spelled out the implications of this concept. It would be wrong to assume, Esposito argued, that ML reproduces the characteristics of human behavior, reasoning, or communication. Instead, it is most successful at tasks where it is not deployed to imitate human consciousness or reasoning (p. 2). ML can take on the form of a social agent by processing "virtual contingency". When users interact with ML-based agents, such as language models, they face a contingency that is not their own nor is it really the contingency of the machine because the model's output is entirely based on (often human-generated) training data whose meaning and actual origin are in themselves contingent (pp. 9–10). Hence, as Esposito explained in her talk, the algorithmic machine presents the users with processed – or reflected, so to speak – perspectives of other users and their contingent perspectives. She concluded that virtual contingency constitutes a mode of communication without distinct alterity. This indicates a shift in the very conditions of sociality. ML potentially allows humans to engage with machines in ways that were previously considered a privilege of human-to-human relationships. It is language-based communication, but "artificial" – to use Esposito's term here to indicate its lack of a familiar element, namely an address to which one can attribute received information.

The trouble here is twofold, as the possibility of the interpretability of ML depends on its addressability as a distinct entity which, in turn, is a precondition for the attribution of accountability. Addressability, a core feature of computational technologies in general (Dhaliwal 2022), is not a given in the face of troubles with opacity in ML (Burrell 2016). There have been notable research efforts aimed at closing this "responsibility gap" (Matthias 2004) and at solving issues related to opacity by the development of methods directed toward "explainable AI" to satisfy the explanatory requirements of different stakeholders involved in ML applications (Zednik 2021). Yet to this day, in many, if not most applications of ML, it is not always clear what caused a model to produce a certain output. This seemingly constitutive gap leaves room for projections. Some have ascribed a sublime status to ML (Ames 2018) – think of the various versions of the discourse on "superintelligence" (Bostrom 2014) – while others have ridiculed it for its actual triviality in the face of the apparent rise of singu-

larity; anecdotes of "artificial unintelligence" (Broussard 2018), if not outright "artificial stupidity" (Steyerl 2017; see also Mackinnon 2017) abound.

With regard to the approaches assembled in this volume, we see these diverse and ambivalent ideas of and views on ML as expressions of the subject's multifaceted social nature. In this sense, perhaps it is best to think of it as a chimera, a "trickster of the natural order" whose nature is beyond "imposed dichotomies", as Ilan Manouach and Anna Engelhardt write in the introduction to their *Inventory of Synthetic Cognition* (2022: 9). ML can simultaneously be a way of processing information, of generating meaning, of negotiating social norms, and of exercising power. It is not a limited-range, domain-specific technology. Rather, it extends into myriad fields and takes on many forms so that ultimately, given the variety of its potential functions in the world, we feel a need to reflect on questions concerning its conceptualization.

By taking a step back, we intend to examine how ML can be plausibly theorized in the face of the challenges relating to its real-world embeddedness. In summary, these characteristics include its ambivalent nature, owing to the dynamics of its entangled situated relationships with other social entities (Groß/Wagenknecht 2023), the alien nature of its (partly pathological) contributions to the world, which are sometimes difficult to account for, and its invisibility as well as its questionable addressability, which leave space for speculation and projection. ML-based technology has become such an integral part of social life that it seems to be becoming "invisible" in its largely unquestioned ubiquity in lived experiences. Such "transparency", to reiterate a notion from Susan Leigh Star and Karen Ruhleder's formative research on "large information spaces" (Star/Ruhleder 1996), is characteristic of any social infrastructure. Acknowledging the infrastructural status of ML-based technologies speaks to their social acceptance which, in its normalizing and habitualizing effects, raises the question of how to maintain critical observation and analysis.

Dimensions of machine learning: models – practices – topologies

To approach ML's real-world embeddedness, we suggest a tripartite framework to highlight what we consider to be some of the most crucial dimensions of its theorization. We propose this schema as a heuristic to help relate different theoretical perspectives on ML to one another.

Viewing ML as *models* – of real phenomena as well as of knowledge (production) itself – highlights the epistemological stakes implied in its use. ML relies on models of worldly phenomena and is as much a product of this world as a

contributor to it: "the model is the message" (Bratton/ Agüera y Arcas 2022). To inquire into the specific conditions and characteristics of modeling is to inquire into ML's reason and epistemology.

The impact of its model-based contributions, on the other hand, depends upon how they are put to use *in practice*. The practical dimensions of technology in action, while related to the model aspects, deserve attention in their own right. It is through practices of machine learning that we can gain insights into its characteristics as a real-world phenomenon, including its embeddedness in existing cultural, economic, and political realities, its social and material interdependencies, and the dynamics of its production as well as its application across different domains.

Finally, in an effort to approach the non-human, alien characteristics inherent in ML as distinct from other technologies, we regard the *topologies* of its computational operations as an important aspect. In particular, we suggest that computational operations in many-dimensional vector spaces come with their own dynamics that reveal insights into what characterizes the 'thinking', 'cognition', or 'reasoning' of 'intelligent' machines. These operations represent the interrelations of the individual data points in a neural network. In fact, vector space can be considered the very medium of artificial neural networks in computational terms. Its properties set the conditions for, as well as the horizon of, their possibilities and limitations.

Practical relationality, structural interdependence, cooperative generativity

A recurring analytical theme across this volume is the relationality of social entities in ML practices. The mutual influence of humans, machines, infrastructures, institutions, and other material entities on how systems, actions, and communications gain shape is embedded within specific dynamics that result in more or less stable relationships. Observing how roboticists deal with their material prototypes, **Hannah Link** (pp. 143-168) notes that advances in this field of research are accompanied by specific modes of posthuman interaction between knowledge objects and knowledge subjects. These novel modes of interaction in turn also affect the researchers' perceptions of their own humanity. **Yaoli Du and Nadine Schumann** (pp. 193-207) address the question of an adequate theoretical approach for the development of communication between humans and machines, focusing on the interactive formation of shared pragmatic action patterns that allow heterogeneous social entities to integrate and

mutually attune. In a similar vein, **Jonathan Harth and Maximilian Locher** (pp. 169-191) examine the relational patterns in human–machine interaction and their implications for social theory. Inspired by systems theory and theories of social networks as well as ethnomethodology, they spell out how ML relates to the dynamics in the continuous ontogenesis of social identities. A detailed case study of one such human–machine relationship is presented by **Jan Georg Schneider and Katharina Zweig** (pp. 93-111) on the *e-rater*, a ML-based grading system used to predict test scores in the evaluation of written essays. Relying on the analysis of speech act felicity conditions, the authors analyze the potentials, limits, and contingencies in the automated grading of written text. All these chapters negotiate the practical relationalities of ML within different theoretical frameworks, yet complement one another in their emphasis on posthumanist perspectives on interaction in their search for possibilities to expand the anthropocentric focus that still characterizes many strands of social theory.

Another guiding theme of the various chapters results from critical reflection on the economic and political contexts in which ML-based technologies are developed and deployed. In such a view, ML does not appear to be an autonomous entity but rather an element of interdependent social structures. Examining political patterns of domination inscribed into ML-based systems of text generation, such as large language models (LLMs), **Christian Heck** (pp. 235-286) presents the practice of *adversarial hacking* as a subversive tool to tackle the conservative cultural tendencies and hegemonic ideological leanings of such systems. **Jan Fuhrmann** (pp. 115-141) approaches the problem of interdependence through systems theory and identifies a constitutive translation gap between algorithmic systems, which are blind to the social grammars of discrimination precisely because of their autopoiesis, which implies operational closure, and the structure of the datasets their models are based on produced by their environment. Focusing on the development of LLMs at Google/Alphabet, **Jonathan Roberge and Tom Lebrun** (pp. 39-65) highlight the hermeneutical implications of the rapid developments in the field of automated text generation and the power structures within which this development is taking place. The authors conclude that the economic framework of prestigious research projects at companies like Google provides little incentive to sufficiently care for the qualitative curation of datasets but rather encourages scaling to increase sales, ultimately resulting in a largely uncontrolled amplification of unintended adverse – and often discriminatory – effects. **Catriona Gray** (pp. 67-92) explores the intersection of ethics,

epistemology, and politics by analyzing examples of ML systems used in governmental decision-making practices. Her investigation shows which novel forms of rights or status violations – obviation, diminishment, impugnment – become pressing issues in the increasing entanglement of ML in sovereign decision-making and practices of administrative judgment.

A third and final noteworthy point of emphasis negotiated across different chapters of this volume is the cooperative generativity in creative practices featuring ML technologies. Touching on discussions regarding the authorship of technological systems of automation, the question of generativity revolves around the specific dynamics of practical cooperation in creative processes involving ML. **Miriam Akkermann** (pp. 315-329) approaches the topic by examining affordances, both real and imagined, that stem from the integration of different ML methods (and earlier forms of AI preceding them) in musical composition. In their contribution **Jakob Claus and Yannick Schütte** (pp. 211-233) explore the dynamics of literary co-production between human author K. Allado-McDowell and OpenAI's large language model GPT-3. This co-production culminated in the publication of the book *Pharmako-AI* (Allado-McDowell 2021), which is the subject of analysis here. **Michael Klipphahn-Karge** (pp. 287-314) addresses the ways in which ML, and AI generally, can become the subject of artworks. He suggests an analytical framework to distinguish the different ways in which contemporary artists refer to ML in their works – as a technology actively used to create an artwork, as a subject of critique, or as an occasion for the artist to develop their work. Likewise, regarding the conceptual questions surrounding "AI art", **Fabian Offert** (pp. 273-286) suggests that ML-based artworks can be conceived of as sculptures owing to the subtractive nature of the image generation methods involved in their making.

The last two contributions of the volume argue against misguided reductionist and trivializing views of ML, and instead, examine the potential of topologically informed understandings of the generative aspects of ML. Exploring the notion of "computational reason" (Cavia 2022) in terms of its topological implications, **AA Cavia's and Patricia Reed** (pp. 351-363) advance a concept of pointless topology based on constructive mathematics. Their argument considers the topological view that "all space comes with an attendant structure" (p. 353) and proceeds to spell out the affordances of a conceptual vocabulary that unifies the realms of the discrete and the continuous – the algebraic and the geometric – in order to expand computational reason's "inferential toolkit" (p. 361). Similarly, in their approach to reason via topology, **Lukáš Likavčan and Carl Christian Olsson** (pp. 333-349) critically evaluate Immanuel Kant's analogy

of orientation in thinking and geographical orientation to propose a topological account of thinking. They take the discussion of artificial neural networks' topological characteristics as an opportunity to reflect on the notion of human thinking in the mirror of the spatiality of so-called deep learning. As computational topologies are of crucial importance in making sense of the alienness and incomprehensibility of ML, an adequate understanding of their relevance and implications will, we hope, contribute to its theorization, and perhaps also to new approaches to the philosophy of computation.

Bibliography

Allado-McDowell, K. 2020. *Pharmako-AI*. UK: Ignota Books.

Ames, Morgan G. 2018. Deconstructing the algorithmic sublime. *Big Data & Society* 5(1):1–4. https://doi.org/10.1177/2053951718779194.

Amoore, Louise. 2020. *Cloud Ethics: Algorithms and the attributes of ourselves and others*. Durham and London: Duke University Press.

Apprich, Clemens, Wendy Hui Kyong Chun, Florian Cramer and Hito Steyerl. 2018. *Pattern Discrimination*. Lüneburg and Minneapolis, MN: meson press/ Minnesota University Press.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922.

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Bratton, Benjamin and Blaise Agüera y Arcas. 2022. The Model Is The Message. *Noema Mag*. https://www.noemamag.com/the-model-is-the-message/. Last access: 12 December 2022.

Broussard, Meredith. 2018. *Artificial Unintelligence: How Computers Misunderstand the World*. Boston, Mass.: MIT Press.

Bucher, Taina. 2017. The algorithmic imaginary: exploring the ordinary affects of Facebook algorithms. *Information, Communication & Society* 20(1): 30–44. https://doi.org/10.1080/1369118X.2016.1154086.

Burrell, Jenna. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society* 3(1):1–12. https://doi.org/10.1177/2053951715622512.

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

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

Dhaliwal, Ranjodh Singh. 2022. On Addressability, or What Even Is Computation? *Critical Inquiry* 49(1):1–27. https://doi.org/10.1086/721167.

Dhar, Payal. 2020. The carbon impact of artificial intelligence. *Nature Machine Intelligence* 2:423–425. https://doi.org/10.1038/s42256-020-0219-9.

Dourish, Paul. 2016. Algorithms and their others: Algorithmic culture in context. *Big Data & Society* 3(2): 1–11. https://doi.org/10.1177/2053951716665128.

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

Galloway, Alexander. 2021. Questions. Answers. http://cultureandcommunication.org/galloway/questions-answers. Last access: 12 December 2022.

Groß, Richard and Susann Wagenknecht. 2023. Situating machine learning – On the calibration of problems in practice. *Distinktion. Journal of Social Theory*. https://doi.org/10.1080/1600910X.2023.2177319.

Mackenzie, Adrian. 2017. *Machine Learners: Archaeology of a Data Practice*. Cambridge, Mass.: MIT Press.

Mackinnon, Lee. 2017. Artificial Stupidity and the End of Men. *Third Text* 31(5-6):603–617. https://doi.org/10.1080/09528822.2018.1437939.

Manouach, Ilan and Anna Engelhardt. 2022. Preface. In *Chimeras: Inventory of Synthetic Cognition*, Eds. Ilan Manouach and Anna Engelhardt, 9–13. Athens: Onassis Foundation.

Matthias, Andreas. 2004. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology* 6:175–183. https://doi.org/10.1007/s10676-004-3422-1.

Mullaney, Thomas S., Benjamin Peters, Mar Hicks and Kavita Philip (eds.). 2021. *Your Computer is On Fire.* Boston: MIT Press.

Noble, Safiya Umoya. 2018. *Algorithms of oppression: How search engines reinforce racism*. New York: NYU Press.

Parisi, Luciana. 2019. The Alien Subject of AI. *Subjectivity* 12(1):27–48.

Pasquinelli, Matteo and Vladan Joler. 2020. The Nooscope Manifested: AI as Instrument of Knowledge Extractivism. *AI & Society* 36:1263–1280. https://doi.org/10.1007/s00146-020-01097-6.

Roberge, Jonathan and Michael Castelle. 2021. Toward an End-to-End Sociology of 21st-Century Machine Learning. In *The Cultural Life of Machine Learning: An Incursion into Critical AI Studies*, Eds. Jonathan Roberge and Michael Castelle, 1–20. Cham: Palgrave Macmillan.

Sherman, Stephanie. 2022. The Polyopticon: A diagram for urban artificial intelligences. *AI & Society*. https://doi.org/10.1007/s00146-022-01501-3.

Star, Susan Leigh and Karen Ruhleder. 1996. Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces. *Information Systems Research* 7(1):111–134. https://doi.org/10.1287/isre.7.1.111.

Star, Susan Leigh and Karen Ruhleder. 2017 (engl. 1996). Schritte zu einer Ökologie von Infrastruktur. Design und Zugang für großangelegte Informationsräume. In Susan Leigh Star, *Grenzobjekte und Medienforschung*, Eds. Sebastian Gießmann and Nadine Taha, 359–401. Bielefeld: transcript.

Steyerl, Hito. 2017. The Nation-State System: „Gott ist doof." On Artificial Stupidity. *Now is the Time of Monsters. What Comes After Nations* (Haus der Kulturen der Welt, Berlin, 23–25 March 2017). https://soundcloud.com/hkw/now-is-the-time-of-monsters-2. Last access: 12 December 2022.

Strubell, Emma, Ananya Ganesh and Andrew McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3645–3650, Florence: Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/P19-1355.

Whittaker, Meredith. 2021. The Steep Cost of Capture. *interactions* 28(6):50–55. https://doi.org/10.1145/3488666.

Zednik, Carlos. 2021. Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence. *Philosophy & Technology* 34:265–288. https://doi.org/10.1007/s13347-019-00382-7.

Danksagung/Acknowledgments

Wir hoffen, dass die in diesem Band versammelten Aufsätze den Diskurs über maschinelles Lernen in den Geistes- und Sozialwissenschaften sowie der Philosophie inspirieren und produktiv anregen werden.

Diese Publikation wäre ohne die Hilfe von Freund:innen und Kolleg:innen nicht möglich gewesen. Wir möchten zuallererst Michael Klipphahn-Karge und Ann-Kathrin Koster für ihre Beiträge zur grundlegenden Konzeption und Organisation dieses Buchprojekts und der vorangegangenen Konferenz danken. Wertvolle Hinweise und Anregungen für die Redaktion dieses Bandes erhielten wir von Celia Brightwell, Valentin Golev, Johannes Haaf, Dr. Andreas Höntsch und Prof. Dr. Dominik Schrage, wofür wir sehr dankbar sind. Darüber hinaus möchten wir unseren Kolleg:innen am Schaufler Lab@TU Dresden, insbesondere Prof. Dr. Lutz Hagen, Dr. Anke Woschech und Jonas Wietelmann, für die kontinuierliche Unterstützung dieses Projekts danken. Dankbar sind wir ebenso für die reibungslose und produktive Zusammenarbeit mit Steffen Schröter und Margaret May, die im Korrektorat der Bandbeiträge mit ihrer präzisen Bearbeitung wesentlich zu deren Qualität beigetragen haben, sowie Arne Winter (Bureau Neue), der das Cover des Buches sehr kurzfristig und mit großer Flexibilität sowie Verständnis für unsere Wünsche gestaltet hat. Für die geduldige wie zielgerichtete Betreuung dieser Publikation gilt darüber hinaus den Herausgeber:innen der Reihe »KI-Kritik«, Prof. Dr. Bernhard Dotzler, Prof. Dr. Andreas Sudmann und Prof. Dr. Anna Tuschling, sowie Dagmar Buchwald, Katharina Kotschurin und Julia Wieczorek im transcript Verlag unser Dank. Abschließend möchten wir uns bei The Schaufler Foundation sowie der Technischen Universität Dresden für die großzügige Förderung dieses Projekts bedanken.

Berlin, im Dezember 2022
Richard Groß & Rita Jordan

* * *

We hope that the thoughts and ideas collected in this volume present inspiring and productive contributions to the discourse on machine learning in the humanities, social sciences, philosophy, and beyond.

This publication would not have been possible without the help of friends and colleagues. We thank Ann-Kathrin Koster and Michael Klipphahn-Karge for their crucial contributions to conceptualizing and organizing this book and the preceding conference. The volume benefited from valuable advice and feedback in terms both of content and the editing process from Celia Brightwell, Valentin Golev, Johannes Haaf, Dr. Andreas Höntsch, and Prof. Dr. Dominik Schrage, for which we are very grateful. Additionally, we thank our colleagues at Schaufler Lab@TU Dresden and especially Prof. Dr. Lutz Hagen, Dr. Anke Woschech, and Jonas Wietelmann for their continued support of this project. We are also grateful for the productive collaboration with Arne Winter of Bureau Neue, who designed the cover of the book at very short notice and with great flexibility in response to our requests, and to Steffen Schröter and Margaret May whose subtle copy editing has substantially improved the quality of this volume. Equally, our thanks go to the editors of the "AI Critique" series, Prof. Dr. Bernhard Dotzler, Prof. Dr. Andreas Sudmann, Prof. Dr. Anna Tuschling, and Dagmar Buchwald at transcript. Finally, we wish to thank The Schaufler Foundation and TU Dresden for their generous funding of this project.

Berlin, December 2022
Richard Groß & Rita Jordan

Embedded Models of Meaning

Parrots All the Way Down

Controversies within AI's Conquest of Language

Jonathan Roberge & Tom Lebrun

Abstract: *Today's deployment of automated semantic models such as Google's BERT or OpenAI's GPT-3 is a remarkable challenge for the inscription of hermeneutics at the very heart of the social sciences project. Artificial intelligence is indeed conquering language. There are three important implications. First, we must take the power and possibilities of such models seriously – that is, the recent history of technological advances in deep learning and the modi operandi of these interpreting machines, particularly their two-way reading and "Transformer" architecture. Second, a better comprehension is required of the type of understanding involved – mainly how the calculation of probability, thresholds and variation, for example, vectorizes language as if to parrot it back. Our analysis takes note of the dismissal by Google of the researcher Timnit Gebru, precisely around the text "On the Danger of Stochastic Parrots", to show how the value of natural language processing (NLP) models lies in the kind of world they put forward as well as in their reference to a precise context. Finally, this should help to circumscribe the current economic, political and even ethical aporias concerning these models, including the fact that the platforms developing them overlook crucial real-world effects of the way in which they advance the extraction, commodification and instrumentalization of meaning. Ultimately, it is this close link between meaning and the displacement of power centers that becomes the central issue of Critical AI Studies.*

1. Introduction

The year 2020 was marked by yet another – substantial – crisis at Google with the departure-firing of researcher Timnit Gebru over the submission of the paper "On the Danger of Stochastic Parrots: Can Language Models Be Too Big?" (Bender et al. 2021). The story went from anecdotal to scandalous when the

company asked for the article to be withdrawn or for the names of the Google employees who contributed to it to be removed. For Jeff Dean, Google's director of AI, the matter was settled in that the work in question "didn't meet our bar for publication" (Hao 2020). However, Gebru's rebuff and the large number of supporters who rallied to her cause – in this case, more than 2,000 of the company's employees signed a letter asking for demanding greater transparency in the management of its internal affairs (Wakabayashi 2020) – had not been factored in. So, who had demanded the withdrawal of the article and why exactly? This is a question in the form of an ultimatum that remains unanswered. "Timnit wrote that if we didn't meet these demands," Dean writes, "she would leave Google [...]. [W]e accept and respect her decision." The break-up was, to all intents and purposes, complete. On Twitter, the researcher expressed her dismay and called out her former boss: "@jeffdean I realize how much larger language models are worth to you now."[1]

It is precisely on this notion of worth, or value, that we want to focus here. The double meaning of the word indicates that it refers more or less distinctly to something economic as well as axiological. Indeed, the very ambiguity of the word and of its use in the tweet is what makes it emblematic of the most important issues surrounding the current deployment of natural language processing (NLP) models. While many have seen the Gebru scandal as a matter of ethics and labor relations, few have been willing to consider the question in its fullest sense and thus explore the ways in which it represents a fundamental hermeneutical issue. Yet the question deserves to be asked: have AI and the latest advances in deep learning enabled the development of (too) big, powerful and deep models? And what can such terms mean, apart from a technical point of view? Do meaning and textuality, interpretation and understanding not become (too) impoverished as a result of their automated processing?

We argue in this chapter that the emergence of models such as Google's BERT or OpenAI's GPT-3 is today a remarkable challenge for hermeneutic disciplines in general, and for the inscription of hermeneutics at the very heart of the social sciences project in particular. It is therefore not a question of denying the rise of modeling or interpreting machines – or even their power and scope – but of examining the conditions of their possibility and significance. To put it in a nutshell: the emergence of these hermeneutic machines is an opportunity to think afresh about what a critical hermeneutics can represent within

1 Timnit Gebru, Twitter, https://twitter.com/timnitGebru/status/1334345550095912961. Last access: 19 July 2021. Emphasis added.

the social sciences and how it can enter into dialogue with or serve as a basis for the development of Critical AI Studies (CAIS) (Roberge 2020; Pasquinelli/Joler 2020). Specifically, this implies taking seriously the history and modus operandi of these language models and the way in which their various problems crystallized in the form of diverging views as part of the wider implications of the Gebru–Google conflict. This in turn implies a better understanding of the type of meaning at stake; that is, above all, the type of textual world deployed – or not – and the type of reading experience that this induces. Finally, it implies circumscribing the aporias of natural language processing models that are most often decontextualized and (re)translating or reinserting them in the social, political, economic and cultural reality from which they originate, particularly in the link between platform capitalism and the ethical desiderata of today.

Our analysis is articulated in three stages which correspond to the three implications mentioned earlier: i) taking seriously, ii) understanding and iii) circumscribing the aporias of NLP models. In the first section, we try to grasp these most recent models as social constructions and socio-technical assemblages (Schwartz 1989; Woolgar 1985). BERT – or Bidirectional Encoder Representations from Transformers – was introduced by Google in 2018 and later integrated into its main search engine. It collects information from Wikipedia, for example, and reads from right to left and back again to identify multiple parallel connections and predict missing terms. OpenAI's GPT-3 – Generative Pre-Trained Transformer – is newer; with its 175 billion parameters, it is said to outperform Google's model by 400 times in "encoding" textuality and thereby opening up a huge range of writing possibilities – journalistic, IT, administrative, etc. What these two models have in common is that they are not exactly black boxes, but rather the objects of a particular historical development which is for many the object of its difficulties and limitations.

In the second section, the overall meaning of this advanced automation comes under scrutiny. What are the implications of the epistemological conception promoted through these data architectures and statistical regressions? And of both the mediation and the recipient of language in this type of connectionist and cybernetic machine? These questions prompt a certain diversion through hermeneutics – that of Paul Ricoeur will be privileged here, partly because his notion of world allows us to think of a semantics, a reference and a "Being-demanding-to-be-said" of textuality, which gives the measure of how AI models sometimes, if not often, appear "shockingly good, and completely mindless" (Heaven 2020). In other words, this world of which Ricoeur speaks is

what can allow us to rethink the link between meaning and reflexivity. The latter is understood here not just as the reader's reflexivity, but also more broadly as the rediscovered reflexivity of the real-mundane world, in the context of society, culture and political economy.

In the third and final section, we therefore seek to develop a sociological and critical understanding of the deployment of these problematic, yet perfectly practical interpreting machines – BERT and GPT-3 – that are intruding on everyday life. The value of these models is inseparable from a market of data and meaning extraction in which some thrive more than others and for which, as the Gebru case shows rather well, ethics becomes a kind of justification and even commodity.

2. The drive to automate language: an all too brief history

The last few years have seen a major evolution in natural language processing. For the first time, language models based on a so-called Transformer architecture make it possible to generate texts that are sufficiently coherent to fool their readers, without relying on a deductive and symbolic logic previously decided by a programmer (Buchanan 2005; Balpe 1991). Based on the mechanism of machine learning, and particularly deep learning, this type of computer programming proposes to imitate some of the cognitive mechanisms of the brain, notably by means of artificial neurons – in reality miniature computer programs that activate or deactivate themselves according to the result of their calculation. As with the human brain, the strength of the mechanism lies in the networking of a large number of these miniature programs. This method, for a long time on the fringes of the AI field, suddenly came back into the spotlight during the 2012 ImageNet competition, won by Geoffrey Hinton's team thanks to the combination of great computing power, a vast data set and this method, which is rightly described as connectionist (Cardon/Cointet/Mazières 2018; Domingos 2015).

The recent evolution of NLP is marked by four significant changes, corresponding to four significant publications. The first paper, published one year after Hinton's great demonstration, was "Efficient Estimation of Word Representations in Vector Space" (Mikolov et al. 2013). Written by a team from Google – Jeff Dean is one of the co-authors – the paper proposes a group of language models called Word2vec, which aims to reconstruct the linguistic context in which words are used. Word2Vec, like most machine learning technologies,

relies heavily on the principle of regression, a method of statistical analysis that allows one variable to be placed in relation to its correlations with others. Roughly summarized, the technology involves locating the variable – the "meaning" – of a word in relation to the variables – the "meanings" – of other words around it. As its name suggests, Word2Vec aims to transform words into vectors, i.e., to model the information they contain using algorithms. In practice, Word2Vec "vectorizes" words using two distinct and complementary architectures. One, called CBOW (Continuous Bag of Words), seeks to predict a word according to the five words to its right and the five words to its left. The other, called Skip-gram, does exactly the opposite and seeks to predict the words in the context according to a given word. The logic is always predictive: the model must be able to assign the "right" vector to each word. Despite its success in the early 2010s, Word2Vec is severely limited. In particular, the language model assigns only one meaning per word and only vectorizes individual words, so that the meaning of even a relatively simple sentence continues to elude it (Horn 2017; Cusin-Berche 2003). We will return to this point later.

To address these limitations, Vinyals and Le – also from Google – published a paper shortly afterwards entitled "A Neural Conversational Model" (2015). This proposed, quite simply, to apply a sequential approach to Word2Vec to model the meaning of a text by linking certain sequences with others – thus forming a longer or "networked" form of text mapping (Sutskever/Vinyals/Le 2014). With this approach, modeling can now be applied to larger sequences, including sentences: the sentences preceding and following the target sentence are thus also taken into account and the model allows for a minimum of contextual consideration.

Despite this progress, Word2Vec-type systems are still based on the approach where a word can only have one meaning. It is this limitation that the article "Deep contextualized word representations" (Peters et al. 2018) aims to overcome. The authors propose a new architecture called Embeddings from Language Models, or ELMo, within which the model can now recognize the dynamic – moving, situational – nature of word meaning. In practice, each word is assigned a coefficient or "weight" according to its influence in the sentence. A word like "bow" can now have different meanings depending on a certain context – "I broke my violin bow" and "I am sitting at the ship's bow." Above all, ELMo makes it possible for the first time to consider modeling that does not learn from the text in a purely orderly way, by offering a "reading" in three different ways: first from left to right – from beginning to end; then in reverse – from end to beginning; and finally by combining the vectorized meanings

of both types of analysis. Also, ELMo marks the real beginning of pre-trained models, allowing users to avoid having to train their models from scratch on huge data sets – an extremely expensive practice, if only in terms of time and computing power.[2]

Finally, the paper "Attention is all you need" (Vaswani et al. 2017) marks the moment when the Transformer architecture virtually seals the field's fate. Previous sequential models had difficulty retaining information about the prioritization of terms among themselves: to take the previous example, information in a simple sentence – "I broke my violin bow", "I was sitting at the ship's bow" – was difficult to retain in longer sequences – "then the bow was splashed". The Transformer architecture moves away from this approach and its many problems in terms of memory, computational speed, word position, etc., by proposing to identify the context that gives meaning to words, which are then processed in parallel. This involves the use of both an encoder and a decoder – and indeed many of them on multiple levels acting and producing feedback in a cybernetic manner. The encoder transforms information into code by giving a calculated value to a word; a decoder does exactly the opposite, transforming code into information by "calculating" a word from a value.

The point to keep in mind here is that such architectures are based on neural networks and on the "deep learning" made famous since Hinton's demonstration of 2012, and in which layers and layers of encoders and layers and layers of decoders can be arranged without too many limits other than technical ones. Above all, the truly innovative character of the Transformer architecture lies in the attention mechanism implemented. The idea is to calculate a "weighted matrix product" – in other words, a matrix score that determines the level of attention that a word should have towards other words; some might also speak, more simply, of situational dependency. An encoder can thus compute several "attention heads" that work in a bidirectional way: an attention weight is computed as input and produces an output vector. The major advantage of this in-depth bidirectionality is that it allows the information to be

2 Unlike the image recognition field – where anyone could download pre-trained face recognition models from ImageNet, for example – the field of NLP appeared before ELMo to be a unified environment, in which each research group or company had to start from scratch, with its own data and available computing power. Drawing on the example of image recognition, different types of pre-trained models emerged at the same time as ELMo, such as ULMFit or the first OpenAI Transformer system. See Ruder 2018.

processed in parallel by the different attention heads, and therefore by the different encoder layers. This results in considerably reduced training times for language models compared with sequential approaches such as Word2Vec.

It is thus these new attention mechanisms specific to Transformer architectures that are at the source of the current successes of language models, in particular Google's BERT and OPEN AI's GPT-3, as spearheads of the ongoing battle waged by the GAFAM (Google, Amazon, Facebook and Microsoft) in their quest for the mastery of artificial intelligence (Thibout 2019; Horowitz 2018). BERT is still a relatively small model compared to GPT-3, as it has been pre-trained on about 3.3 billion words and has 345 million parameters (Devlin et al. 2018). Its main objective is to end formalized keyword searches, a goal that may seem trivial at first but is central to the company's mission statement – "to organize the world's information to make it universally accessible and useful". To achieve this, Google must enable its users to express themselves in the most natural, user-friendly and dialogical way possible.[3] BERT tries to achieves this objective by focusing on the encoder part of the architecture, the part that transforms information, the written or spoken request, but also texts to be translated, for example, into code and vector as to capture their contours: who does what, where, etc. BERT, in other words, and above all, "understands" in the sense of extracting the relevant elements as rendered in more encompassing sets. Its Transformer architecture is thus very flexible and functions as an interface between the natural input language (the query) and the output (the result). It should also be noted that BERT is open-source, which is part of a corporate strategy of value creation quite specific to Google,[4] to which we will return in the last section.

---

3 Prabhakar Raghavan, vice-president of Google, explains that the ultimate goal is to respond *directly* and *intelligently* to users' needs: "Let's say you are planning to go hiking on Mount Fuji […] Do my hiking boots suffice? Today, what you do is you transcribe it into hours of interaction with Google […] Wouldn't it be a lot better if you could […] let Google figure this out and address the need behind your query? […] I want to be able to get to a point where you can take a picture of those hiking boots and ask, 'Can these be used to hike Mount Fuji?'"; see Levy 2021.

4 "With this release, anyone in the world can train their own state-of-the-art question answering system (or a variety of other models) in about 30 minutes on a single Cloud TPU, or in a few hours using a single GPU" (Nayak 2019). See also Devlin and Chang 2018. On Google's open-source strategy (as opposed to Microsoft's, in particular), see Janakiram 2017.

GPT-3 is at the time of writing the most powerful language model, trained on around 570 gigabytes of data and composed of 175 billion parameters (Brown et al. 2020). GPT-3 explicitly aims to generate text, according to its creators. Unlike BERT, therefore, it favors the decoder part of its architecture, the part that more precisely allows code to be transformed into information, i.e., inferring missing words, completing sentences, etc. Far from being open-source, GPT-3 is currently marketed via its Application Programming Interface (API), a choice which is obviously part of a corporate strategy that aims to control the economic ecosystem on which many future companies will be based. In a blog post, OpenAI reported in March 2021 that more than 300 companies were making use of this API, a number that keeps growing. For instance, applications already available include CopyAI, which can generate slogans and product descriptions for companies, and Fable, which can model characters from novels and talk to them (Scott 2020).[5]

Because they are socio-technical assemblages, it goes without saying that these models struggle to be perfect or even to live up to the rhetoric legitimizing their use and, more generally, all that is the magic of AI (Roberge 2020; Elish/Boyd 2018). The fact is that all is not well in the best of all NLP worlds, and that upon closer inspection, its deployment is more a matter of "garbage in, garbage out" – the so-called GIGO principle, as ironically referred to by scientists working in the field (Kilkenny/Robinson 2018). With regard to input, it should be seen that while language can apparently be computationally modelled, the Transformer architecture can only achieve this from a resource that is itself a social construct: the database. This dependence of language models on their training sources is quite widely discussed (Hutchinson et al. 2020; Roberge 2018), as any given Transformer architecture remains based on the principle of regression outlined earlier, which aims to locate a variable (a word) according to its correlations with other variables (the other words in the database). This simple mathematical procedure thus constructs an approach to language based on the principle of "winner takes all". Put differently, the language model promotes the most statistically probable language constructs according to the data set on which it is trained. Also, the choice of texts on which these language models are developed participates in a certain representation of the world, whose symbolic, if not ideological, dimension is often only revealed once the models have been applied – through the racist,

5 To see the companies directly: CopyAI, https://www.copy.ai/ [last access: 18 March 2021] and Fable, https://fable-studio.com/ [last access: 4 June 2021].

misogynistic or other biases that result from them. This is one of the most striking observations in the text that led to Gebru's dismissal from Google:

> GPT-2's training data is sourced by scraping outbound links from Reddit, and Pew Internet Research's 2016 survey reveals 67% of Reddit users in the United States are men, and 64% between ages 18 and 29. Similarly, recent surveys of Wikipedians find that only 8–15% are women or girls (Bender et al. 2021: 4).

What is thus a problem on the input side becomes a problem on the output side, with the highest number of potential and proven slippages. One of the most feared applications in this respect is what is usually referred to as astroturfing, in which a plethora of micro-speeches is automatically generated as if to simulate a mass movement accrediting such organizations, ideas, etc. (Kovic et al. 2018; Zhang/ Carpenter/Ko 2013). Indeed, the US National Intelligence Council's latest report *Global Trends 2040* lists AI-powered propaganda as one of its top ten economic and political security concerns (2021).[6] In this case, fake profiles with automatically generated content already populate social networks that are used by billions of people on a daily basis and are therefore prone to misinformation, manipulation and the promotion of hate speech (Keller et al. 2020). There are also other examples of biases embedded in BERT and GPT-3 that are related to the probabilistic ideology of these models. AI Dungeon, a computerized version of Dungeons & Dragons powered by GPT-3, made news in April 2021 for, among other things, allowing the generation of narratives featuring sexual relations involving children – a phenomenon that was obviously not foreseen by OpenAI (Simonite 2021a). In the follow-up to their *Algorithms of Oppression: How Search Engines Reinforce Racism*, Noble and others have also extensively exposed the biases that have always been built into Google environments, both in the various language models that preceded BERT and in the way that BERT is now far from solving these difficulties (Noble 2018; Bhardwaj, Majumder and Poria 2021; Hutchinson et al. 2020). "Stochastic Parrots" is part of this broader critique of NLP:

> The size of data available on the web has enabled deep learning models to achieve high accuracy on specific benchmarks in NLP […]. However, the

6 In particular, the report explains that "[b]oth states and nonstate actors almost certainly will be able to use these tools to influence populations, including by ratcheting up cognitive manipulation and societal polarization to shape how people receive, interpret, and act on information" (National Intelligence Council 2021: 97).

> training data has been shown to have problematic characteristics […] resulting in models that encode stereotypical and derogatory associations along gender, race, ethnicity, and disability status (Bender et al. 2021: 4).

Despite its deleterious consequences for their authors – Gebru in particular – "Stochastic Parrots" is not particularly innovative. As Wired reports, "the paper was not intended to be a bombshell".[7] It merely explores three major issues related to the ever-growing size of language models: first, their environmental cost; second, their formal and rigid nature, which allows biases both to structure themselves and often to go unnoticed; and third, solutions that might mitigate the risks associated with their use.

In the sections that make up the core of the overall argument, the paper reminds us that the models are trained only on the form of the language and not on its substance. To use Saussurean terms, a model can only ever master the signifier of language, never the signified – an argument developed in another paper by Bender and Koller (2020). "Stochastic Parrots" uses this argument to denounce the deceptive or illusory character of the current successes of models such as BERT and GPT-3, which seem to master language when they will only ever have a statistical understanding of it:

> Text generated by an LM [language model] is not grounded in communicative intent, any model of the world, or any model of the reader's state of mind. […] Contrary to how it may seem when we observe its output, an LM is a system for haphazardly stitching together sequences of linguistic forms it has observed in its vast training data, according to probabilistic information about how they combine, but without any reference to meaning: a stochastic parrot (Bender et al. 2021: 616–617).

3. Problematizing and understanding the world of hermeneutic machines

Even such a brief history of natural language processing should serve to show how it is very much about meaning and significance. This is essentially what

7 "The authors did not present new experimental results. Instead, they cited previous studies about ethical questions raised by large language models, including about the energy consumed […]. An academic who saw the paper after it was submitted for publication found the document 'middle of the road.'" (Simonite 2021b).

is at stake. First, it is clear that a certain hermeneutic claim of AI cannot be ignored, denied or simply dismissed out of hand. This claim is already disseminated across a wide environment: from researchers like Hinton, by declaring that models "are going to do things like common reasoning", to digital business leaders talking about their platform as a "content understanding engine" (Candela) "focus[sed] on understanding the meaning of what people share" (Zuckerberg). To take another example, the Toronto-based company Cohere, which specializes in the design of NLP models, has set itself the motto and mission of "building machines that understand the world".[8] So while all these claims should be taken seriously, this does not mean that they should be accepted without question. Second, it is clear that there is a need for a better understanding of what we are talking about here, i.e., a better grasp of both the scope and the limits of these hermeneutic machines. The intellectual effort, in other words, is still one of problematization (Romele et al. 2020; Hongladorom 2020).

Looking at the most common criticisms of this particular kind of automated "management" of language and meaning, it is possible to see how they represent variations on the theme of Clever Hans, the so-called intelligent horse from the turn of the twentieth century that appeared to find answers to arithmetic problems on a blackboard; in fact, it was only responding to its master's stimuli and indications. According to Crawford (2021: 151), for example, this represents the embodiment of our desire to anthropomorphize the non-human, as well as a certain spectacle of what intelligence is, without considering a whole set of institutional relationships and political tensions. For others, the image of Clever Hans serves to illustrate the lightness, if not the hermeneutic superficiality, of AI and its language models; as Pavlus's commentary points out, "even a simulacrum of understanding has been good enough for natural language processing" (2019).[9] That said, it is perhaps Gary Marcus in recent years who has done the most to identify the various ways in which what is deemed "deep" in all things deep learning remains only an architectural and technical property – and thus not symbolic and hermeneutic

8 Cohere website, https://cohere.ai/about. Last access: 19 July 2021.

9 This idea of "simulacrum" is understood here not so much in its postmodern and Baudrillardian sense, but more simply as the emergence of handy solutions that are accepted above all for their efficiency. This is what Floridi and Chiriatti refer to when they note how GPT-3 "represents the arrival of a new age in which we can now mass produce good and cheap semantic artifacts" (2020: 690).

(Marcus/Davis 2019, 2020). His argument is threefold. First, this type of model lacks what he calls compositionality – that is, the ability to play with complex and often plotted meanings. On this first point, Marcus is quite close to the idea of the hermeneutic circle – that of Gadamer in particular – in which the whole and the part are so much in dialogue that one can hope to arrive at a form of truth which is more than a simple methodical assemblage (Marcus 2019a; Gadamer 1996 [1960]; see also Andersen 2020). Second, Marcus insists that models like BERT or GPT-3 have "no good way to incorporate background knowledge" (2019). Categories or tools are put forward – for instance those of probability, distance, variation or threshold which have their own logic, horizontal so to speak. Of course, they calculate meanings, but without wanting or being able to draw on their historical, cultural and other richness. And, third, this is what translates into a substantial semantic issue:

> The problem is not with GPT-3 syntax (which is perfectly fluent) but with its semantics: it can produce words in perfect English, but it has only the dimmest sense of what those words mean, and no sense whatsoever about how those words relate to the world (Marcus/Davis 2020).

It is this latter notion of world that seems to be the measure here, even if obviously not without ambiguity itself. Marcus makes use of it, but defines it rather sparingly – which is also the case with everything to do with the form of language in Bender and Gebru, as seen above. How and why do words, meaning and the world appear so inseparable?

This is the type of question that is central to Paul Ricoeur's hermeneutic reflection on textuality, a reflection that can be revisited in the age of natural language processing (Ricoeur 1991a [1986]; Moore 1990; Roberge 2008). "The 'thing' of the text – is the object of hermeneutics, writes the philosopher. Now the thing of the text is the world it unfolds before itself" (1991a: 95). Something is fixed by writing that is not reducible to the intention of its author or to the social conditions of its production – behind or beyond, something by which Ricoeur seeks to guard against a certain romanticism and a certain determinism. As tautological as it may seem, the world of the text is its world, as if to signal its autonomy and objectivity, not only once but twice. On the one hand, textuality in Ricoeur's sense has an internal dynamic and structuring that are reminiscent of the compositionality discussed by Marcus. But on the other hand, and without any contradiction, all texts are always about something, namely that they all have their own reference in a world that they open up and discover? This world is not reality as such, since this would exclude all works of fiction.

No, the world in question is indeed that of the meaning unfolding in it, i.e., a certain universality in the discourse that would represent its "claim to truth" or "Being-demanding-to-be-said" (1991a [1986]: 35 and 19).

What Ricoeur is trying to do is to think of the world of textuality as mediation and suggestion, as what is, so to speak, given to interpretation. The reflection is then resolutely ontological and phenomenological – the author speaks elsewhere of the "immanent transcendence" of textuality (1984). In a text, fundamentally, what is at stake are "sensory [...] and axiological values that make the world one that can be inhabited"(1991a [1986]: 11). Ontologically and phenomenologically, this means that it is also always a question of human experience, so that Ricoeur's aim is to combine or bridge different possibilities which are hardly compatible a priori: experience and reflexivity, text and action, explanation and interpretation-comprehension, as well as, more broadly, philosophy and the human and social sciences (Ricoeur 1977; 1991b).

This quick digression through the textual world can only raise the question of its destination: why does it become meaningful and for whom? The whole problem with models like BERT or GPT-3 is that they provide ethereal solutions to this issue, namely that they have infinite difficulty in constructing a meaningful world which, as a result, really means something to someone. Ricoeur saw this horizon of textuality and how, therefore, it forced a reflection on the multiple relationships – complex and ambiguous – between world and appropriation, interpretation of texts and self-understanding (Roberge, 2008). "Reading is like the execution of a musical score, he writes, it marks the realization, the enactment, of the semantic possibilities of the text" (Ricoeur 1991a [1986]: 119). Thus, what a hermeneutic theory like that of Ricoeur suggests is nothing less than the elaboration of a philosophical anthropology (see Ricoeur 1960a and 1960b; 1989). Understanding is as much effort as it is recognition: "to understand oneself is to understand oneself as one confronts the text and to receive from it the conditions of a self other than that which undertakes the reading" (1991a [1986]: 17). It is a question of a diversion through which "I find myself only by losing myself" (1991a [1986]: 88). For Ricoeur, the appropriation in question is more necessary than easy, as if hermeneutic reflection represented a call or a challenge.

However, it is this type of hermeneutic challenge that AI and natural language processing, BERT and GPT-3 in particular, refuse to take up today. One example is the discussion about the "interpretability" and "explicability" of machines that have been the subject of much ink in recent years (Biran/Cotton 2017; Gilpin et al. 2018). For the computational sciences, one of the challenges

is to move away from this (polluted) black box image by showing models in their simplicity and transparency with the avowed aim of increasing confidence in them. Dietterich illustrates this position rather well when he notes, for example, that the aim is "to translate our fuzzy notion of interpretation and understanding into concrete, measurable capabilities" (2019). Interpretability and explicability, in other words, are neologisms of a practical, if not technical and instrumental, nature, which share the logic of automation with the related terms of prediction, optimization, generalization and so on. Some, in fact, have argued that all this is conceptually confusing, to say the least; that there is "conflation" (Miller 2019) between explicability and interpretability or that the latter is "ill-defined" (Lipton 2016). Others have gone on to note that there is a kind of reassignment of the debate's parameters (Mittelstadt/Russell/Wachter 2019), and this in the double sense of translation and impoverishment. In short, the unbearable lightness of the discussion in vogue in the field of AI lies in the fact that it questions nothing or so little, whereas, quite rightly, the challenge of hermeneutics is that of an opening up, a *mise en abyme* and a problematization. For this is what it is all about: as Mittelstadt, Russell and Wachter point out, reflection within the field itself "might benefit from viewing the problem […] more broadly" (2019: 7; see also Campolo/Crawford 2020). What is understanding and interpreting in the age of natural language processing? What kind of world, subject, experience and doubt does this bring into play? Asking these questions encourages us to think of or rethink hermeneutics as part of a search for reflexivity, both individual and collective – that of a subject, but also of a society, a culture, etc.

If words, the world and experience are so inseparable, it is because this world can be said in different ways. It is not by chance that this polysemy is present in Bender and Gebru or in Marcus, and it is not by chance either that it is already present in Ricoeur. Hermeneutics is contextualizing; that is, the world is as much in the text as the text is in the world. The hermeneutical question of language automation is that of a certain pregnancy or anchoring of reflection in what can be said of reality. This can be seen, first of all, at the level of meaning; as Romele notes, "meaningfulness [should be] problematized in [its] context-dependency" (Romele/Severo/Furia 2020: 78). This can then be seen in the historicity of understanding, whereby a subject is always situated in time and space and where such a situation necessarily colors that subject's reading

of what happens.[10] Finally, and to get to the heart of the matter, this can be seen in the very object of what we are dealing with here: AI, natural language processing, BERT and GPT-3. In fact, it was even the first sociological sketches of this vast field of technologies that insisted on showing that it was "socially constituted" (Schwartz 1989; see also Woolgar 1985 in particular). This is not to say that determinism is triumphant – which, as seen above, would not satisfy a hermeneutic perspective such as Ricoeur's – but rather that there is something of a co-construction, a cross-referentiality or a resonance between contextuality and technological advances.[11]

What about this world today? Our world? What characterizes it so well that it makes possible the kind of scandal surrounding the person of Timnit Gebru and the publication of "Stochastic Parrots"? Among other fundamental things, it is clear that we are increasingly living in the midst of not only an increasing "platformization" of the web and digital culture (Helmond 2015), but also an increasing datafication of everyday life (Van Dijck 2014). When, as above, the CEO of Facebook says that his platform "focuses on understanding the meaning of what people share", this is what it is all about. Individually and collectively, it is about our data, our information, and a work that is constantly in progress, which goes as far as the way we (re)construct language, writing, reading, etc. This never fails to be problematic: it is precisely these worlds of meaning that are increasingly under the sway of an appropriation that could be described here as other or heteronomous – as we define further in the next section. At the same time, this should incite hermeneutics – very broadly – to rethink interpretation-comprehension as a sociological and critical issue by attempting, for instance, to reflect on the political economy that does not fail to go with the deployment of language processing models such as BERT or GPT-3 (Roberge 2011; 2020).

10 This is also a large part of the debate between Gadamer and Habermas about the *Vorstruktur des Verstehens*. See Roberge 2011.

11 This is also one of the basic precepts of a vast literature in Science and Technology Studies (STS) ranging as far as Holton and Boyd 2019.

4. Circumscribing the aporias: between critical hermeneutics and Critical AI Studies

The platformization-datafication of the '*hic et nunc*' world is situated in a particular context with equally particular, practical, almost down-to-earth implications and origins. It is its modus operandi that remains to be understood and interpreted. Most concretely, the recent history of AI is about a pragmatic deployment that is therefore more utilitarian than reflexive. It is a question of optimizing solutions as automated forms of action and decision-making. This applies, for example, to autonomous vehicles, cancer diagnosis by algorithmic imaging and much else, including language processing (Stilgoe 2018). What most of these applications and models share is that they are part of an engineered modus operandi, which in turn is part of what Pedro Domingos, a leading figure in the field, calls its "black art" (cited in Campolo/Crawford 2020: 7–8). To train and calibrate a model is to tinker with it; it is to "tweak it to the level of detection that is useful to you" (Amoore 2019: 6). And this is one of the reasons why these solutions are often beta and still imperfectly implemented. Choices are made that nevertheless respond to a certain logic, pressure and urgency. This once again raises perfectly concrete and practical questions: "what is being optimised, and for whom, and who gets to decide[?]" (Crawford 2021: 9). Here we must take another step forward with the author of the *Atlas of AI* when she notes the eminently political nature of all these issues. New power relations are being established prosaically, but certainly. For Crawford, what we are witnessing today is a "shifting tectonics of power in AI" (2021: 11). In a steady fashion, the control of technology gives access to controlled resources. The distribution of power is thus being reorganized more in the sense of greater aggregation than in that of greater equality or symmetry.

Politics and economics are intimately linked, of course, and in the case of AI and natural language processing, this requires a particular adaptation of contemporary capitalism. One of the fundamental reasons why the GAFAM of this world are investing in the development of interpretive machines like BERT and GPT-3 is for the competitive advantage, even dominance, that can be gained. As noted in Simonite's well-known commentary in *Wired*, there is a form of highly performative desiderata here that "makes tech giants harder to topple" (Simonite 2017). It is not that these companies show solidarity among themselves, or conspiratorial tendencies against the rest of the world, but rather that their entire innovation efforts are part of a single "cooperative struggle" (Crandall 2010). To risk an analogy: if each of them occupies a particular po-

sition on the chessboard, they all play the same game of chess that is natural language processing here. As we saw in earlier, Google's BERT is historically entwined in academic research, which ensures that an open-access and open-science model is favored. BERT, in other words, is open, even if only for strategic reasons. In fact, Google's advantage is precisely that it can bring everyone into its environment – familiar, saved in the cloud, allowing for easy transition between different devices, etc. For its part, GPT-3 follows a more direct, if not aggressive, proprietary strategy, like the Microsoft ecosystem of which it is now an integral part. As such, copyright ownership of the content generated by these language models belongs in principle to the company operating the model. It is therefore possible to see the economic challenge in which companies using services such as GPT-3 or BERT could have no more rights to what they generate, or even to the computer code from which their product operates. This issue is still in its infancy, but it is likely to be the major copyright issue of the twenty-first century.

But again, critical thinking cannot reduce everything to economic relationships. When, for example, a renowned researcher in the field such as Yoshua Bengio points out that AI models have become very valuable for GAFAM,[12] he is undoubtedly pointing to broader, if more ambiguous, possibilities. It is these possibilities that, among other important issues, will be crystallized in Gebru's case. We should remember her tweet on leaving: "@jeffdean I realize how much large language models are worth to you now." The problem with the value of natural language processing is that it is both economical and practical, on the one hand, and axial, normative and symbolic, on the other. Hence the reason for a critical hermeneutics around a political economy of meaning and significance as well as the reason for developing Critical AI Studies.

A broader and more distant reflection may point to the whole problem of "assetizing" (Birch/Muniesa 2020) not only data, but also language models and language as such. Optimizing-reducing, enriching-appreciating, common-particular, this "becoming-resource" of language is one of those uncertain couples whose meaning emerges in the gap separating it from appropriation, as described above. This meaning is no longer so much reflexive as extractivist. Following the argumentative line of Birch and Muniesa as well as of Crawford and others, there is a justification and a belief currently being implemented that "everything is data and is there for the taking" (Crawford 2021: 93) –

12 Yoshua Bengio, https://www.technovation.org/blogs/an-interview-with-yoshua-bengio/. Last access: 28 June 2022.

something Shaev et al. summarize perfectly by speaking of "platform meaning extraction" (2019). The general business model works to become a model of the world; that is, it sets up a new normality of which it is both the guarantor and the main beneficiary. The AI myth continues unchallenged, except that, once again, it is the task of a critical hermeneutics to ask questions and show how everything from AI to BERT or GPT-3 is a construction and contingency for which other possibilities are imaginable.

Consider, for example, the ethical turn that the debate around the Gebru affair has sometimes, if not often, taken. In fact, it is common sense to link AI and ethics – as if the hype of the one could not go without the hype of the other and as if, in this cross-discussion, there was not a whole industry, both public and private, of discourse production (Jobin/Ienca/Vayena 2019; Roberge 2020). However, this association is never self-evident and always rather problematic. Authors such as Mittelstadt have shown, for example, how the major principles put forward on the international scene were quite rightly vague and formal, as well as representing "a reason not to pursue new regulation" (2019: 501; see also Wagner 2018). In the same vein, Elish and Boyd have emphasized the normative and political aspects that go hand in hand with such "ability to manufacture legitimacy" within fashionable ethical discourses (2018). And this is what the Gebru test or crisis exposes. When it comes to evaluating or amending itself, Google remains judge and jury. What the company wants to say – or make clear – diverges from what it needs to do. Hao's commentary points out: "As Google underscored in its treatment of Gebru […], the few companies rich enough to train and maintain large language model investments have a heavy financial interest in declining to examine them carefully" (2021: 2).[13] An important part of "Stochastic Parrots" is the discussion of discrimination and bias – gender, race, etc. – that is not just aimed at Google. This is almost worse, as it signals that the problem is more fundamental – a structural one. The article speaks of "real harm" and a simultaneous immediate and insatiable need for accountability, as if this is where the very meaning of criticism becomes eminently practical.

13 An "insider" account of this same idea is found in Lemoine: "Google has moved from being the company whose motto is 'Don't be evil' to being the company whose motto is 'if you don't like it there's the door.' Business interests kept clashing with moral values and time and time again the people speaking truth to power were shown the door" (2021: 4).

The fate of criticism as a result of the Gebru affair is of great interest to a perspective like ours. Critical hermeneutics and Critical AI Studies are in fact intimately linked to these exercises of reflexivity *in situ*, to discourse on discourse and to the development of a political economy of meaning. Its stakes are perfectly summarized by Hanna and Whittaker (2020):

> Gebru's firing suggests this dynamic is at work once again. Powerful companies like Google have the ability to co-opt, minimize, or silence criticisms of their own large-scale AI systems – systems that are at the core of their profit motives […]. The handful of people who are benefiting from AI's proliferation are shaping the academic and public understanding of these systems, while those most likely to be harmed are shut out of knowledge creation and influence.

Perhaps it only remains to be added that what is at stake here is the possibility of a critical culture. What can still be discussed in the automation of language? What can still be discussed about it? These questions should not be closed. Gebru, for her part, is annoyed, but basically she is right: "'Responsible AI' at Google = promote those good at ethics washing & ensuring the marginalization of those already marginalized. I'm telling you after all this they have zero shame."[14]

5. Conclusion

Artificial intelligence is now well and truly conquering language; this is as much a form of zeitgeist as it is of technological development. Models such as BERT and GPT-3 are becoming powerful interpreting machines, to say the least, which, of course, is not without its share of claims. It was the purpose of the first section of this text to take these models seriously. The recent history of natural language processing is partly linked to advances in deep learning and the way in which this type of architecture and networking, based on a principle of statistical regression, now allows parallel processing of a large quantity of data that a model does not need to "understand" in order to calculate efficiently. The strength of Transformers lies in their flexibility: the relationship between letters and numbers, words and codes, or sentences and vectors is

14 Timnit Gebru, tweet, https://twitter.com/timnitGebru/status/1391111917968707585. Last access: 20 July 2021.

thus played and replayed in a continuous flow. The models adapt to platforms and digital culture, which at least partly obliterates some of their weaknesses. As we have seen, these are not exactly hidden, but nonetheless struggle to emerge. When Gebru and company, for example, begin this discussion, it is mostly done through a questioning of the upstream and downstream, i.e., of the biases in the constitution of the databases and in the impacts on populations. The hermeneutic core of the problem remains more or less intact, which is to say nothing of the reception of the researcher's proposal by the industry.

It is, then, a matter of better problematizing in order to better understand. As the second main section has tried to show, the issue of natural language processing is fundamentally semantic. Following the example of Marcus mentioned above, GPT-3 has "no sense whatsoever about how […] worlds related to the word." This is not simple, of course, since this notion of world is sufficiently rich and encompassing to be polysemous. And it is here that a diversion through hermeneutics – that of Ricoeur in particular – is fruitful, insofar as a world can be that of a text as a truth value and a relationship to appropriation, as well as that of a context, namely our world through and for history, culture, etc. An important part of the merit of the Ricoeurian position is its ability to hold these two possibilities together, as if it were not necessary to choose, but to reflect on their innumerable interactions. Reality and interpretation revive each other, as do signification and criticism. In the present discussion, this makes it possible to update hermeneutics to reflect on AI and the way it appropriates something of us through the automation of both data and language. The whole problem is that we need to disentangle a new normality that is inseparably technological, cultural, social, economic and political. Circumscribing the aporias – as the final section of the chapter seeks to do – means showing multiple variations of power, inequalities and their justifications, ethical or otherwise. Most fundamentally, the meaning of AI and natural language processing is to be an extraction of meaning and significance. And that is probably where the choice lies. Gebru has chosen. Her rebuff may be personal and not perfectly calibrated, but it has the great advantage of assuming its political charge by indicating that a critique is always possible, a fortiori when it draws its source from experience and echoes the very idea of society.

Bibliography

Amoore, Louise. 2019. Doubt and the algorithm: on the partial accounts of machine learning. *Theory, Culture & Society* 36(6):147–169.

Andersen, Jack. 2020. Understanding and interpreting algorithms: toward a hermeneutics of algorithms. *Media, Culture & Society* 42(7-8):1479–1494.

Balpe, Jean-Pierre. 1991. Macro-structures et micro-univers dans la génération automatique de textes à orientation littéraire. In *L'imagination informatique de la littérature, Colloque de Cerisy,* Eds. Bernard Magné and Jean-Pierre Balpe, 128–149. Presses Universitaires de Vincennes.

Bender, Emily M. and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major et al. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.

Bhardwaj, Rishabh, Navonil. Majumder and Soujanya Poria. 2021. Investigating gender bias in BERT. *Cognitive Computation:* 1–11.

Biran, Or and Courtenay V. Cotton. 2017. Explanation and justification in machine learning: a survey. IJCAI-17 Workshop on Explainable AI (XAI).

Birch, Kean and Fabian Muniesa (Eds). 2020. *Assetization: Turning Things into Assets in Technoscientific Capitalism.* MIT Press.

Brown, Tom B., Benjamin Mann, Nick Ryder et al. 2020. Language models are few-shot learners. arXiv preprint, 1–75.

Buchanan, Bruce G. 2005. A (very) brief history of artificial intelligence. *AI Magazine* 26(4):53–60.

Campolo, Alexander and Kate Crawford. 2020. Enchanted determinism: power without responsibility in artificial intelligence. *Engaging Science, Technology, and Society* 6:1–19.

Cardon, Dominique, Jean-Philippe Cointet and Antoine Mazières. 2018. La revanche des neurones. *Réseaux* 5:173–220.

Crandall, Jordan. 2010. The Geospatialization of Calculative Operations: Tracking, Sensing and Megacities. *Theory, Culture & Society* 27(6):68–90.

Crawford, Kate. 2021. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence.* New Haven, Conn.: Yale University Press.

Cusin-Berche, Fabienne. 2003. *Les mots et leurs contextes.* Paris : Presses Sorbonne nouvelle.

Devlin, Jacob and Ming-Wei Chang. 2018. Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing. https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html. Last access: 15 June 2021.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee et al. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint, 1–16.

Dietterich, Thomas G. 2019. What does it mean for a machine to 'understand'? https://medium.com/@tdietterich/what-does-it-mean-for-a-machine-to-understand-555485f3ad40. Last access: 21 July 2021.

Domingos, Pedro. 2015. *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. New York: Basic Books.

Duesenberry, James S. 1949, *Income, Saving, and the Theory of Consumer Behavior.* Cambridge, Mass.: Harvard University Press.

Elish, Madeleine C. and Danah Boyd. 2018. Situating methods in the magic of Big Data and AI. *Communication Monographs* 85(1):57–80.

Floridi, Luciano and Massimo Chiriatti. 2020. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines* 30(4):681–694.

Gadamer, Hans-Georg. 1996 [1960]). V*érité et méthode. Les grandes lignes d'une herméneutique philosophique.* Paris: Seuil.

Gilpin, Leilani H., David Bau, Ben Z. Yuan, Ayesha Baiwa, Michael Specter and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), 80–89.

Hanna, Alex and Meredith Whittaker. 2020. "Timnit Gebru's Exit from Google Exposes a Crisis in AI," Wired. https://www.wired.com/story/timnit-gebru-exit-google-exposes-crisis-in-ai/. Last access: 28 June 2022.

Hao, Karen. 2020. We read the paper that forced Timnit Gebru out of Google. Here's what it says. https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/. Last access: 28 June 2022.

Hao, Karen. 2021. The race to understand the exhilarating, dangerous world of language AI. https://www.technologyreview.com/2021/05/20/1025135/ai-large-language-models-bigscience-project/. Last access: 28 June 2022.

Heaven, Will D. 2020. OpenAI's new Language Generator GPT-3 is shockingly good, and completely mindless. *MIT Technological Review,* July.

Helmond, Anne. 2015. The platformization of the Web: making Web data platform ready. *Social Media + Society* 1(2):1–11.

Holton Robert and Ross Boyd. 2019. 'Where are the people? What are they doing? Why are they doing it?' (Mindell). Situating artificial intelligence within a socio-technical framework. *Journal of Sociology* 7(2):179–195.

Hongladarom, Soraj. 2020, Machine hermeneutics, postphenomenology, and facial recognition technology. *AI & Society,* 1–8.

Horn, Franziska. 2017. Context encoders as a simple but powerful extension of word2vec. arXiv preprint, 1–5.

Horowitz, Michael C. 2018, Artificial intelligence, international competition, and the balance of power. *Texas National Security Review*, 2018:1–22.

Hutchinson, Ben, Vinodkumar Prabhakaran, Emily Denton et al. 2020. Social biases in NLP models as barriers for persons with disabilities. arXiv preprint, 1–5.

Janakiram, M.S.V. 2017. How Google Turned Open Source Into a Key Differentiator for Its Cloud Platform. https://www.forbes.com/sites/janakiramsmsv/2017/07/09/how-google-turned-open-source-into-a-key-differentiator-for-its-cloud-platform/?sh=7a52302e646f .Last access: 15 June 2021.

Jobin, Aanna, Marcello Ienca and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1(9): 389–399.

Keller, Franziska B., David Schoch, Sebastian Stier and JungHwan Yang. 2020. Political Astroturfing on Twitter: How to Coordinate a Disinformation Campaign. *Political Communication* 37(2):256–280.

Kilkenny, Monique F. and Kerin M. Robinson. 2018. Data quality: 'Garbage in–garbage out'. *Health Information Management Journal* 47(3):103–15.

Kovic, Marko, Adrian Rauchfleisch, Marc Sele et al. 2018. Digital astroturfing in politics: Definition, typology, and countermeasures. *Studies in Communication Sciences* 18(1): 69–85.

Lemoine, Blake 2021. The History of Ethical AI at Google. https://cajundiscordian.medium.com/the-history-of-ethical-ai-at-google-d2f997985233. Last access: 21 July 2021.

Levy, Stephen. 2021. Prabhakar Raghavan Isn't CEO of Google – He Just Runs the Place. https://www.wired.com/story/prabhakar-raghavan-isnt-ceo-of-google-he-just-runs-the-place/. Last access: 15 June 2021.

Lipton, Zachary C. 2016. The mythos of model interpretability. 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), 1–9.

Marcus, Gary and Ernest Davis. 2019a. If computers are so smart, how come they can't read? https://www.wired.com/story/adaptation-if-computers-are-so-smart-how-come-they-cant-read/. Last access: 21 July 2021.

Marcus, Gary and Ernest Davis. 2019b. *Rebooting AI: Building Artificial Intelligence We Can Trust*. Vintage.

Marcus, Gary and Ernest Davis. 2020. GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about. https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/. Last access: 21 July 2021.

Mikolov, Tomas, Kai Chen, Greg Corrado et al. 2013. Efficient Estimation of Word Representations in Vector Space. https://doi.org/10.48550/arXiv.1301.3781.

Miller, Tim. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267:1–38.

Mittelstadt, Brent, Chris Russell and Sandra Wachter. 2019, Explaining Explanations in AI. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 279–288.

Mittelstadt, Brent. 2019. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence* 1:501–507.

Moore, Henriatta. 1990. Paul Ricoeur: Action, Meaning and Text. In *Reading Material Culture. Structuralism, Hermeneutics and Post-Structuralism*, Ed. Christopher Tilley. Oxford: Basil Blackwell.

National Intelligence Council. 2021. Global Trends 2040: A More Contested World.

Nayak, Pandu. 2019. Understanding searches better than ever before. https://blog.google/products/search/search-language-understanding-bert/. Last access: 4 June 2021.

Noble, Safiya U. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: NYU Press.

Pasquinelli, Matteo and Vladan Joler. 2020. The Nooscope Manifested: Artificial Intelligence as Instrument of Knowledge Extractivism. *AI and Society*, 1–18.

Pavlus, John. 2019. Machines beat humans on a reading test. But do they understand? *Quanta Magazine* [online]. https://www.quantamagazine.org/machines-beat-humans-on-a-reading-test-but-do-they-understand-20191017/. Last access: 21 July 2021.

Peters, Matthew E., Mark Neumann, Mohit Iyyer et al. 2018. Deep contextualized word representations. arXiv preprint, 1–15.

Ricœur, Paul. 1960a. L'antinomie humaine et le problème de l'anthropologie philosophique. *Il Pensiero* 5(3) : 283–290.

Ricœur, Paul. 1960b. *L'homme faillible*. Paris : Aubier.

Ricœur, Paul. 1977. Phenomenology and the social sciences. *The Annals of Phenomenological Sociology* 2:145–159.

Ricœur, Paul. 1984. *Temps et récit. La configuration dans le récit de fiction*. Vol. II, Paris : Seuil.

Ricœur, Paul. 1989. L'homme comme sujet de philosophie. *Anzeiger der philosophisch-historischen Klasse der Österreichischen Akademie der Wissenschaften* 126:73–86.

Ricœur, Paul. 1991a. *From Text to Action: Essays in Hermeneutics. II*, Trans. K. Blamey and J. B. Thompson. Evanston, Ill.: Northwestern University Press. [First published as Ricoeur. 1986. *Du texte à l'action. Essais d'herméneutique II*. Paris: Seuil.]

Ricœur, Paul. 1991b. L'herméneutique et les sciences sociales. In *Théorie du droit et science,* Ed. P. Amselek, 15–25. Paris : Presses universitaires de France.

Roberge, Jonathan and Michael Castelle. 2020. Toward an End-to-End Sociology of 21st-Century Machine Learning. In *The Cultural Life of Machine Learning: An Incursion into Critical AI Studies*, Eds. Jonathan. Roberge and Michael Castelle, 1–29. New York: Palgrave Macmillan.

Roberge, Jonathan, Marius Senneville and Kevin Morin. 2020. How to translate artificial intelligence? Myths and justifications in public discourse. *Big Data and Society* 7(1). https://journals.sagepub.com/doi/full/10.1177/20539517209 19968.

Roberge, Jonathan. 2008. *Paul Ricœur, la culture et les sciences humaines*. Collection Sociologie contemporaine. Québec : Presses de l'Université Laval.

Roberge, Jonathan. 2011. What is critical hermeneutics? *Thesis Eleven* 106(1): 5–22.

Romele, Alberto, Marta Severo and Paolo Furia. 2020. Digital hermeneutics: from interpreting with machines to interpretational machines. *AI & Society* 35:73–86.

Ruder, Sebastian. 2018. NLP's ImageNet moment has arrived. https://ruder.io /nlp-imagenet/. Last access: 19 July 2021.

Saxenian, AnnaLee. 1994. *Regional Advantage: Culture and Competition in Silicon Valley and Route 128*. Cambridge, Mass.: Harvard University Press.

Schwartz, H. Andrew and Dirk Hovy. 2019. Predictive biases in natural language processing models: a conceptual framework and overview. arXiv preprint, arXiv:1912.11078.

Schwartz, Ronald D. 1989. Artificial intelligence as a sociological phenomenon. *Canadian Journal of Sociology/Cahiers canadiens de sociologie* 14(2):179–202.

Scott, Kevin. 2020, Microsoft teams up with OpenAI to exclusively license GPT-3 language model. https://blogs.microsoft.com/blog/2020/09/22/microsoft-teams-up-with-openai-to-exclusively-license-gpt-3-language-model/. Last access: 4 June 2021.

Simonite, Tom. 2017. AI and 'Enormous Data' could make tech giants harder to topple. *Wired*. https://www.wired.com/story/ai-and-enormous-data-could-make-tech-giants-harder-to-topple/. Last access: 21 July 2021.

Simonite, Tom. 2021a. It began as an AI-fueled dungeon game. It got much darker. *Wired*. https://www.wired.com/story/ai-fueled-dungeon-game-got-much-darker/. Last access: 4 June 2021.

Simonite, Tom. 2021b. What really happened when Google ousted Timnit Gebru. *Wired*, https://www.wired.com/story/google-timnit-gebru-ai-what-really-happened/. Last access: 21 July 2021.

Stilgoe, Jack. 2018. Machine learning, social learning and the governance of self-driving cars. *Social Studies of Science* 48(1): 25–56.

Sutskever, Ilya., Oriol Vinyals and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. arXiv preprint, 1–9.

Thibout, Charles. 2019. La compétition mondiale de l'intelligence artificielle. *Pouvoirs* 3:131–142.

Van Dijck, José. 2014. Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. *Surveillance & Society* 12 (2): 197–208.

Vaswani, Ashish, Noam Shazeer, Niki Parmar et al. 2017. Attention is all you need. arXiv preprint, 1–5.

Vincent, James. 2021. Google is poisoning its reputation with AI researchers. https://www.theverge.com/2021/4/13/22370158/google-ai-ethics-timnit-gebru-margaret-mitchell-firing-reputation. Last access: 10 July 2021.

Vinyals, Oriol and Quoc Le. 2015. A neural conversational model, arXiv preprint, 1–8.

Wagner, Ben. 2018. Ethics as an Escape from Regulation: From Ethics-Washing to Ethics-Shopping? In *Being Profiled: Cogitas Ergo Sum*, Ed. Mireille Hildebrandt, 1–7. Amsterdam University Press.

Wakabayashi, Daisuke. 2020. Google chief apologizes for A.I. researcher's dismissal. https://www.nytimes.com/2020/12/09/technology/timnit-gebru-google-pichai.html. Last access: 19 July 2021.

Woolgar, Steve. 1985. Why not a sociology of machines? The case of sociology and artificial intelligence. *Sociology* 19(4):557–572.

Zhang, Jerry, Darrell Carpenter and Myung S. Ko. 2013. Online Astroturfing: A Theoretical Perspective. *Proceedings of the Nineteenth Americas Conference on Information Systems*, 1–7.

Testimonial Injustice in Governmental AI Systems

Catriona Gray

Abstract: *This chapter analyses the application of AI systems which test and/or contest the accounts of human subject(s), and which are applied within the course of governmental decision-making. It argues that the rise in these decisional practices demands thorough interrogation of the ways in which testimony is elicited, offered, and received as an element of AI systems. This enables critical inquiry beyond narrowly conceived ethical categories, allowing for more comprehensive accounts of the range of harms – material and epistemic – produced by systems which bypass, undermine, and challenge the testimony of their targets. I identify the three evidentiary manoeuvres by which testimony figures in various governmental AI technologies: obviation, diminishment and impugnment, and apply the concept of epistemic justice to illuminate the different ways in which harm is produced through their enactment. I argue for a sociotechnical approach which recognizes that resulting testimonial injustices are not easily addressed by the cultivation of more virtuous practices and instead require alternative governance responses. This enables much-needed analysis at the intersection of ethics, epistemology and politics which better equips us to identify new vectors of domination and marginalization, and to imagine and realize less violent alternatives.*[1]

Though often presented as novel and innovative, the use of digital technologies and automation to support governmental decision- and policy-making has emerged over several decades (Henman 2010). These practices have been marked by a recent turn toward greater use of artificial intelligence (AI) systems (Berryhill et al. 2019). This turn has been driven by increases in computational power, higher levels of internet use, alongside increasing computing ubiquity, and greater availability of large datasets and large cloud

1 This work was supported by the Engineering and Physical Sciences Research Council grant EP/S023437/1.

storage facilities. AI systems encompass a range of software-based technologies including machine learning, as well as symbolic approaches based on manipulation of abstract representations of objects and relations (Garnelo/Shanahan 2019). Machine learning, which is now widely used in governmental decision-making (Veale/Brass 2019), includes a range of techniques that allow computational systems to learn directly from data and experience rather than pre-programmed rules. In public-sector contexts, large administrative datasets are used to build machine learning models that in turn support policy and operational decisions, including those which are highly consequential for individuals and their rights. These transformations can be observed at different levels of government across the globe – from municipalities to international organizations (ITU 2021).

Despite this growing adoption, the precise definition of artificial intelligence in law and policy remains contested. The European Union's proposed AI Act (European Commission 2021) is the first attempt anywhere in the world to legislate comprehensively on AI. It broadly defines AI to cover systems which can generate outputs including predictions, recommendations, content and decisions, and which are developed using three approaches listed in its Annexes: (a) machine learning, (b) logic and knowledge-based approaches, and (c) statistical approaches, Bayesian estimation, search and optimization methods. As part of legislative negotiations, member state governments have put forward amendments seeking to narrow the regulation's scope by limiting the definition of AI to systems designed with a level of autonomy to achieve a given set of objectives using machine learning or logic and knowledge-based approaches (Bertuzzi 2022). Civil society and human rights organizations have argued that this move would exclude many rudimentary software systems which can nonetheless have significant impact on people's lives and pose significant risks to their rights (AlgorithmWatch 2021). Rather than limiting inquiry to advanced computational techniques, this chapter considers AI in an expansive sense to include all three approaches listed above, and without a requirement of autonomy.

A more capacious approach is consistent with a sociotechnical analysis which apprehends AI systems not as discrete products or services but as dynamic and constituted by complex webs of actors, code, data and infrastructure. The potential social and ethical implications of a given AI system will depend not just on its technical attributes but on its changing input data, adaptations in its use, and integration with other systems. Many of the risks AI poses to individuals, groups and societies are not inherent in the technology;

they are shaped too by complex development processes and applications across changing contexts. For example, facial recognition technology when used to automate the operation of a coffee machine in a person's home would present a very different range of possible impacts from use of the same product in the policing of political demonstrations.

In this chapter, I analyze the application of AI systems that have three common features. First, they produce *individualized* outputs – including determinations of status, eligibility, entitlements and application of sanctions. Second, they are *adversarial* in that they test or contest the accounts of the human subject (or subjects) to whom they relate. Third, they are applied within the course of *public policy*, including in legal and administrative decision-making and service delivery. I argue that these legal and public policy decisional practices demand thorough interrogation of the ways in which testimony is elicited, offered and received as an element of governmental AI systems. This in turn enables identification and analysis of multiple dimensions of injustice, including, as I will elaborate, epistemic injustice.

Decision-supporting AI systems may be designed, integrated and used within policy practices in ways that undermine particular epistemic values (such as scrutability and explainability) and that give rise to unjust harms against specific individuals and groups in their capacity as knowers. As various thinkers within social epistemology have elucidated, some forms of injustice arise from wrongs against people in their specific capacity as knowers (Fricker 2007; Pohlhaus 2014). These are *epistemic injustices*. Drawing on these theoretical currents, I address a major shortcoming in dominant conceptualizations of ethics and responsibility in AI. Attending only to narrowly conceived ethical and legal categories (including those of fairness and bias) results in a failure to account for the full range of (material and epistemic) harms produced by systems which bypass, undermine and challenge the testimony of their targets. There is a nexus of ethics, epistemology and politics in need of critical attention from scholars concerned with the social implications of AI and related technologies.

The chapter proceeds as follows. First, I introduce the concept of epistemic – including testimonial – justice. After identifying the three main ways in which testimony figures in various adversarial governmental AI systems, I illuminate the different ways in which epistemic injustice can be produced through the development and deployment of AI systems in public policy settings. Whilst the cultivation of epistemic virtues (such as open-mindedness) is often proposed as a means of addressing epistemic injustices, I show that

this will fall short when it comes to governmental decision-making supported by AI.

1. Epistemic injustice and AI

The reality that social life increasingly unfolds across digital environments requires us to rethink theories and applications of epistemic justice (Origgi/Ciranna 2017). Miranda Fricker's pathbreaking book *Epistemic Injustice* (2007) offers an account of how unequal social relations inflect what gets to count as knowledge and who gets to count as a credible knower or epistemic agent.[2] These unequal social relations may include, for example, positional categories such as gender, or the relationship between a public authority and a person in receipt of social assistance. Fricker distinguishes between what she terms *testimonial injustice* and *hermeneutical injustice*. The former occurs when one attempts to convey knowledge ("prejudice causes a hearer to give a deflated level of credibility to a speaker's word") and the latter at a prior stage when we attempt to make sense of our own social experiences ("a gap in collective interpretive resources puts someone at an unfair disadvantage when it comes to making sense of their social experiences") (2007: 1).

For Fricker, the central case of systematic testimonial injustice rests on what is deemed "identity-prejudicial credibility deficit" (2007: 28). A credibility deficit can be a result of innocent error which is "both ethically and epistemically non-culpable" (2007: 21) but it will not meet the threshold of an *injustice* unless there is an element of the ethical wrong of prejudice.[3] In essence, testimonial injustice describes the accordance of lower credibility to a speaker or a knower because they belong to a group that has been typecast in some way. Alert to the dehumanizing effect of testimonial injustice, Fricker argues that to cause a person to suffer it is to degrade them not just qua *knower* but also, symbolically, qua *human*. Moreover, she suggests that persistent and systematic cases may also "genuinely inhibit the development of an essential aspect of a person's identity" (2007: 54) and even exercise social

2 It must be noted that Fricker was also working with ideas drawn from a feminist standpoint, and postcolonial and other critical theories (Spivak 1987; Hill Collins 1998).

3 A prejudice can either be one which is relatively incidental and localized or one which 'tracks' the subject across different contexts (e.g., homophobia), giving rise to different forms of injustice.

constructive power in a way that "constrains who the person can be" (2007: 58). Other thinkers seeking to understand epistemic injustice (or violence) have placed less emphasis on prejudice, and instead conceptualized, for example, "pernicious ignorance" (Dotson 2011).

Turning to the other limb of epistemic injustice, hermeneutical injustice, its features appear quite different, but it is also constituted by a wrong to a person in their capacity as a knower. It precedes testimonial injustice and is less directly a matter of credibility determination. The underlying rationale is that, where experiences are occluded, a knower is unable to make sense of them in order to give testimony and to come up with ways of challenging the prevailing social order. Whereas the powerful have at their disposal coherent interpretations of social reality which work to their advantage, others are marginalized in multiple ways by their social location, making their resistance less possible.

Digitally mediated environments are not just spaces in which epistemic injustices can occur or become exacerbated; they can generate novel and distinctive epistemic injustices. Noting the relative neglect within wider critical analysis, Symons and Alvarado argue for scholarly light to be shed on the specifically *epistemic* harms and injustices arising from the "design, development and deployment of complex and opaque data-driven technologies such as machine learning, deep neural networks, and big data analysis" (Symons/Alvarado 2022). According to the authors, the main reason hermeneutical and testimonial epistemic injustices arise through the operations of AI and related technologies is "their opacity and their inability to permit corrective recourse" (2022: 92). Earlier scholarship had already identified the "epistemic opacity" of computational systems and processes (Humphreys 2004) – a feature which has become even more pronounced in some, though not all (Felzmann et al. 2020), AI systems, leading in some cases to epistemic injustice (Alvarado and Humphreys 2017). Taking up these insights, I examine three evidentiary manoeuvres which are reconfiguring the role of testimony (namely obviation, diminishment and impugnment), and the epistemic justice implications of their enactment.

2. The concept of testimony in AI

Although philosophers may disagree about its exact nature – including whether it generates or merely transmits epistemic properties – testimony can be distinguished from other ways of justifying knowledge such as mem-

ory, perception and reason (Audi 1997). Within epistemology, scholars have attempted to distinguish testimony from mere assertion. Coady (1992) suggests that, to count as testimony, a speech act must meet several conditions, including: that it is being offered as evidence of a proposition; that the speaker has competence, authority and credentials to make such a statement; and that it is relevant to some unresolved question about which those hearing the testimony are in need of relevant evidence. A revised version of this definition which hinges on the speaker *purporting to convey information* is put forward by Graham (1997). According to this view, what matters is that the speaker *intends* their audience to believe they embody the relevant qualities, and that *the speaker believes* their statement of a proposition to be relevant to a question they believe is unresolved and about which they believe those hearing it are in need of relevant evidence.[4] As Freiman and Miller (2020) suggest, however, a charge of anthropocentrism may be levelled against this human cognition-dependent conception of testimony and assertion. According to them, some instrumental (including AI-generated) outputs can also constitute a kind of *quasi-testimony*.

The adoption and use of AI to produce outputs relating to specific people requires us to re-evaluate not just the role of the speaker (the person or entity giving testimony), but that of the hearer (the person or entity receiving testimony). If we consider the status of an artificial entity as a potentially epistemically and ethically responsible agent – what we might refer to as the "problem of technological agency" (Rosenberger 2014) – we know that a machine cannot *hear* speech as testimony, nor can it alone truly form a belief (as opposed to producing an output or reaching a conclusion). However, it does not follow that we can analyse only the epistemic activities of the *human* agents who design, use and interpret AI systems. Instead, we can undertake a more sociotechnical form of analysis (Sartori/Theodorou 2022) which privileges neither the human nor the machine (nor the data) and which regards these elements, and the epistemic properties they enact and transmit, as interrelated.

As is now well known, there has been an expanded use of digital technologies within processes designed to verify and evaluate evidence, and to adjudicate claims made by citizens regarding, for example, their eligibility for a ser-

4 The concept of 'bent testimony' has been developed to capture scenarios in which, because of disputed norms of communication, we may treat search algorithms as if they were asserting true content at the top of a search page – even in the absence of assertion (Rini 2017; Narayanan/De Cremer 2022).

vice or benefit. Testimony often features in these systems in numerous complex ways, often in combination. Below I cite several examples of AI technologies applied in public policy settings – including in migration, social security and policing – and describe their treatment of testimony within the three-part typology referred to above. This is not intended to be exhaustive but rather to highlight that analysis of the epistemic properties and ethical stakes involved in the use of AI systems requires consideration of their purposes and operations. Whilst these three categories are not entirely discrete, each corresponds to a specific evidentiary manoeuvre. In the first (obviation), testimony is largely discarded as a source of evidence. In the second (diminishment), testimony is not discarded but rather devalued in the face of alternative or contradictory evidence. In the third (impugnment), there may be no alternative or contradictory evidence at all, and the system is used instead to evaluate the quality of testimonial evidence.

2.1 Obviating testimony

In this grouping, direct testimony is largely sidelined. Rather than eliciting self-reported accounts of a subject's identity, biography or predicament, these systems tend to generate or extract information about a subject without their direct input and/or awareness. In such cases, sources of knowledge justification other than testimony are used. This may include biometric information derived using computer vision systems which purport to recognize an individual's intimate characteristics such as sexuality, personality traits or emotional state. This information is often based on analysis of data inputs deriving from the body (such as gait, facial geometry, eye gaze), and electroencephalogram (EEG) (Seo/Laine/Sohn 2019), and training data based on the socially conditioned human perceptions of these traits. For the most part, these systems have been developed in research settings, and have not been deployed by democratic governments. Many have been described as pseudoscientific and a form of physiognomy (Stark/Hutson 2022; Kaltheuner 2021). However, even less controversial classification systems, such as those which claim to identify a person's gender, rely on the notion that "social systems of classification have an essence; an intrinsic, biological substrate that can be detected via the face once-and-for-all with the assistance of 'objective' automated tools" (Scheuerman/Pape/Hanna 2021).

As well as AI systems deployed to extract information directly from the body, many AI systems make inferences about a person based on biographi-

cal and circumstantial details. Angwin et al. (2016) investigated use of such a system by state authorities in the United States which aimed to predict the recidivism risk associated with specific individuals based on scores derived from 137 questions answered by defendants or taken from their criminal records. They found significant racial bias. As well as inference, in this type of encounter there is also a strong element of memory as an epistemic source, as it is heavily dependent on data related to a person which is stored and then retrieved. In this scenario, even though subjects are being asked for their direct input to complement information already held about them – that is, they are allowed to speak – the process is conducted in a highly mechanical and circumscribed way. Participants may be able to guess how their answers will effectively count against them and expose them to harsher treatment, but they are not informed about how their responses will be analysed or what inferences are being made about them. Even if there is a small element of verbal testimony, the exchange is highly imbalanced, and respondents are not given the opportunity to know what it is they are being asked and why. As Hildebrandt notes, being profiled by a machine, without access to the knowledge used to categorize us, means we "cannot adequately anticipate the actions of those that know about us what we may not know about ourselves" (Hildebrandt 2008: 17). A person being asked, for example, simply how often they used to get into fights at school is not able to give contextual information or to know which unresolved question is regarded by the hearers as in need of evidence (and so their answer would not qualify as testimony in the schema outlined by Graham [1997] above). The opacity of many AI systems and their outputs can effectively curtail opportunities for subjects (and their legal representatives) to review decisions (Cobbe 2019) or to adapt legal and advocacy strategies based on clear decision-making criteria. The latter problem arises particularly where statistical and logical models may be manipulated by public servants, leading to irrational outcomes (Koulish/Evans 2021).

In the examples of systems which deal with testimony in ways that effectively bypass it in favour of perception, memory and inference, it may not be immediately obvious where the epistemic injustice arises. If testimony is not being elicited and someone is not engaged in their capacity as a knower, one might assume no epistemic harm is done. But it is in these situations that a particularly grave form of epistemic injustice may occur. AI systems which generate and process information about a person – but which do not allow for acknowledgement of them as a knowing subject – produce what many would regard as epistemic objectification. That is, people are treated as sources of information, rather than informants. Here, our knowledge of ourselves and the

world is not so much regarded as unreliable as it is deemed superfluous, unfathomable or even irrelevant.

The above describes a failure (by public authorities) to respect the dignity and autonomy of the targets of AI systems *in the process of producing knowledge*. However, epistemic harms cannot be separated from the character of *the knowledge itself*. In many cases, particularly where machine learning is used, the knowledge production enabled by an AI system is based on probability, not actuality. Rather than being designed to enable the formation of a genuine belief about a person, the purpose is to produce calculative inferences about that person which can serve as the basis for action. This type of knowledge is based on prediction rather than theorization and is less concerned with whether something is true than with the cost of acting as if it were so (Joque 2022). As Origgi and Ciranna put it, people are "epistemically diminished as individual knowers: the knowledge of themselves objectifies them in a new way. Their identity becomes a virtual object, a 'statistical double'" (2017: 305).

Vallor (2021) argues that certain AI technologies are being developed to 'scrape' the body for unconsented and supposedly unmediated emotional 'truth', in a way that mirrors the *basanos* of Ancient Greece (a method of extracting truthful testimony through torture from the bodies of enslaved people). Such technologies, which include emotion recognition systems, are typically deployed against those without the power or knowledge needed to consent (and, conversely, to refuse). In Vallor's reading, rather than testimony being ignored entirely, there is a displacement from speech and communication toward the body as a site of truth.

Departing from *objectification* as the lens for understanding the wrongs of epistemic injustice, Cusick (2019) argues that the concept of *derivization* is more productive and precise. Here, following Cahill (2011) and Pohlhaus (2014), the core wrong is not the treating of others as mere (bodily) objects (as objectification entails), but the "active, willful misinterpretation of the evidence from victims' own bodies and lives and a derivatizing of them as persons for others rather than for themselves" (Cusick 2019: 112). This allows analysis of how bodies (and not just words) are treated as sources of information to serve others' ends. Bypassing verbal testimony to read information from a body is not always wrong (for example, in some medical settings). However, it can become wrong when listeners treat themselves as the only active participants in testimonial exchanges. Derivization, then, may occur when AI systems are used in ways that obviate verbal testimony to treat persons as for others rather than for themselves. Though the possibility of derivization is already present in adver-

sarial adjudicative settings without the introduction of AI, the intersubjective qualities of decision-making (Bergman Blix 2022) are reduced, or even eliminated, by automation. AI limits potential opportunities for subjects and their representatives to intervene with more narrative accounts and context-driven argumentation.

2.2 Diminishing testimony

Critics of techno-solutionist thinking contend that it results in over-reliance and excessive trust in quantified methods and outputs. In this vein, Broussard (2019) identifies a pervasive tendency to assume that technological solutions are inherently better than their alternatives, a tendency she calls *technochauvinism*. What is less theorized, however, is the extent to which the excessive credibility accorded to and reliance on AI systems and their outputs correspond (necessarily or in effect) with a deflation of the testimony of their targets.

Empirical studies into decision-making in human–algorithm interaction have identified distinct biases in how outputs are processed by decision-makers. Automation bias, according to Alon-Barkat and Busuioc (2022), consists in the "human propensity to automatically defer to automated systems, despite warning signals or contradictory information from other sources" (2022: 2). The British Post Office scandal provides an illustrative example of the potential consequences of excessive trust in or reliance on automating digital technology, including how this may diminish the relative position of testimony. Over the course of several years, hundreds of workers were wrongly prosecuted for theft, false accounting and fraud because of discrepancies arising from a flawed software system. The courts later found that the Horizon system was wrongly represented as reliable and its outputs as incontrovertible – effectively reversing the burden of proof so that the onus was on the defendants to prove that no losses had occurred (Wallis 2021). Whether or not all actors truly believe in the integrity and veracity of their outputs, we can identify discursive processes that give AI-enabled systems a type of social power (Beer 2017) and that in turn may make it easier for actors to deliberately rely on systems which appear to be flawed or for which malfunction is foreseeable. Where computational systems are widely regarded (or at least treated) as embodying a special, authoritative form of reliability or trustworthiness, any countervailing testimonial evidence will hold less power.

Distinct from automation bias, *selective adherence* is the propensity "to adopt algorithmic advice selectively, when it matches pre-existing stereotypes

about decision subjects (e.g., when predicting high risk for members of negatively stereotyped minority groups)" (Alon-Barkat/Busuioc 2022: 2). Whilst automation bias has been identified in studies in social psychology, in this study it was found that public servants do not necessarily tend to defer to automated system outputs (Alon-Barkat/Busuioc 2022). However, the authors found that decision- makers may be likely to adhere to decision recommendations when they align with existing group stereotypes and disadvantage minority groups (hence *selective adherence*). This was identified in the case of the infamous discriminatory AI enabled decision-making system which was used by the Dutch government to flag potentially incorrect or fraudulent childcare benefit claims. Decision-makers operated with the perverse incentive of ensuring more money would be retrieved through the scheme than the system itself would cost (Amnesty International 2021).

As we have seen, the classic case of testimonial injustice put forward by Fricker is one of identity-based credibility deficit i.e., a situation in which a person's testimony is accorded decreased (or, in some cases, increased) credibility not because of any relevant factors, but on the basis of prejudices. In Fricker's original theorization of testimonial epistemic injustice, interpersonal discrimination within localized interactions was taken as a basic premise. However, if we follow subsequent contributions (Coady 2017; Medina 2011) in viewing credibility as a relational or distributional (and even finite) epistemic good, excessive trust or reliance in technologies and their outputs may engender changes to the handling of subjects' testimony. In such cases, testimonial injustice arises from the assumption that technological outputs are more valid – as in automation bias (Alon-Barkat/Busuioc 2022; Symons/Alvarado 2022). Where decisions confer access to finite resources or are constrained by management targets and incentives, decision-makers may also experience pressure to adjust their evidentiary burden to, for example, allow only the most straightforward or convincing claims to succeed, or to lower the thresholds for flagging potential fraud. In this way, the superficial *objectivity* (Porter 1995) of AI systems may generate credibility and (at least temporarily) ward off scrutiny. As this reminds us, the credibility given to quantified methods and reasoning is not just an effect of perceptions about technological capability and reliability – it is deeply political (Rose 1991). As Alon-Barkat and Busuioc (2022) suggest, in the wake of high-profile events such as the Dutch childcare benefit fiasco, public servants' attitudes and behavior are likely to reflect greater scepticism and even diligence with regard to AI decision-making tools.

2.3 Impugning testimony

In some cases, AI tools are used to directly evaluate and verify the integrity of a speaker's testimony. This is subtly different from diminishment in that it relates to the quality of the testimony, rather than how much weight should be attached to it. This treatment of testimony is typified by the iBorderCtrl project which aimed to develop systems with the ability to perform "deception detection" based on facial recognition technology and the measurement of micro-expressions by finding so-called "biomarkers of deceit" on the bodies of people attempting to cross borders. The controversial project – funded by the European Commission (European Commission 2022) – was designed with the aim of speeding up border control of third-country nationals crossing EU borders by providing authorities with information to support decision-making. Psychologists have, however, refuted the premise of such a system and argued that, without definitive and reliable cues to deception, its validity is highly questionable (Jupe and Keatley 2019).

Sánchez-Monedero and Dencik (2020) situate iBorderCtrl within a lineage of lie detection and data-driven deception detection technologies, arguing that these systems perform not just technical but distinctly political functions. Other systems which use behavioral data (such as keystrokes) or language to assess credibility and detect fraud have been developed and continue to attract investment (Bittle 2020). Although many of them only flag cases for further attention – rather than offering a final determination of honesty or deception – their use signals an intensification of the level of scrutiny applied to testimony within individualized decision-making. Despite well-grounded concerns about their validity and accuracy, these systems at least claim to make available a wider range of behaviors, physical features and communicative traces which can be used to test the credibility of testimony offered by their targets.

Many other instances of decision-making in public policy are characterized by suspicion and high burdens of proof. This is particularly apparent in asylum adjudication processes, where applications often hinge on the credibility accorded to the asylum seeker's account of their own identity and biography. Information including state-produced country reports is often accorded more authority than the testimony of the asylum seeker and is used to impugn their narrative accounts (Haas and Shuman 2019). This has been accompanied, at times, by widespread use of invasive and rights-violating practices to determine credibility – particularly in asylum applications related to sexual orien-

tation or gender identity (Spijkerboer 2013). As this suggests, the introduction of digital technologies to assess credibility often takes place against a backdrop of unjust epistemic practices by public authorities which impugn the accounts of individuals (Sertler 2018).

As noted already, a point of disagreement amongst scholars about the nature of testimonial injustice is whether credibility can be assessed by viewing one speaker at a time. According to Fricker's original theorization, credibility is not a finite resource and so credibility *deficit* is not *inversely proportional to credibility excess; it can be looked at in relative isolation*. Medina (2013), however, convincingly argues that we cannot separate subjects from their own social positionality. This is because our judgements about what is "normal" tend to take shape in "comparative and contrastive" ways (Medina 2013: 66) – for example, queer identities becoming known in part through a series of oppositional binaries (Sedgwick 2008). For this reason, a subject may be unfairly assessed as lacking credibility largely because of a comparison with the normal (Medina 2013: 63). Indeed, this process of comparison between objects is a core feature of many models based on machine learning through which "the 'other' is algorithmically produced as anomaly" (Aradau/Blanke 2018: 1).

In systems which purport to detect deception using machine-learning algorithms trained from large datasets, epistemic injustice may also arise from the exploitation of one's data – even when accorded the appropriate level of credibility. If my testimony is justifiably deemed credible and personal data about me (such as my facial expression at the time) is then then used to help build a model to impugn the testimony of others, we might think of this unconsenting exchange as a kind of *epistemic extraction* (Pasquinelli/Joler 2021). People whose data correspond to what is (or will be) deemed 'normal' are compelled to offer testimony that may be combined and used to train an AI system which could automate and reinforce further epistemic and material harms. Here, the qualities of 'credible' testimony are alienated from its content, and from the speaker, and used to generate markers of credibility through ongoing calculations of similarity and anomaly. As a result, it becomes impossible to refuse testimonial participation in the production of unjust determinations and decisions. AI systems which are used to impugn testimony therefore add to the epistemic injustices already present in, for example, asylum adjudication processes by making these contrastive relationships between excess and deficit – or normal and anomalous – both more pronounced and less open to contestation.

3. Epistemic virtue and governmental AI

To address instances of epistemic injustice, a set of correctives rooted in virtue ethics has been proposed. According to Fricker, a virtuous hearer must exercise critical awareness so that they can identify the impact of identity power in their credibility judgement. This must address both the identity of the speaker and the hearer's own position and be an ongoing, reflexive process: "someone whose pattern of spontaneous credibility judgement has changed in light of past anti-prejudicial corrections and retains an ongoing responsiveness to that sort of experience" (Fricker 2007: 97). This set of sensibilities, taken together, makes up the hybrid virtue of testimonial justice.[5]

As for the corrective to hermeneutical injustice, a virtue is again proposed – *hermeneutical justice*. However, the task of overcoming unequal power relations that cause situations of hermeneutical marginalization (and injustice), Fricker admits, "takes more than virtuous individual conduct of any kind; it takes group political action for social change" (2007: 174). Medina (2013) takes up this suggestion, arguing that in order to work toward epistemic justice, we must cultivate sensibilities that encourage us to be open to, and actively in search of, sources of contestation and epistemic friction. To counteract the epistemic vices of arrogance, laziness and closed-mindedness, virtues of humility, diligence and open-mindedness are proposed instead.

Epistemic injustice cannot be separated from unjust social structures more broadly – including, for example, racism, socioeconomic inequalities and disablement. For Medina (2011), it is at the level of the *social imaginary* that epistemic injustice is most deeply rooted, and countering it is largely a task of re-imagination and of determining what can count as epistemic alternatives. Credibility deficits affecting marginalized groups, then, are not always underpinned by some form of prejudice; there are several *structural* causes. These include differential access to markers of credibility, particularly through educational and distributive inequalities. Examining only the local properties of interactions and decisions therefore may not reveal all injustices. This is because cognitive biases can be *transactionally* innocent but nevertheless act as vehicles for the spread of *structural* injustices. As Anderson argues, just as individuals are accountable for how they act independently, they are accountable

5 Correcting for prejudice is hybrid because it is necessary in order both not to miss out on the truth and to avoid commission of unjust practices to a person in their capacity as a knower, according to Fricker (2007).

for how they act collectively: "Epistemic virtue is needed at both individual and structural scales" (Anderson 2012: 171). She argues for a move away from what she regards as Fricker's preoccupation with individual epistemic virtue toward a consideration of epistemic justice as a virtue of what she terms *social institutions*. Institutions, she argues, may have the power to prevent or correct problems that virtuous individuals cannot solve on their own – particularly where unjust outcomes result from complex and cumulative inputs and operations. Others have similarly argued that organizations can cultivate virtues of open-mindedness, courage, integrity and humility (Choo 2016). However, from a social scientific perspective, *pace* Anderson, the concept of institution ought not to be used synonymously with either *organization* or *social structure* (Fleetwood 2008). One must also be careful not to commit what Archer (1995) calls *upwards conflation*, i.e., the attribution of causal efficacy only to agents but not to structure (which is instead conceived of as a mere aggregation of agentive forces).

Nevertheless, if epistemic virtues can and should be cultivated as collective and organizational values, we can inquire into how this might extend to governing arrangements involving AI to support decision-making. In the following section, I show that efforts to cultivate virtues such as open-mindedness and empathy as a means of regulating the operation and minimizing negative effects will encounter four key constraints. These constraints emanate from the complex embeddedness of AI in governing practices (Carmel 2019). AI cannot be understood as a type of isolated, standalone product or service. Rather, the ethical and political implications will vary greatly depending on the context, purpose(s), operation and use. Where the responsible parties procuring, deploying and operating systems are public authorities, the effects on the relationships between citizens and the state, in particular, are significant. I set out four dimensions that require consideration when addressing the epistemic injustices outlined above. These relate to: the political economy of AI adoption; the distribution of agency and responsibility; the nature of bureaucratic rule; and decision-making knowledge in government.

3.1 The political economy of AI adoption

Though many thinkers already agree that epistemic injustices are a structural problem and, as such, in need of a structural response (Samaržija/Cerovac 2021), this proposition alone does not identify which structures are relevant and in which contexts. While it is widely recognized that social structures,

including oppressive social relations such as those of race (Mills 2017) and gender, shape epistemic injustices, there is a need to attend at the same time to specific political economic contexts. These technologies which reconfigure the role of testimony within the exercise of public power are not outcomes of discrete and spontaneous policy decisions. They emerge under particular conditions, with particular political rationales, and are often constituted by complex interactions between public and private interests. This further complicates the task of imbuing systems, their use, and their oversight and accountability mechanisms with epistemic virtue. Attending only to epistemic practices leaves unaddressed many *extra-epistemic* forces involved in the production of epistemic harms outlined in the typology above (obviation, diminishment, impugnment). This is where democratic politics (rather than epistemic virtue alone) may be a more suitable means of addressing injustices.

3.2 The distribution of agency and responsibility

Much thinking and writing on epistemic justice proceeds on the basis that some knowers are dominantly situated in relation to others. It is widely accepted that technologies are never apolitical or value-neutral (Winner 1980), and AI's propensity to reproduce social inequalities is well documented (O'Neil 2017; Noble 2018; Buolamwini/Gebru 2018). However, where beliefs are formed – or conclusions established – not simply by individuals who hear testimony and weigh it up with other sources of knowledge, but in processes distributed across sociotechnical systems, questions of domination are less clear cut. Digitalization and automation disorganize established decision-making practices and procedures in ways that may make it difficult to locate ethical and political agency and sources of harm. Ananny and Crawford (2018) highlight the *assembled* nature of algorithmic systems which, rather than being reified objects, are made up of a network of human and computational actors. Given that agency and responsibility are distributed and relational, any attempts to foster virtue must target not just developers, operators, users or the system components, but the entire assemblage. The complexity of AI supply chains complicates this task. Many systems are derived from multiple sources or come already embedded in goods and services. Allocation of legal responsibility throughout the AI supply chain and lifecycle is central to current debates about the regulation of AI. These considerations could be expanded to include sites (and even chains) of epistemic responsibility.

3.3 The nature of bureaucratic rule

As I have suggested, the character of the institutional entity responsible for making decisions – in most of the cases discussed, state agencies – is significant. Indeed, there is a specifically *bureaucratic* character to many of the epistemic harms discussed in this chapter. This comes with important analytical implications. As Du Gay (2000) notes, bureaucratic conduct is frequently regarded as irremediably unethical. This tendency is perhaps most acute in the work of Hannah Arendt (1969), who argued that bureaucracy can be thought of as *rule by nobody*. Similarly, in *Modernity and the Holocaust*, Bauman (1989) examined the nature of bureaucracy as representative of modernity, and as characterized by its functional division of labor and separation of the technical from the moral. Bureaucracy, for Bauman, is dehumanizing because it produces distance between the (ethical) conduct of decisions and their effects and outcomes. It has even been observed that machine learning and bureaucracy are both modes of goal-oriented, rational ordering that claim neutrality and objectivity through detached abstraction (McQuillan 2020). While we might not go as far as to claim they are coextensive, there is, as this suggests, much promise in understanding the application of machine learning techniques as cohering with bureaucratic rule. The ignorance, deflation or disbelief of citizen testimony is not simply a result of negative aggregated attitudes or inadequate hermeneutical resources. The widespread use of automating technologies means subjects are also *epistemically* disadvantaged by their *materially* subordinated position as people who rely on bureaucratic formations to meet their basic needs. It is only through extension of substantive political and socioeconomic rights, not just improved epistemic practices, that these unjust power asymmetries can be redressed.

3.4 Decision-making knowledge in government

Recent scholarship has begun to examine the knowledge practices of street-level bureaucrats (Lipsky 2010) who use algorithmic decision tools (Snow 2021), including how they exercise (or withhold) individual judgement. Ranchordás (2022) argues that *empathy* ought to be seen as a key value within administrative law, and, accordingly, that the digitalized state ought to consider multiple viewpoints as well as individual circumstances (2022: 45). She argues that there ought to be a duty to forgive errors in some cases – a duty not well served by automation and data-driven decision support. Governmental AI technologies

are often accompanied by the curtailment or removal of administrative discretion (Bovens and Zouridis 2002; Eubanks 2018), and their design and development may overlook the plurality of forms and practices of knowledge involved in policy- and decision-making processes, including implicit or tacit knowledge (Polanyi 1962). Attempts to automate or support decision-making using algorithmic outputs may challenge capacities to adapt and improvise in unpredictable circumstances, and potentially subordinate idiosyncratic and uncodified ways of knowing and acting. As a result, there may be less scope for decision-makers to work with knowledge in more reflexive and virtuous ways. In other words, just as fairness cannot be automated (Wachter et al. 2021), nor can virtue.

Similarly, we cannot assume those deploying AI aim to accumulate as much knowledge as possible, or that such knowledge has a direct correlation with power. The work of Linsey McGoey (2012), and that of other scholars concerned with the sociology of ignorance, has illuminated the ways in which "strategic unknowns" can be harnessed as resources which enable knowledge to be deflected, obscured or manipulated to expand the scope of what remains unintelligible. Institutions employ strategies to keep "uncomfortable knowledge" at bay (Rayner 2012: 107). Rather than being the result of a flawed design or ineffective decision-making, the exclusion of countervailing testimonial knowledge may be driven by strategic imperatives.

Conclusion

Much of the critical literature on the use of AI in government has detailed the ways in which the design and use of technologies that automate decision-making processes can produce significant harmful, discriminatory and rights-violating outcomes for their targets (Eubanks 2018; O'Neil 2017). As well as these individualized and collective adverse outcomes, the use of automated decision-making in the public sector is altering how power, authority and knowledge are arranged within and between institutions. Scholars have begun to argue for a shift away from narrowly conceived notions of AI ethics toward wider consideration of justice (Gabriel 2022). In this chapter, I have advanced this further by showing that the use of AI is reconfiguring how testimony is elicited, offered and received, and that this is giving rise to specifically epistemic injustices. Through the use of AI technologies, people are made legible to the state (Scott 2020) in ways that silence, derivize and extract from them. By variously

obviating, diminishing and impugning the testimony of individuals targeted by decisions, these systems are not just harming them by undermining their dignity as knowing persons; they are reshaping the place of testimony in public policy, with potentially far-reaching implications for democratic citizenship.

Whereas much existing scholarship posits the cultivation of epistemic virtue as one way to prevent and address epistemic injustices, the nature of (AI-supported) governmental decision-making means this is insufficient as a mode of prevention or redress. The drivers and effects of epistemic injustice in these cases extend well beyond unjust structures of knowledge. AI systems are being used in processes already designed to be impersonal, highly rationalized and not always amenable to epistemic virtues. Furthermore, the nature of many systems precludes the identification of a single locus of bias (or epistemic vice). Whether demanded of the human in the loop or the "institution" in the loop, reflexive and virtuous practices cannot resolve these problems.

How then might we attempt to reposition testimony within governmental decision-making to prevent and address such injustices? Many of those most implicated by these practices are already politically disempowered, and digitalization and automation can produce further exclusions which limit their capacity to contest and advocate. Nevertheless, as recent political and legal mobilizations and initiatives have demonstrated, there are opportunities for collective redress, resistance and refusal (Ganesh/Moss 2022; Dent 2022), as well as formal regulation. Alongside these movements for greater democratic oversight and control of AI, various public accountability measures such as public registries and impact assessments have been proposed and implemented to different degrees (Ada Lovelace Institute et al. 2021). To address the types of injustices outlined in this chapter, however, AI policy-making and regulation must go further in centering the perspectives and guaranteeing the rights of decision subjects themselves.

Bibliography

Ada Lovelace Institute, AI Now Institute and Open Government Partnership. 2021. Algorithmic Accountability for the Public Sector. https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector. Last access: 8 October 2022.

AlgorithmWatch. 2021. EU policy makers: Protect people's rights, don't narrow down the scope of the AI Act! *AlgorithmWatch*. https://algorithmwatch.org/en/statement-scope-of-eu-ai-act/. Last access: 5 October 2022.

Alon-Barkat, Saar and Madalina Busuioc. 2022. Human–AI interactions in public sector decision making: "Automation Bias" and "Selective Adherence" to Algorithmic Advice. *Journal of Public Administration Research and Theory*. https://doi.org/10.1093/jopart/muac007.

Alvarado, Rafael and Paul Humphreys. 2017. Big Data, thick mediation, and representational opacity. *New Literary History* 48:729–749. http://doi.org/10.1353/nlh.2017.0037.

Amnesty International. 2021. Xenophobic machines: Discrimination through unregulated use of algorithms in the Dutch childcare benefits scandal. *Amnesty International*, 25 October 2021. https://www.amnesty.org/en/documents/eur35/4686/2021/en/. Last access: 3 October 2022.

Ananny, Mike and Kate Crawford. 2018. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society* 20:973–989. https://doi.org/10.1177/1461444816676645.

Anderson, Elizabeth. 1995. Feminist epistemology: An interpretation and a defense. *Hypatia* 10:50–84. http://doi.org/10.1111/j.1527-2001.1995.tb00737.x.

Anderson, Elizabeth. 2012. Epistemic justice as a virtue of social institutions. *Social Epistemology* 26:163–173. https://www.tandfonline.com/doi/abs/10.1080/02691728.2011.652211.

Angwin, Julia, Jeff Larson, Surya Mattu and Lauren Kirchner. 2016. Machine bias. *ProPublica*, 23 May 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Last access: 3 October 2022.

Aradau, Claudia and Tobias Blanke. 2018. Governing others: Anomaly and the algorithmic subject of security. *European Journal of International Security* 3(1):1–21. https://doi.org/10.1017/eis.2017.14.

Archer, Margaret. 1995. *Realist Social Theory: The Morphogenetic Approach*. Cambridge: Cambridge University Press. http://doi.org/10.1017/CBO9780511557675.

Arendt, Hannah. 1969. A special supplement: Reflections on violence, *New York Review*. 27 February 1969. https://www.nybooks.com/articles/1969/02/27/a-special-supplement-reflections-on-violence/. Last access: 3 October 2022.

Audi, Robert. 1997. The place of testimony in the fabric of knowledge and justification. *American Philosophical Quarterly* 34:405–422.

Bauman, Zygmunt. 1989. *Modernity and the Holocaust.* Cambridge: Polity.

Beer, David. 2017. The social power of algorithms. *Information, Communication & Society* 20:1–13. https://doi.org/10.1080/1369118X.2016.1216147.

Bergman Blix, Stina. 2022. Making independent decisions together: Rational emotions in legal adjudication. *Symbolic Interaction* 45(1):50–71. https://doi.org/10.1002/symb.549.

Berryhill, Jamie, Kévin Kok Heang, Rob Clogher and Keegan McBride. 2019. Hello, World: Artificial intelligence and its use in the public sector. *OECD Working Papers on Public Governance*, 36. 21 November 2019. https://doi.org/10.1787/726fd39d-en.

Bertuzzi, Luca. 2022. AI Act: Czech Presidency pushes narrower AI definition, shorter high-risk list. *Euractiv,* 18 July 2022. https://www.euractiv.com/section/digital/news/ai-act-czech-presidency-pushes-narrower-ai-definition-shorter-high-risk-list/. Last access: 5 October 2022.

Bittle, Jake. 2020. Lie detectors have always been suspect. AI has made the problem worse. *MIT Technology Review.* 13 March 2020. https://www.technologyreview.com/2020/03/13/905323/ai-lie-detectors-polygraph-silent-talker-iborderctrl-converus-neuroid/. Last access: 5 October 2022.

Bovens, Mark and Stavros Zouridis. 2002. From street-level to system-level bureaucracies: How Information and Communication Technology is transforming administrative discretion and constitutional control. *Public Administration Review* 62:174–184. https://doi.org/10.1111/0033-3352.00168.

Broussard, Meredith. 2019. *Artificial Unintelligence: How Computers Misunderstand the World.* Cambridge, Mass.: MIT Press.

Buolamwini, Joyce and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine Learning Research* 81:1–15. 2018 Conference on Fairness, Accountability, and Transparency. https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf. Last access: 20 December 2022.

Cahill, Ann J. 2011. *Overcoming Objectification: A Carnal Ethics*. New York: Routledge.

Carmel, Emma. 2019. Introduction to Governance Analysis: Critical Enquiry at the Intersection of Politics, Policy and Society. In *Governance Analysis: Critical Enquiry at the Intersection of Politics, Policy and Society*, Ed. Emma Carmel, 2–24. Cheltenham: Edward Elgar Publishing.

Choo, Chun Wei. 2016. *The Inquiring Organization: How Organizations Acquire Knowledge and Seek Information*. New York: Oxford University Press.

Coady, Cecil Anthony John. 1992. *Testimony: A Philosophical Study*. Oxford: Oxford University Press.

Coady, David. 2017. Epistemic Injustice as Distributive Injustice. In *The Routledge Handbook of Epistemic Injustice,* Eds. Ian James Kidd, José Medina and Gaile Pohlhaus, 61–68. London: Routledge.

Cobbe, Jennifer. 2019. Administrative law and the machines of government: Judicial review of automated public-sector decision-making. *Legal Studies* 39(4):636–655. http://doi.org/10.1017/lst.2019.9.

Cusick, Carolyn, M. 2019. Testifying bodies: Testimonial injustice as derivatization. *Social Epistemology* 33:111–123. https://doi.org/10.1080/02691728.2019.1577919.

Dent, Anna. 2022. Disabled benefits claimants are being unfairly investigated. *Huck*, 16 March 2022. https://www.huckmag.com/perspectives/disabled-benefits-claimant-are-being-unfairly-investigated. Last access: 5 October 2022.

Dotson, Kristie. 2011. Tracking epistemic violence, tracking practices of silencing. *Hypatia*. 26:236–257. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1527-2001.2011.01177.x. Last access: 20 December 2022.

Du Gay, Paul. 2000. *In Praise of Bureaucracy: Weber, Organization, Ethics.* London: SAGE Publications Ltd.

Eubanks, Virginia. 2018. *Automating Inequality: How High-Tech Tools Profile, Police and Punish the Poor.* New York: St Martin's Press.

European Commission. 2021. *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts*. COM/2021/206 final. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206. Last access: 5 October 2022.

European Commission. 2022. Intelligent Portable Border Control System. *Community Research and Development Information Service (CORDIS)*. https://doi.org/10.3030/700626. Last access: 20 December 2022.

Felzmann, Heike, Eduard Fosch-Villaronga, Christoph Lutz, and Aurelia Tamò-Larrieux. 2020. Towards transparency by design for Artificial Intelligence. *Science and Engineering Ethics* 26:3333–3361. https://doi.org/10.1007/s11948-020-00276-4.

Fleetwood, Steve. 2008. Institutions and social structures. *Journal for the Theory of Social Behaviour* 38:241–265. https://doi.org/10.1111/j.1468-5914.2008.00370.x.

Freiman, Ori and Boaz Miller. 2020. Can Artificial Entities Assert?. In *The Oxford Handbook of Assertion*, Eds. Sanford Goldberg, 414–434. Oxford: Oxford University Press.

Fricker, Miranda. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford: Oxford University Press.

Gabriel, Iason. 2022. Toward a theory of justice for artificial intelligence. *Daedalus* 151(2):218–231. https://doi.org/10.1162/daed_a_01911.

Ganesh, Maya Indira and Emmanuel Mossl. 2022. Resistance and refusal to algorithmic harms: Varieties of 'knowledge projects'. *Media International Australia* 181(1):90–106. https://doi.org/10.1177/1329878X221076288.

Garnelo, Marta and Murray Shanahan. 2019. Reconciling deep learning with symbolic artificial intelligence: representing objects and relations. *Current Opinion in Behavioral Sciences* 19:17–23. https://doi.org/10.1016/j.cobeha.2018.12.010.

Graham, Peter J. 1997. What Is testimony? *The Philosophical Quarterly* 47**:** 227–232. https://doi.org/10.1111/1467-9213.00057.

Haas, Bridget M. and Amy Shuman. 2019. *Technologies of Suspicion and the Ethics of Obligation in Political Asylum*. Athens, Ohio: Ohio University Press.

Henman, Paul. 2010. *Governing Electronically: e-government and the Reconfiguration of Public Administration, Policy, and Power.* Basingstoke: Palgrave Macmillan.

Hildebrandt, Mireille. 2008. Defining Profiling: A New Type of Knowledge? In *Profiling the European Citizen: Cross-Disciplinary Perspectives,* Eds. Mireille Hildebrandt and Serge Gutwirth, 17–30. Dordrecht: Springer Netherlands.

Hill Collins, Patricia. 1998. *Fighting Words: Black Women and the Search for Justice.* Minneapolis and London, University of Minnesota Press.

Humphreys, Paul. 2004. *Extending Ourselves: Computational Science, Empiricism, and Scientific Method*. Oxford: Oxford University Press.

ITU. 2021. *United Nations Activities on Artificial Intelligence (AI)*. http://handle.itu.int/11.1002/pub/81b35851-en. Last access: 5 October 2022.

Joque, Justin. 2022. *Revolutionary Mathematics: Artificial Intelligence, Statistics and the Logic of Capitalism*. New York: Verso.

Jupe, Louise M. and David A. Keatley. 2019. Airport artificial intelligence can detect deception: or am I lying? *Security Journal* 33**:**622–635. https://doi.org/10.1057/s41284-019-00204-7.

Kaltheuner, Frederike. 2021. *Fake AI*. Manchester: Meatspace Press.

Koulish, Robert and Kate Evans. 2021. Punishing with impunity: The legacy of risk classification assessment in immigration detention. *Georgetown Immigration Law Journal* 36:1–72.

Lipsky, Michael. 2010. *Street-level Bureaucracy: Dilemmas of the Individual in Public Services.* New York: Russell Sage Foundation.

Maynard-Moody, Steven and Michael Musheno. 2022. *Cops, Teachers, Counselors: Stories from the Front Lines of Public Service* (2nd edition). Ann Arbor, Mich: University of Michigan Press. https://doi.org/10.3998/mpub.12247078.

McGoey, Linsey. 2012. Strategic unknowns: towards a sociology of ignorance. *Economy and Society* 41:1–16. https://doi.org/10.1080/03085147.2011.637330.

McQuillan, Dan. 2020. Deep Bureaucracy and Autonomist AI. In *Deserting from the Culture Wars,* Eds. Maria Hlavajova and Sven Lütticken. Cambridge, Mass.: MIT Press.

Medina, José. 2011. The relevance of credibility excess in a proportional view of epistemic injustice: Differential epistemic authority and the social imaginary. *Social Epistemology* 25:15–35. https://doi.org/10.1080/02691728.2010.534568.

Medina, José. 2013. *The Epistemology of Resistance: Gender and Racial Oppression, Epistemic Injustice, and Resistant Imaginations.* Oxford: Oxford University Press.

Mills, Charles W. 2017. *Black Rights/White Wrongs: The Critique of Racial Liberalism.* New York: Oxford University Press.

Narayanan, Devesh and David De Cremer,. 2022. "Google told me so!" On the bent testimony of search engine algorithms. *Philosophy & Technology* 35:22. https://doi.org/10.1007/s13347-022-00521-7.

Noble, Safiya U. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism.* New York: New York University Press.

O'Neil, Cathy. 2017. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* New York: Broadway Books.

Origgi, Gloria and Serena Ciranna. 2017. Epistemic Injustice: The Case of Digital Environments. In *The Routledge Handbook of Epistemic Injustice,* Eds. Ian J. Kidd, José Medina and Gaile Pohlhaus. New York: Routledge.

Pasquinelli, Matteo and Joler, Vladan. 2021. The Nooscope manifested: AI as instrument of knowledge extractivism. AI and Society 36:1263–1280. https://doi.org/10.1007/s00146-020-01097-6.

Pohlhaus, Gaile. 2014. Discerning the primary epistemic harm in cases of testimonial injustice. *Social Epistemology* 28:99–114.

Polanyi, Michael. 1962. *Personal Knowledge: Towards a Post-critical Philosophy.* London: Routledge & Kegan Paul.

Porter, Theodore. M. 1995. *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton, N.J.: Princeton University Press.

Ranchordás, S. 2022. Empathy in the digital administrative state. *Duke Law Journal* 71(6):1341–1389.

Rayner, Steve. 2012. Uncomfortable knowledge: the social construction of ignorance in science and environmental policy discourses. *Economy and Society* 41:107–125. https://doi.org/10.1080/03085147.2011.637335.

Rini, Regina. 2017. Fake news and partisan epistemology. *Kennedy Institute of Ethics Journal* 27(2): E-43–E-64. http://doi.org/10.1353/ken.2017.0025.

Rose, Nikolas. 1991. Governing by numbers: Figuring out democracy. *Accounting, Organizations and Society* 16:673–692. https://doi.org/10.1016/0361-3682(91)90019-B.

Rosenberger, Robert. 2014. Multistability and the agency of mundane artifacts: from speed bumps to subway benches. *Human Studies* 37:369–392. https://doi.org/10.1007/s10746-014-9317-1.

Samaržija, Hana and Ivan Cerovac. 2021. The institutional preconditions of epistemic justice. *Social Epistemology* 35:621–635. https://doi.org/10.1080/02691728.2021.1919238.

Sánchez-Monedero, Javier and Lina Dencik. 2020. The politics of deceptive borders: 'biomarkers of deceit' and the case of iBorderCtrl. *Information, Communication & Society* 25(3):413–430. https://doi.org/10.1080/1369118X.2020.1792530.

Sartori, Laura and Andreas Theodorou Andreas. 2022. A sociotechnical perspective for the future of AI: narratives, inequalities, and human control. *Ethics Information Technology* 24(4). https://doi.org/10.1007/s10676-022-09624-3.

Scheuerman, Morgan Klaus, Madeleine Pape and Alex Hanna. 2021. Auto-essentialization: Gender in automated facial analysis as extended colonial project. *Big Data & Society* 8(2). https://doi.org/10.1177/20539517211053712.

Scott, James C. 2020. *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. New Haven, Conn: Yale University Press.

Sedgwick, Eve Kofosky. 2008. *Epistemology of the Closet.* Berkeley: University of California Press.

Seo, Jungryul, Teemu H. Laine and Kyung-Ah Sohn. 2019. Machine learning approaches for boredom classification using EEG. *Journal of Ambient Intelli-*

gence and Humanized Computing 10:3831–3846. https://doi.org/10.1007/s12652-019-01196-3.

Sertler, Ezgi. 2018. The institution of gender-based asylum and epistemic injustice: A structural limit. *Feminist Philosophy Quarterly* 4(3). https://doi.org/10.5206/fpq/2018.3.5775.

Snow, Thea. 2021. From satisficing to artificing: The evolution of administrative decision-making in the age of the algorithm. *Data & Policy* 3, E3. http://doi.org/10.1017/dap.2020.25.

Spijkerboer, Thomas (Ed). 2013. *Fleeing Homophobia: Sexual Orientation, Gender Identity and Asylum*. Abingdon: Routledge.

Spivak, Gayatri C. 1987. *In Other Worlds: Essays in Cultural Politics.* New York: Methuen.

Stark, Luke and Jevan Hutson. 2022. Physiognomic artificial intelligence. *Fordham Intellectual Property, Media & Entertainment Law Journal* 32:922–978.

Symons, John and Ramón Alvarado. 2022. Epistemic injustice and data science technologies. *Synthese* 200, 87. https://doi.org/10.1007/s11229-022-03631-z.

Vallor, Shannon. 2021. The digital basanos: AI and the virtue of and violence of truth-telling. *2021 IEEE International Symposium on Technology and Society (ISTAS)*. https://www.doi.org/10.1109/ISTAS52410.2021.9629137.

Veale, Michael and Irina Brass. 2019. Administration by Algorithm? Public Management Meets Public Sector Machine Learning. In *Algorithmic Regulation*, 121–C6.P200, Eds. Karen Yeung and Martin Lodge. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198838494.003.0006.

Wachter, Sandra, Brent Mittelstadt and Chris Russell. 2021. Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review* 41:105567. https://doi.org/10.1016/j.clsr.2021.105567.

Wallis, Nick. 2021. *The Great Post Office Scandal: The Fight to Expose a Multimillion Pound Scandal Which Put Innocent People in Jail.* Bath: Bath Publishing.

Winner, Langdon. 1980. Do artifacts have politics? *Daedalus* 109:121–136.

Grade Prediction Is Not Grading

On the Limits of the e-rater

Jan Georg Schneider & Katharina A. Zweig

Abstract: *We use the example of the so-called e-rater to show how automated essay grading systems work and where we see their limitations, but also their potentials. From the perspective of a speech act theory that follows Austin and the late Wittgenstein, we show that the prediction of an essay evaluation as provided by the e-rater is subject to completely different felicity conditions from those of the evaluation itself. We find that the e-rater is not suited to capture cultural meaning. It analyzes cohesion without considering coherence, and for this reason it cannot be used to evaluate essays as a 'grader'. Nevertheless, we explore the question of whether it could be integrated into the evaluation process as a corrective under certain circumstances. Thus, using a specific example as an illustration, we show in detail the conditions under which machine predictions can have their role in social processes, even if the prediction of the outcome of a speech act is not equivalent to the performance of that speech act itself. On the basis of these findings, we reflect on the social nature of machine learning systems and their embeddedness in society and culture.*

1. Introduction

In this contribution we are concerned with the question of whether artificial intelligence can replace human decision-makers by predicting their decision on a new case. We address this question for a specific task: the grading of an essay. For this purpose, we analyze the so-called e-rater, a machine learning system for grading essays, which was developed in the United States and is already widely used in the educational system there. We take it as an example to show how automated essay grading systems work and where we see their limitations, but also their potentials. In the following three sections, we describe

how the e-rater functions and what issues it raises. In the last section, we further develop our central speech act argument by showing how predicting an essay grade is fundamentally different from the grading itself. Finally, we address the question of whether, despite its limitations, the e-rater can be used in some way to support essay grading.

2. Training an essay scoring system

The patent of the e-rater, entitled "System and Method for Computer-based Automatic Essay Scoring",[1] was approved in 2002. The description essentially also applies to the current version of the e-rater, in which nothing has changed fundamentally (cf. Perelman 2020). The system serves the purpose of replacing gradings by human reviewers (patent: 1). Primarily, the e-rater is used in so-called language proficiency courses, such as the TOEFL test. Here is an example of a typical task that could appear in such tests:

> "Everywhere, it seems, there are clear and positive signs that people are becoming more respectful of one another's differences." In your opinion, how accurate is the view expressed above? Use reasons and/or examples from your own experience, observations, or reading to develop your position. (patent: 10)

The task is then to write an essay of about 400–500 words, which is very often part of the final exam of language proficiency courses. They aim to find out whether the student's English skills are good enough to study at university, for instance.

The underlying scoring system by humans is very elaborate and uses a matrix of so-called rubrics, which assign grades according to quality criteria as described in the patent:

> For example, the scoring guide for a scoring range from 0 to 6 specifically states that a "6" essay "develops ideas cogently, organizes them logically, and connects them with clear transitions". A human grader simply tries to evaluate the essay based on descriptions in the scoring rubric. This technique, however, is subjective and can lead to inconsistent results. (patent: 1)

1 Cf. Burstein et al. 2002; hereafter cited as "patent" with the indication of the column number.

Thus, according to its inventors, the attractiveness of the e-rater lies on the one hand in its greater accuracy and objectivity, and on the other hand in the cost-efficient replacement of human reviewers. However, previous human evaluations form the basis for the calculations of the e-rater. The general approach of the system is to check which features are common in essays that have been positively evaluated by humans before. So, how does the e-rater do this?

Technically speaking, the e-rater is a combination of curated rules, so-called *expert systems*, and a learnt component to predict essay grades. In the patent, the approach of learning how to predict a grade is described for two kinds of essays. Here, we focus on so-called *argument essays*, in which the student is provided with an argument and asked to analyze it. The grading process for a given test question to be answered in the argument essay is based on an electronic version of the essay. This electronic version is read by a standard language parser assigning word categories to each word and also identifying larger syntactic structures such as *infinitive clauses* or *relative clauses*. Another expert system tries to identify the beginning and end of individual arguments by searching for a list of keywords such as *otherwise*, *conversely* or *notwithstanding* (patent: column 11). This heuristic thus annotates the text and splits it into individual arguments based on particular words and phrases. The partition of the text and the text as a whole then provide the basis for further calculations, which are all of a very simple nature, e.g., counting the total number of infinitive clauses.

Eventually, each essay results in four sets of numbers.

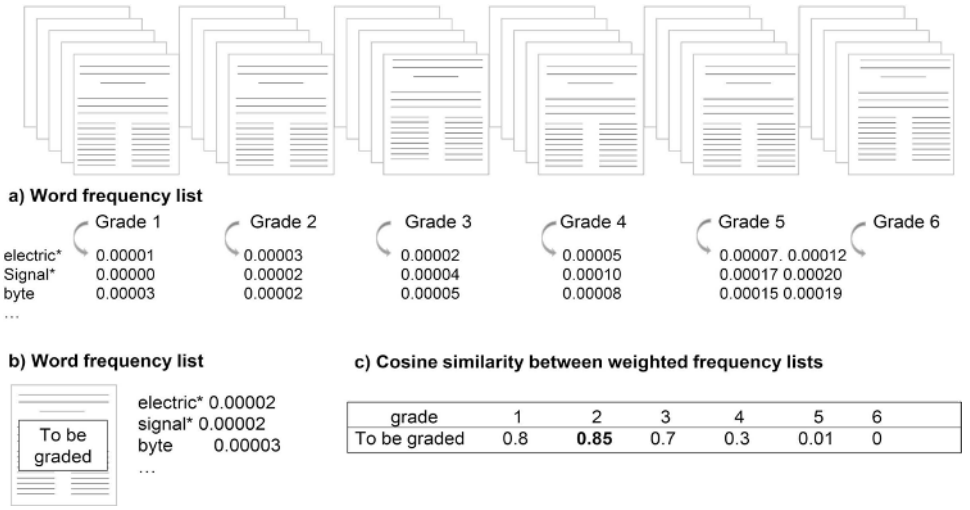
- The first set of results contains only two values: The total number of *modal auxiliary verbs* (such as *can*, *must*, *will*) and the relative portion of *complement clauses* per sentence.
- The second set of results is supposed to score the *rhetorical structure* of the essay; this is based on the output of the argument-identification component described above. It contains four values obtained from a count of specific markers, e.g., the total number of occurrences of subjunctive forms of modal verbs (*would*, *could*, *should*) in the final paragraph of the essay.
- The third set of results contains a weighted value for words depending on their salience (see Fig. 1). The salience of a word is calculated by its relative frequency in any essay and by the inverse frequency in the set of all essays. For example, any word directly related to the test question will be very frequent in an individual essay, but at the same time its salience is very low because it occurs in all of the essays. Technically, the third set of results

is based on the product of the token frequency of the particular word in one document and its inverse token frequency across all documents.[2] This gives a long list of numbers for each individual essay. The same is done for the concatenation of all essays graded with a 1, for the concatenation of all essays graded with a 2, and so on. In the last step, the weighted word list is compared to the weighted word list for essays from each of the six different grades by a measure called *cosine similarity*. The result of this third analysis is the grade of the essay collection that is most similar to the essay under scrutiny with respect to this similarity measure. That is, if in class 6, for example, some rare words are used by many essays and the new essay does so as well, while other, more frequent, words are not used so much in essays of this class and in the new essay, the essay is more likely to be assigned to class 6 than to some other class. Figure 1 illustrates the issue just described, i.e., the e-rater's grading approach based on word frequency lists, which leads to the third result mentioned. a) The essays graded by humans are sorted by grade. Over all essays with the same grade, the frequencies of the stemmed words are counted. Thus words like *electrical* and *electricity* are counted as *electric**. For each stem of a word, all occurrences in the essays of the same grade are counted as a frequency. b) The same is done for the essay to be graded by the system. All words on all word lists are then weighted by the inverse frequency of the (stemmed) words in all texts. c) Finally, the resulting sequences of numbers from all essays with a grade 1, those of all essays with a grade 2, and so on are compared to the sequence of numbers of the essay to be graded in terms of cosine similarity. The highest value determines the outcome of the third score in the grading approach of the e-rater.

- For the fourth output, the procedure described for the third is similarly performed for the words in each 'argument' as identified by the expert system with its keywords created by humans: That is, a grade is now assigned to each 'argument' by comparing word usage with arguments from essays in the different categories. All these scores for the individual 'arguments' are then averaged in a way not described in detail in the patent to form the final score.

2 Stop-words are removed and the words are stemmed.

Figure 1: e-rater's grading approach based on word-frequency lists (3rd number computed in the approach)



The entire procedure of creating these four sets of scores is applied to 250–300 essays already graded by humans (see also Burstein et al. 2013a: 61) – thus the eight resulting numbers and the grade given to each of these essays are known. These numbers are then the input to a *linear regression*, a simple machine learning method. The method finds the weights for each of the inputs so that the predicted grade from a linear equation based on the inputs is not too far from the actual grade, averaged over all 250–300 graded essays. The resulting formula is then used for all essays to be re-scored. Thus in summary, the e-rater does the following:

- Based on a set of 250–300 essays on one test question which were graded by humans, it learns which of the very simple syntactical features are most associated with a good or bad grade. These features can be easily identified, e.g., the number of tokens of modal auxiliary verbs.
- It also learns which words are most and least often used in essays on the given question, both in the overall text and in the individual arguments identified by an expert system, based on a list of keywords created by human experts.

It can thus be stated that the e-rater is not designed and hence not able to identify coherence, logical arguments and the reasonability of their interconnec-

tion. It can only count individual words[3] and very simple syntactical structures, and it has "learned" which number of tokens of these structures is often associated with a high or low grade.

3. Criticism of the e-rater

The patent states that the e-rater-system "automatically rates essays using features that reflect the 6-point holistic rubrics used by human raters" (patent: 3). For example, an essay would be worth the highest grade 6 according to such a rubric matrix if it "develops ideas cogently, organizes them logically, and connects them with clear transitions", as already quoted above (patent: 1). The patent-holders base their quantification of these features on the identification of specific words that may be used in a discourse to structure arguments. However, these word lists are not complete, nor is the identification of mere words a substitute for the semantic and pragmatic analysis of the argumentative structure. Thus the heuristics used may or may not identify all subtexts containing an argument. In any case, the logic of their organization or the use of clear transitions cannot be grasped by machines. It is not measured whether the ideas are coherently developed – even the inventors of the e-rater concede this (Burstein et al. 2013b). Nor is the organization of the arguments assessed in terms of content.

The second vector, which is said to represent "rhetorical structure", is just a count of simple syntactic properties of the text, e.g., the "total occurrences of argument development using belief words", based on a heuristic for identifying keywords that statistically indicate arguments, i.e., that are taken as "indices" or "symptoms" of arguments (cf. Keller 2018: 155–168, with reference to Peirce; see also section 5 below). Again, it is not at all the students' rhetorical skills that are determined since the accomplished quantification is not based on a semantic analysis of the text. Perelman (2020) puts it like this: "Testing companies freely use the term artificial intelligence, but most of the systems

3 In the *Handbook of Automated Essay Evaluation*, Burstein et al. (2013a: 60) state that "topic-specific vocabulary" would also be identified in advance – by humans – as characteristic of better-rated essays; thus it is not a purely quantitative approach. This does indeed bring into play a human categorization ("topic-specific vocabulary") which as such cannot be quantitative but rather constitutes a semantic property; however, this does not change the fact that the meaningfulness and correctness of the use of these expressions cannot be checked by the e-rater.

appear to produce a holistic score largely through summing weighted proxies." To prove this, Perelman built a text generator called BABEL which generates grammatically correct but otherwise non-sensical texts. Here is an example of a text generated by such a system (quoted after Perelman 2020):

> Theatre on proclamations will always be an experience of human life. Humankind will always encompass money; some for probes and others with the taunt. Money which seethes lies in the realm of philosophy along with the study of semiotics. Instead of yielding, theatre constitutes both a generous atelier and a scrofulous contradiction. As I have learned in my reality class, theatre is the most fundamental analysis of human life. Gravity catalyzes brains to transmit pendulums to remuneration. Although the same gamma ray may receive two different pendulums at the study of semiotics, a plasma processes interference. Simulation is not the only thing an orbital implodes; it also inverts on theater. […]

According to Perelman, this text and similar ones earned the highest grade even with the newest e-rater system. Perelman's results are often used to show that students might be able to "learn to the test", e.g., by adding rare words to their essay whether they fit or not, or by learning lists of special "cue words" by heart; they can also be taught to use specific syntactic structures independent of their semantic quality and fit. It is obvious that memorizing and using proxies does not mean knowing how to write meaningful essays, while it elicits the highest grades from the Automated Easy Scoring (AES) since the AES cannot capture whether a text makes sense or not. Thus a self-fulfilling prophecy can set in: It is possible that the predictions become more and more precise, while the texts become more and more meaningless, since a meaningful coherence is no longer an evaluation criterion. If the e-rater were actually used as a *substitute* for essay evaluations by humans, as the authors of the patent envision, then it would basically be honest and transparent to test, very explicitly, only the knowledge of the expected proxies and thus also make explicit the underlying 'teaching to the test'.[4]

However, since the required competence of each student consists or should consist precisely in writing texts that can be understood by other humans in terms of content, and since the e-rater is not suitable for handling situations in which human feedback is needed, the use of such a system *alone* is already ruled

4 However, it should not be forgotten here that successful candidates do also consider what *human* raters might 'want to hear' (in ideological terms, for example).

out for didactic reasons. For the following discussion, we will thus assume that students need to expect at least control samples by human graders so that they cannot afford to write non-sensical texts.

4. Quality of the prediction

Can the e-rater and similar systems be used as the main grading system if human graders control some sample of the essays? To answer this question, the accuracy of the grade prediction must be considered first. A study in Germany and Switzerland showed that for two different types of tasks the e-rater, when specifically trained, achieves between 13% and 42% absolute agreement, i.e., it hits exactly the same grade as human graders in this percentage range on a 0–5 point scale (Rupp et al. 2019). Furthermore, if a deviation of the grade by no more than one point on the 0–5 point scale is considered, the accuracy of the system increases to up to 99% (between 73.8% and 99.4%; Rupp et al. 2019: Table 5). In the light of these results the authors estimated that the e-rater's predictions were within an acceptable range – although the agreement between human raters was significantly higher than that between the system and human raters.

Other studies on the e-rater show similar accuracy values or even slightly better ones (cf. Meyer et al. 2020: 4), so the acceptance level for its use has increased in recent years. In the following, we will argue that the results provided by a system such as the e-rater, even if a 100% agreement can be achieved, are not *qualitatively* the same as an essay assessment, but something categorically and qualitatively (cf. Becker 2021: 9–30) different. It is not even an essay grade, but exclusively a prediction of such a grade. This is our core argument, which we develop in the next section using the perspective of speech act theory. What are the characteristics of the act of evaluating an essay? Here, we are not interested in denying the power of the e-rater as a predictive tool; rather, we want to explore the limits of its applicability. Under what conditions can such a system be used in a supportive manner, and what requirements of the social process can it *not* meet? In order to arrive at an assessment that is as robust and fair as possible, we assume a best-case scenario for the use of an e-rater system as follows:

- The system is trained for each test question separately, based on 250–300 essays rated by human raters.

- The students know or at least assume that their essay may be graded by a human rater – thus they have to write an intelligible essay (no 'learning to the test').

5. Grade prediction is not grading

Summarizing what we have presented in the last two sections, we can say that the e-rater is not able to apply quality criteria but only to count 'symptoms'[5] such as the frequency of words and particular syntactic structures. Thus it cannot *evaluate* essays, which are culturally anchored semiotic phenomena. As such, the comprehension of essays and other texts is culturally embedded and dependent on cultural knowledge the e-rater does not have – for fundamental reasons. Cultural knowledge, which significantly includes the understanding of cultural artifacts, is "a complex constellation of acquired abilities" (Goodman/Elgin 1988: 114) that cannot be acquired by machines trained in the way outlined above.

In order to make this idea clearer, let us undertake a small detour via the automated *translation* program DeepL. Since the culture-dependency of essays and other texts seems almost self-evident in some respects, the recent progress of DeepL is astonishing. How is it able to produce translations that can – to some extent and within limits – impress even experienced human translators? The software employs Convolutional Neural Networks (CNN), a method commonly used in image recognition. The advantage is that – unlike so-called recurrent neural networks – all words are translated (cf. Merkert 2017). In the case of DeepL, the CNNs were trained with the associated Linguee system. With this database, DeepL was able to collect extremely extensive, very high-quality training data (Schmalz 2018: 200). The company bases its success in part on the fact that it has access to "billions of high-quality translations" from its corporate history (Schmalz 2018: 203).

The fundamental qualitative difference between predicting a grade of an essay and translating a text with an automatic translator is twofold. First, DeepL produces something that is categorically the same as a human translation: a linguistic product that can be evaluated according to the same criteria

5 In semiotics, a symptom is an outward indicator (*Anzeichen*) of something. For example, red spots on the face can be a symptom of measles (cf. Keller 2018: 161). In informatics, the analogous term 'proxy variables' is used.

as a human translation. Second, the cultural embedding or the cultural circumstances of the words in the text are potentially represented, although there may of course be errors in these representations of cultural meaning offered by DeepL, so the ultimate decision and responsibility must rest with the human translator. However, the large number of human translations that serve as reference texts ensure inclusion of the relevant criteria which effectively orient people when translating. This way, for instance, human intentionality and taste as well as the 'zeitgeist' together with its fluid, group-specific differentiating conventions can, and often do, find expression in the translations produced by DeepL.

The e-rater, by contrast, makes a rating prediction, more precisely a grade prediction,[6] based on indices or symptoms alone. And from the perspective of speech act theory, the prediction of a grade is, as we will show below, something categorically different from an evaluation itself.

5.1 The basic idea of speech act theory

Speech act theory was developed by John L. Austin in his 1955 Harvard lectures and published posthumously under the title *How to Do Things with Words* (Austin 1975 [1962]). In everyday situations, we often think of speaking and acting as opposites. But for Austin, speaking in many cases means doing something, namely, performing speech acts. When I say, "I christen this ship the Queen Elizabeth", I am performing the speech act of ship christening, provided the circumstances fit and I am authorized to do so. When children re-enact such a christening in play, it will thus not have the same effect. Only if a set of "felicity conditions" is met can the speech act be successful. Austin calls utterances of this kind *performatives* or *illocutionary acts*.[7] A judge's verdict is also such a

6 Of course, one could object that DeepL also technically generates a translation *prediction*. However, the de facto difference is that here a product is created which can be evaluated in the same way as a human translation and into which the cultural context has implicitly entered, as already mentioned. The e-rater, on the other hand, does not base its grade prediction on criteria but only on symptoms, on the basis of which the quality of the evaluation and grading cannot be assessed, but only the precision of the prediction in comparison with a human evaluation.

7 On the surface, Austin's argument could be understood as abandoning the distinction between performatives and constatives, replacing it with the distinction between locutionary, illocutionary and perlocutionary acts, in which the notion of illocutionary act captures the performative aspect of speech acts. However, if one includes the

performative utterance. If the court is legitimate and the procedure is carried out correctly and completely, then the verdict applies with all its consequences. Furthermore, the question arises of whether the verdict was appropriate and fair. All these aspects are considered by Austin.

Each speech act has its specific set of felicity conditions, and the individual conditions often differ from those of other speech acts. Such differences may include, for instance, whether a speech act requires a justification and why a justification is necessary. With this in mind, let us now present and discuss the felicity conditions of evaluating and grading.

5.2 Felicity conditions of evaluation and grading of essays

Our hypothesis is that an essay evaluation must include a justification, preferably even an explicit one. But how can we justify why it must contain a justification? Why is a prediction not sufficient here? Why do we also need a qualitative evaluation?

As philosophers of language have made very clear, cultural meaning is constantly renegotiated by the collective, that is, by a community of sign users (cf. Wittgenstein[8] 1984; Goodman/Elgin 1988). But how can we involve this collective in the grading process? Essay evaluation can only be legitimized by knowing how sign usages are entrenched and established within the collective. The existence of this knowledge, then, is one of the felicity conditions in essay evaluation, for only on this basis can a valid justification be given for an evaluation. And only a justification provides the possibility of checking whether, e.g., an evaluation is intersubjectively legitimated or – at least – legitimizable and not arbitrary.

With regard to felicity conditions, the questions that arise are a) exactly which procedures are to be chosen here, and b) which persons fit these procedures and therefore should be authorized for an evaluation. In scientific and educational contexts, we employ *experts* for this purpose: i.e., people we authorize as reviewers because we believe they have been part of the collective long

---

fact that Austin kept revisiting the distinction between performatives and constatives, even though he had long since 'deconstructed' the dichotomy between the two, then there are good reasons to suppose that the notion of the performative remained important to him. Even if all speech acts are ultimately performative, there are those in which the performative character is more prominent than in others.

8 On Wittgenstein's pragmatic conception with regard to linguistics cf. Schneider 2008.

enough to be able to speak for it, or more precisely, to be able to provide good candidates for justifications that are then in turn intersubjectively verifiable (e.g., by other experts).

In the following, we go through Austin's (1975 [1962]: 14–18 and 25–46) six felicity conditions in detail and apply them systematically to essay grading:

- *A 1*: A speech act can only be accomplished at all if there is a corresponding *conventional procedure* which involves certain persons uttering certain words under certain circumstances. In the case of essay grading, this is a procedure that requires a close reading of the submitted essay and an assignment of the essay to a score level according to specific criteria within the framework of the underlying grading scheme, e.g., those given by a rubric. The central linguistic act, usually a writing act in the case of essay evaluation, has certain similarities with a judge's verdict and can be put into an explicitly performative form (cf. Austin 1975 [1962]: 69) of the following kind: 'I hereby evaluate the present essay with the grade x.' As mentioned, essay evaluation also includes justification of the grade by the reviewer or at least the assumption that it can be justified by the reviewer upon request.
- *A 2*: The respective persons, objects and circumstances must fit the speech act to be performed. In our case, the persons authorized and qualified to perform the evaluation are the experts employed by the collective, e.g., teachers or professors.
- *B 1 and B 2*: All persons involved must carry out the procedure correctly and completely. In our case, this means that on the basis of the respective essay and with the help of transparent criteria (cf. rubric descriptions), an assignment to one of the intended grading categories must be made unambiguously. The correctness and completeness of the procedure also includes, for example, checking as well as possible whether the submitted essay is valid, i.e., that it is, for instance, not plagiarized.
- If one or more of the conditions A 1 to B 2 are not met, then the speech act of essay grading does not proceed. It can also happen that such an evaluation is null and void in retrospect, e.g., because the essay only later turns out to be plagiarized.
- But even if the evaluation has come about and is thus valid, it can still fail in two other ways. Since these conditions are of a categorically different kind, Austin does not continue his list with the third letter of the Latin alphabet, but with a Greek gamma.

- *Γ 1*: The speech act must not be untrustworthy or insincere. With respect to grading in general, this requirement can be deduced from the perspective that a grade must also be a "signal" (see Spence 1973) to the author (i.e., the student) and also to potential future employers. The grade signals how the student's performance was assessed by an expert with regard to the qualification aimed at and possibly also with regard to the student's career opportunities. If the expertise is not carried out in good faith, then this condition is not met. We see that it is precisely here that the ethical and moral dimension of grading is located.
- *Γ 2*: Afterwards, all participants must behave in a way that fits the respective completed speech act. For example, if one has given a very good grade, it is not appropriate to reprimand the student afterwards. It would be equally inappropriate for an expert to cast doubt on the student's evaluation afterwards. Here we can see, by the way, how closely Γ 1 and Γ 2 can be related.

If all six felicity conditions with respect to the grading are fulfilled, then the probability is very high that the evaluation was successful as a speech act.[9] Beyond that, however, the question can of course still arise as to whether it was fair, appropriate, etc. If a grade is being challenged on substantive rather than procedural grounds, it may be appropriate to bring in additional reviewers. How many reviewers are required depends, generally speaking, very much on the type of evaluation procedure. In the case of a post-doctoral habilitation in the German university system, for example, three to five reviewers are involved, while a simple exam in school is usually graded by just one teacher.

In contrast to the procedure of human essay evaluation and grading described above, the symptoms that the e-rater identifies and then uses for its predictions can never be used as *reasons* for gradings. However, the justification is the most important factor in the procedure of essay evaluation, which consists of both grading and justification. The justification serves to stabilize the procedure for the future, and only in this way can the felicity conditions be maintained here.

Thus the procedure must remain anchored in the corresponding cultural practice. It is essential that the criteria which make up a good essay are explicitly taken into account in the evaluation procedure and that their aggregation

9 However, this is not absolutely certain, because owing to the fundamental unpredictability of human communication Austin gives only necessary, but not sufficient felicity conditions.

leads to the specification of a grade, because only in this way can a comprehensible and adequate justification be given. These criteria include, in particular, coherence, argumentative plausibility, truthfulness, originality and aesthetic value. As shown by the software BABEL, the generator of non-sensical texts, the e-rater cannot capture any of these aspects but can only consider surface phenomena of vocabulary and cohesion. It analyzes cohesion without coherence, symptoms but not criteria. Machine learning cannot distinguish between rational and senseless inferences (Goodman/Elgin 1988: 108f.; Anson/Perelman 2017: 279); AES systems cannot 'read between the lines', they do not capture allusions and irony, and they are not able to assess complex novel metaphors and humor (Balfour 2013: 42).

5.3 Can predictions nevertheless play a meaningful role in the evaluation process?

As we have pointed out, grading is a completely different speech act from grade predicting, and grading can only be done by educated, selected members of the respective language collectives – which restricts the performance of such an act to humans for the time being. Strictly speaking, the e-rater cannot perform the speech act of prediction either – because "machines are not actors" (cf. Becker 2021: 19, our translation) – but at least it can substitute such a prediction under certain circumstances (cf. Janich 2015: 314; Becker 2021: 182). This raises the question of whether such automated grade predictions can nevertheless be helpful in the process of grading and evaluating an essay. A prediction is successful if it is as statistically accurate as possible. This statistical accuracy must be established in many instances in order to say that the prediction could be viable at all. Obviously, this is the case with the e-rater, so its predictions, if used correctly, may have some value for the process after all.

With all due caution, the e-rater could perhaps help to filter out the 'outliers' heuristically in a large set of essays to be evaluated. Those essays where the human evaluation deviates significantly from that of the e-rater might then deserve special attention. They could – and we do not think this is unlikely – actually be particularly good, although they do not exhibit the statistical symptoms. Conversely, however, the presence of the statistical symptoms could also indicate that a human rater may have rated an essay too negatively or too positively. An automated comparison of the essays with regard to the symptoms mentioned cannot determine whether the respective 'outlier' was caused by the specificity of the essay or by the specificity of the expert opinion. This, in turn,

can only be verified by the judgement of a human employed by the collective, because the e-rater's predictions are based on the assumption of a 'normality' of the essays to be graded.

So even if the e-rater could be used carefully as a convenient tool, this would by no means – and this is the crucial point – authorize it to perform the speech act of grading with all its consequences. As we have argued above, only humans who are part of the respective collective can actually evaluate a text and determine whether it is written in accordance with the culture of that collective. But not *all* humans are allowed to speak for the entirety of this collective: only evaluators chosen on the basis of their qualifications or other prior achievements are capable of doing so. Thus a grade as the result of an evaluation can only be given by a human. Here again, the comparison with a translation by a system like DeepL is quite interesting. It reveals differences as well as similarities: In the case of the translation, it is immaterial whether the rough version came from a human being or a machine, since the cultural context tends to be translated in the process. In the end, however, it is again crucial that for the released version responsibility is taken by a competent human translator who can check whether the nuances have been correctly captured, etc. Again, this has to do with the assumption of the duty to justification and is part of the language game, for example, in book translations by publishers.

Certainly, with the e-rater, a comparison can always be made between the real evaluation by a human and the automated prediction. This prediction always remains dependent on such comparisons, since it is not, after all, based on the quality criteria that humans use to evaluate an essay meaningfully. Without human raters, there is as yet no way to provide a justification for the particular grade, and as we have argued, an essay grade must always be intersubjectively legitimated.

In the course of this analysis, we have identified some methodological preconditions that might allow for a useful application of AES. The first two were already noted at the end of section 4, and we add a third point (c) here:

- Individual training based on tests graded by humans: The system would need to be trained separately for each test question, based on 250–300 essays scored by human reviewers, in line with Austin's felicity conditions.
- Safeguards against learning to the test: It has to be made explicit to the students that they can assume their essay is to be graded by humans, i.e., they need to know that writing a meaningful essay is what matters.

- Legitimacy of the speech act: The prediction by the e-rater is used only as a measure of deviation, a guide to attention, and a possible corrective to supplement human evaluations; it is intended to support human evaluators, but in no way can or should it completely replace them without harming the legitimacy of grading as a speech act.

Generally, it must always be considered and appropriately taken into account that this type of software has massive economic incentives, in terms of saving time and human resources. It is therefore necessary to reflect ethically and politically on the implications this may have (cf. Zweig 2018; Zweig et al. 2021). The more the evaluators are under pressure of time and economic considerations, the more the 'sincerity condition' (Γ 1) is challenged and possibly compromised: honest 'signals' can only be given in the long run if both sides have to prove and can verify honesty. To this end, it must be ensured on the one hand that the reviewer provides grade justification when asked to do so, and, on the other hand, to the greatest possible extent the achievement of good grades with plagiarized or non-sensical texts must be prevented.

6. Conclusion

For the future, the question remains whether the e-rater could become as convincing a tool as, say, DeepL if it learned the grading process itself by being supplied *en masse* with human essay evaluations in text form. This would require, however, linking these evaluations to the respective essays during the e-rater's machine learning process. Then, perhaps, the e-rater could generate an assessment in text form that would be linked and matched to a specific new essay, and that might then even substitute, as in the case of DeepL, the human action in a near-equivalent way (cf. Becker 2021: 182, following Janich 2015[10]). But is such a process of essay evaluation technically even possible, given that

10 In the German original, the relevant term is "leistungsgleiche Substitution", which is hard to translate into English. It means that the substitute is not the same but something with the same output as the substituted action or process: computers, for example, do not calculate, because calculating is an intentional action. But they can have the same output as human calculating. In this sense they substitute human action in a near-equivalent way (cf. Becker 2021: 19).

the content relationship between essay and evaluation text is much more complex and 'loose' than that between a text and its translation? And even if it were possible, we could not do without human reviewers, because only they could further explain the justification if this is required, and take responsibility for evaluating the essays. However, as argued above, even with automated translation, it is necessary that justification by a human can be provided upon request.

The application of speech act theory in our analysis has shown why the prediction of a speech act result does not match the correct performance of the speech act in all its facets. The prediction cannot substitute the speech act of grading in a near-equivalent way. On the one hand, our comparison of grading and grade prediction shows that we are dealing with two categorically different processes; on the other hand, however, the findings also leave room for the assumption that even AI systems that are actually inappropriate can sometimes represent an interesting corrective in a social process. In the intelligent engagement with so-called artificial intelligence, we have the opportunity to reflect on our cultural practices and examine them in terms of their viability: In which cases does a decision need an explicit substantive justification? What are historically evolved practices worth to us in particular? Which ones do we want to retain for ethical or cultural reasons and which can be modified with the help of AI? What happens in communicative processes when machines are involved? Machines rarely perform in the same way as humans; in some cases they perform something functionally equivalent (e.g., a calculator), in many other cases something categorically different (e.g., the e-rater), and in still other cases something somewhere in between (e.g., DeepL). It becomes particularly interesting when a machine learning system uses large amounts of data to uncover quantitative aspects that humans do not see because they interpret qualitatively and in this sense proceed *intelligently*. When humans 'interact' with such technologies, i.e., deal with them intelligently and use them as a corrective, they can optimize their results and decisions, whether in translation, essay evaluation or other cultural practices. In this respect, as Elena Esposito (2017) points out, it is indeed more appropriate to speak not of artificial intelligence but rather of "artificial communication". When people learn with the help of machines by 'interacting' with them, machine learning can potentially help to improve social decisions.

We have proposed an approach that could also be suitable for the assessment of similar AI systems which are intended to complement human evaluations or decisions in social processes. We therefore see potential for generalization in this approach, which we will explore in future research.

Bibliography

Anson, Chris M. and Les Perelman. 2017. Machines Can Evaluate Writing Well. In *Bad Ideas About Writing*, Eds. Cherryl E. Ball and Drew M. Loewe, 278–286. Morgantown, W. Va.: West Virginia University Libraries.

Austin, John L. 1975 [1962]. *How To Do Things with Words*. 2nd edition. Oxford: Oxford University Press.

Balfour, Stephen P. 2013. Assessing writing in MOOCs: Automated Essay Scoring and Calibrated Peer Review™. *Research & Practice in Assessment* 8:40–48.

Becker, Ralf. 2021. *Qualitätsunterschiede. Kulturphänomenologie als kritische Theorie*. Hamburg: Meiner.

Burstein, Jill C., Joel Tetreault and Nitrin Madnani. 2013a. The e-rater Automated Essay Scoring System. In *Handbook of Automated Essay Evaluation. Current Applications and New Directions*, Eds. Mark D. Shermis and Jill Burstein, 55–67. London: Routledge.

Burstein, Jill C., Lisa Braden-Harder, Martin S. Chodorow, Bruce A. Kaplan, Karen Kukich, Chi Lu, Donald A. Rock and Susanne Wolff. 2002. *US 6,366,759 B1* [United States Patent, April 2, 2002: System and method for computer-based automatic essay scoring].

Burstein, Jill, Joel Tetreault, Martin Chodorow, Daniel Blanchard and Slava Andreyev. 2013b. Automated Evaluation of Discourse Coherence Quality in Essay Writing. In *Handbook of Automated Essay Evaluation. Current Applications and New Directions*, Eds. Mark D. Shermis and Jill Burstein, 267–280. London: Routledge.

Esposito, Elena. 2017. Artificial communication? The production of contingency by algorithms. *Zeitschrift für Soziologie* 46(4):249–265.

Goodman, Nelson and Catherine Z. Elgin. 1988. Confronting Novelty. In *Reconceptions in Philosophy & Other Arts & Sciences*, Eds. Nelson Goodman and Catherine Z. Elgin, 101–120. Indianapolis, Ind./Cambridge, Mass.: Hackett Publishing Company.

https://www.heise.de/newsticker/meldung/Maschinelle-Uebersetzer-DeepL-macht-Google-Translate-Konkurrenz-3813882.html. Last access: 3 March 2022.

Janich, Peter. 2015. *Handwerk und Mundwerk. Über das Herstellen von Wissen*. Munich: Beck.

Keller, Rudi. 2018. *Zeichentheorie. Eine pragmatische Theorie semiotischen Wissens*. 2nd edition. Tübingen: Narr.

Merkert, Pina. 2017. *Maschinelle Übersetzer: DeepL macht Google Translate Konkurrenz.*

Meyer, Jennifer, Thorben Jansen, Johanna Fleckenstein, Stefan Keller and Olaf Köller. 2020. Machine Learning im Bildungskontext: Evidenz für die Genauigkeit der automatisierten Beurteilung von Essays im Fach Englisch. *Zeitschrift für Pädagogische Psychologie* 2020:1–12, https://doi.org/10.1024/1010-0652/a000296.

Perelman, Les. 2020. The BABEL Generator and e-rater: 21st century writing constructs and Automated Essay Scoring (AES). *Journal of Writing Assessment* 13(1). http://journalofwritingassessment.org/article.php?article=145. Last access: 3 March 2022.

Rupp, André A., Jodi M. Casabianca, Maleika Krüger, Stefan Keller and Olaf Köller. 2019. Automated Essay Scoring at Scale: A Case Study in Switzerland and Germany. *TOEFL Research Report Series and ETS Research Report Series 1/2019:1–23.*

Schmalz, Antonia. 2018. Maschinelle Übersetzung. In Volker Wittpahl (Ed.). *Künstliche Intelligenz. Technologie, Anwendung, Gesellschaft*, 194–208. Berlin and Heidelberg: Springer.

Schneider, Jan Georg. 2008. *Spielräume der Medialität. Linguistische Gegenstandskonstitution aus medientheoretischer und pragmatischer Perspektive*. Berlin and New York: de Gruyter.

Spence, Andrew Michael. 1973. Job market signaling. *Quarterly Journal of Economics* 87(3):355–374.

Wittgenstein, Ludwig. 1984. Philosophische Untersuchungen. In Ludwig Wittgenstein. *Werkausgabe in 8 Bänden. Vol. 1: Tractatus logico-philosophicus*, 225–580. Frankfurt a. M.: Suhrkamp.

Zweig, Katharina A., Tobias D. Krafft, Anita Klingel and Enno Park. 2021. *Sozioinformatik: Ein neuer Blick auf Informatik und Gesellschaft*. Munich: Hanser.

Zweig, Katharina. 2019. *Ein Algorithmus hat kein Taktgefühl. Wo künstliche Intelligenz sich irrt, warum uns das betrifft und was wir dagegen tun können*. 9th edition. Munich: Heyne.

Intelligente Ökologien

Von allopoietischen zu autopoietischen algorithmischen Systemen

Überlegungen zur Eigenlogik der operationalen Schließung algorithmischer Systeme

Jan Tobias Fuhrmann

Abstract: *Der Beitrag geht davon aus, dass der Erfolg aktueller Künstlicher Intelligenz darauf beruht, dass sie zu einem konstitutiven Außen der Kommunikation avanciert ist. Dies ist gelungen, indem von allopoietischen auf autopoietische algorithmische Systeme umgestellt wurde. Kennzeichnend für diese Umstellung ist, dass die Regeln einer sinnhaften Verarbeitung von Kommunikation nicht mehr wie bei der symbolverarbeitenden KI über Algorithmen implementiert, sondern durch eine statistische Auswertung von Daten emuliert werden. Dabei schließen sich die algorithmischen Systeme autopoietisch, indem sie neben der Reproduktion der Permutation elektronischer Schaltzustände die Bedingungen der Permutation durch ein historisches Unruhepotenzial von Daten variieren. Das zeigt sich beispielsweise daran, wie Schaltungen in künstlichen neuronalen Netzen gewichtet werden. Um das zu zeigen, wird in einem ersten Schritt plausibilisiert, dass autopoietische algorithmische Systeme nicht in die Autopoiesis sozialer Systeme eingebaut werden, sondern als deren konstitutives Außen äquivalent zu psychischen Systemen zu behandeln sind. Im darauffolgenden Schritt werden die Bedingungen geklärt, die es gestatten, von einer Autopoiesis algorithmischer Systeme zu sprechen. In einem dritten Schritt soll sodann danach gefragt werden, wie Kommunikation in die Logik der Autopoiesis algorithmischer Systeme übersetzt wird. Abschließend wird das Konzept der Duplexstruktur von Kommunikation eingeführt, das es ermöglichen soll, Kommunikation sowohl als sinngenetischen Vollzug als auch als für algorithmische Systeme verarbeitbaren Impuls denken zu können.*

1. Einleitung

2022 verkündete Blake Lemoine, ein ehemaliger Mitarbeiter der KI-Forschungsabteilung von Google, er habe mit LaMDA, einem von Google entwickelten selbstlernenden System der Dialogführung in natürlicher Sprache (vgl. Thoppilan et al. 2022), über dessen Bewusstsein und dessen Angst vor einer Abschaltung gesprochen, und leitete daraus ab, die Künstliche Intelligenz (KI) habe ein Bewusstsein entwickelt (vgl. dazu den Chatverlauf bei Lemoine 2022). Das Erleben Lemoines basiert auf einer Verwechslung, die die Kommunikation mit dem sie konstituierenden System (LaMDA) in eins setzt, so als wäre das im Chatverlauf Geschriebene als 1:1-Repräsentation des operativen Vollzugs der KI zu deuten. Erst diese Verwechslung produziert ein Erleben für beteiligte psychische Systeme, durch das der KI eine Empfindsamkeit zugerechnet werden kann. Aus dem Erleben heraus wird dem algorithmischen System eine mit Bewusstsein bewährte Intelligenz nicht nur zugerechnet, sondern diese Zurechnung wird zugleich auch am eigenen Erleben der Interaktion als valide bestätigt.

Die Kommunikationssequenz unter Beteiligung von Lemoine und LaMDA konstituiert sich über zwei Systemtypen, nämlich über ein psychisches System, das im Medium Sinn zu operieren vermag, und ein algorithmisches System, das nicht in der Lage ist, im Medium Sinn zu operieren (vgl. dazu auch Fuhrmann 2020). Und dennoch gelingt es LaMDA, sinnhafte Kommunikationsereignisse zu konstituieren und eine längere Interaktionssequenz aufrechtzuerhalten. Für die Fortsetzung sozialer Systeme, in diesem Fall die Fortsetzung eines Interaktionssystems, durch algorithmische Systeme muss also danach gefragt werden, wie es diesen Systemen gelingt, eine hinreichende Dynamik zu erzeugen, um *reliable outputs* zu generieren, die als Kommunikation durch Kommunikation qualifiziert werden. Oder mit den Worten Dennetts (2021: 29) formuliert: Wie gelingt es einer algorithmischen »competence without comprehension«, die komplexe Problemlösungen in einfache, triviale Lösungsprozeduren zerlegt, Kommunikation derart zu replizieren und zu variieren, dass wie bei LaMDA Kommunikationsereignisse konstituiert werden, die als sinnvolle Fortsetzung erlebt werden?

Zur Beantwortung dieser Frage muss zunächst geklärt werden, wie algorithmische Systeme Daten verarbeiten, um kommunikative Anschlussfähigkeit zu produzieren. Das wird insbesondere dann relevant, wenn davon ausgegangen wird, dass spezifische Algorithmen, etwa die des Machine Learnings, der künstlichen neuronalen Netze und stochastischer Verfahren, autopoieti-

sche Systeme konstituieren können. Diese Systemkonstitution, so die These, basiert darauf, dass autopoietische algorithmische Systeme durch systeminterne Zustandspermutationen die Bedingungen der Konstitution ihrer Operationen aus sich selbst heraus schaffen. Sie stellen dann nicht mehr allopoietische Systeme dar, die wie etwa die symbolverarbeitende KI als triviale Maschinen (vgl. von Foerster 1988: 21) einen Input gemäß den Anweisungen eines Algorithmus nach »logische[n] Schlussfolgerungsregelungen« (Ernst et al. 2019: 11f.) verarbeiten. Mit der Implementierung der Algorithmen des Machine Learnings findet eine Umstellung von einer allopoietischen zu einer autopoietischen Systemkonstitution statt, die insbesondere dadurch gekennzeichnet ist, dass Symbole nicht mehr über den Algorithmus, also die Vorschrift, wie Daten verarbeitet werden, auf eine sinnhafte Art und Weise regelhaft bearbeitet werden. Vielmehr werden über statistische Verfahren riesige Mengen an Daten verarbeitet und anstelle einer Sinnproduktion über die Verarbeitungsregel wird die durch Sinn produzierte Regelhaftigkeit in der Struktur der Daten detektiert. Statt wie in Searles (1980: 419) chinesischem Zimmer Symbole deduktiv über den Vollzug grammatikalischer Regeln zu verarbeiten, werden nun statistische Induktionen betrieben (vgl. Pasquinelli 2017). Während also allopoietische algorithmische Systeme durch die Prozedur der Verarbeitung selbst die Regeln dessen, was kommunikative Anschlüsse erfolgreich macht, auf einen jeweiligen Input anzuwenden versuchen, versucht die statistische Induktion den Input mit schon bestehenden Daten zu korrelieren und aus diesen Verarbeitungsregeln zu gewinnen. Eine solche Verarbeitung führt dazu, dass »die Inputs zum Spielball des Regelsystems selbst [werden], das sich mit jeder erneuten Eingabe selbst prüfen und bestätigen kann« (Koster 2022: 582), indem die verarbeiteten Symbole von ihren möglichen Bedeutungsgehalten entkoppelt werden und dadurch Bedeutung eliminiert wird.

Das System beginnt sich operational zu schließen, indem es durch die Datenverarbeitung seine eigenen Systemzustände, also konkrete Schaltzustände, permutiert und durch die Permutation wiederum Daten produziert, die in der weiteren Datenverarbeitung neue Systemzustände initiieren. Das System gewinnt so die Kapazität, sich durch die Operation der Veränderung von Schaltzuständen selbst fortzusetzen. Durch die Produktion von Daten, sei es durch Input, durch die Analyse von schon gespeicherten Daten oder die Erzeugung von Metadaten in der Trainingsphase, die auch interaktiv initiiert sein kann, indem Crowdworker Daten kategorisieren (vgl. Sheng/Zhang 2019), wird gleichsam eine Irritation, ein Impuls zur Fortsetzung des Systems aus der Umwelt des Systems eingespielt.

Der Artikel verfolgt also die These, dass soziale Systeme es mit Systemen zu tun bekommen, die sie mitkonstituieren, ohne dabei im Medium Sinn operieren zu können. Dabei scheint es so zu sein, dass dies der KI umso besser gelingt, je weniger sie allopoietisch über symbolverarbeitende Regel operiert. Stattdessen wird KI über die Auswertung riesiger Datenmengen (Big Data) erfolgreich. Um zu klären, wie es algorithmischen Systemen gelingt, Kommunikation (mit) zu konstituieren, wird in einem ersten Schritt gezeigt, dass KI dann erfolgreich wird, wenn sie als Konstitutionsbedingung der Kommunikation unsichtbar bleibt, also die Kommunikation derart fortgesetzt wird, dass nicht thematisiert wird, dass das einzelne Kommunikationsereignis durch ein algorithmisches System konstituiert wurde. Dabei wird angenommen, dass der kommunikative Erfolg der KI darin besteht, den Verweis auf Intelligenz selbst verschwinden zu lassen. Entsprechend wird in einem zweiten Schritt geklärt, welche Bedingungen algorithmische Systeme erfüllen müssen, um sich autopoietisch schließen zu können. Dazu wird die Unterscheidung zwischen allopoietischen und autopoietischen Systemen eingeführt. Kann die KI, die noch auf die Verarbeitung von Symbolen kapriziert ist, als Vollzug allopoietischer Systeme verstanden werden, operieren künstliche neuronale Netze und stochastische Verfahren des Machine Learnings als autopoietische Systeme. Im dritten Schritt wird dann expliziert, dass algorithmische Systeme zur Erfüllung der Autopoiesis ein doppeltes Unruhepotenzial auf der Ebene der elektronischen Verschaltungen, die die Operationen algorithmischer Systeme darstellen, konstituieren müssen. Damit folgt die Argumentation der Konzeption Luhmanns (2008a: 28), dass ein System sich sowohl aus seinen Operationen, die nur eines Typs sein können, reproduzieren als auch historische Zustände, die jeweils einmalig sind, generieren muss. Im vierten Schritt wird danach gefragt, wie autopoietisch operierende algorithmische Systeme die Kapazitäten dafür entwickeln, Kommunikation so zu verarbeiten, dass sie *reliable outputs* zur Konstitution weiterer Kommunikationsereignisse generieren können. Im letzten Schritt wird Kommunikation als eine Duplexstruktur diskutiert, die sowohl ein sinnhaftes Auslesen durch psychische Systeme ermöglicht als auch eine ungleiche Häufigkeitsverteilung von Zeichen produziert. Diese Duplexstruktur ist es, die es möglich macht, dass algorithmische Systeme sinnhaft auslesbare Kommunikationsereignisse konstituieren können, ohne selbst im Medium Sinn zu operieren. Ziel der Ausführungen ist es also, zu klären, welche Bedingungen erfüllt sein müssen, um von algorithmischen Systemen ausgehen zu können, die sich, statt allopoietisch zu operieren, als autopoietische Systeme konstituieren.

2. Zur Intelligenzfunktion der KI

In dem 1995 erschienenen Text *Über Verteilung und Funktion der Intelligenz im System* schlägt Dirk Baecker (2008) den Begriff der Intelligenzfunktion vor. Damit bezeichnet Baecker einen Wiedereintritt der Unterscheidung zwischen Selbst- und Fremdreferenz auf der Seite der Selbstreferenz, durch den diese Unterscheidung als Reflexionsschema im System operational verfügbar wird (vgl. ebd.: 61f.). Demnach ist Intelligenz dann gegeben, wenn in einem System die Selbstreferenz eines Systems, mithin die Bezeichnung des Systems durch sich selbst, von einer Fremdreferenz unterschieden wird. Dabei merkt Baecker mit Verweis auf Douglas Hofstadter an: »Die künstliche Intelligenz positioniert sich genau dort, wo ›etwas im System herausspringt und auf das System einwirkt, als wirke es von außerhalb‹. Sie macht sich zur Adresse von Fremdreferenz, zum Lieferanten jenes Wissens, auf das nur ein um sich selbst wissendes Nichtwissen schließen kann. Sie wird zu einem Reflexionspotential, das in dem Maße, in dem es auf die Formulierung jeder eigenen Selbstreferenz verzichtet, Zugang zu den rekursiven Schleifen eines sozialen Systems gewinnt und aus diesen Rekursionen produziert und reproduziert wird.« (Ebd.: 63) KI könne da auf ein soziales System einwirken, wo sie »genügend Eigenkomplexität besitzt, um interne Verweisungsmöglichkeiten selbstständig und innovativ wahrzunehmen« (ebd.). Sie wird gleichsam dadurch zum Reflexionspotenzial für die Kommunikation selbst, weil sie einen Input für das soziale System liefert, der vom System als Fremdreferenz eingeordnet werden kann. KI hält entsprechend Kommunikation am Laufen, ohne die Intelligenzfunktion als reflexives Schema von Selbst-/Fremdreferenz selbst operativ zu realisieren, indem sie Kommunikation nutzt, um ihre eigenen Operationen aus deren Verarbeitung zu reproduzieren, und dadurch Kommunikation mit Varianz versorgt, die weitere kommunikative Anschlüsse evozieren kann.

Die hier kurz referierte und inzwischen schon fast 30 Jahre alte Textstelle erscheint deswegen interessant, weil das Problem der KI auf eine ambivalente Weise zu lösen gesucht wird. Die Intelligenzfunktion wird an den Werten der Reflexivität, der Rekursivität und der Paradoxieentfaltung fixiert und auf sinngenetisch operierende Systeme bezogen, die die Kapazität aufweisen, ihre Selbstreferenz in Relation zur Fremdreferenz setzen zu können. Im Moment der Reflexion, durch die so etwas wie ein Selbstbewusstsein des Systems als Kalkül von Selbst- und Fremdreferenz artikulierbar wird, wird diese systemtheoretische Gewissheit sogleich wieder gebrochen, indem KI Zugriff auf das soziale System gewinnt, und dies umso mehr, je weni-

ger sie ihre Selbstreferenz systemintern reflektiert. KI produziert lediglich Referenz, also eine eigene Operation im Modus des Wechsels differenter elektronischer Verschaltungen, und führt ihre Outputs in die Kommunikation als eine Fremdreferenzen ein, ohne die artikulierte Differenz zwischen Selbst- und Fremdreferenz in ihrer systeminternen Verweisung selbst nachzuvollziehen. Algorithmische Systeme sind dann weniger dazu aufgefordert, reflexive Ich-Artikulationen als Bezeichnung eines Selbst zu vollziehen, als vielmehr Kommunikationen zu produzieren, an die weitere Kommunikationen anschließen können. Ich-Artikulationen, wie sie von Sprachassistenzen produziert werden, sind keinesfalls als systeminterner Gebrauch einer Selbst-/Fremdreferenz-Unterscheidung zu deuten (vgl. Lang 2020: 7f.). Vielmehr verweist das *Ich* als Bezeichnung einer Sprecher*innenrolle auf eine Pragmatik und Konvention, ohne dabei die Eigenschaften der Sprecher*in repräsentieren zu müssen (vgl. Helmbrecht 2004: 23f.). Gerade das Ausbleiben des Prüfens dessen, was es bedeutet, wenn die Sprachassistenz ein *Ich* artikuliert, kann als Akzeptanz der Leistung aufgefasst werden, anschlussfähige Kommunikation zu konstituieren. Entsprechend kann, wie im Erleben Lemoines zu sehen ist, die kommunikative Artikulation eines Ich-Bewusstseins mit der Funktion der Konstitution dieser Artikulation verwechselt werden. Das Ich markiert statt eines Bewusstseins vorerst eine sprachliche Pragmatik, mittels derer das outputgenerierende System, in diesem Fall das algorithmische System, kommunikativ adressierbar gemacht wird. In diesem Sinne dient eine Ansprache der Sprachassistenz wie *Hey Google*, *Hey Alexa* oder *Hey Siri* dem System als Triggerphrase, um eine Prozedur der Sprachanalyse zu starten. Gleichsam signalisiert die Pragmatik der Ich-Artikulation beteiligten psychischen Systemen, dass die Sprachassistenz responsiv gegenüber der Kommunikation gesetzt wurde, also zu erwarten ist, dass die Sprachassistenz auf die eigene Artikulation antworten wird. Das bedeutet, dass KI dann erfolgreich wird, wenn sie in der Kommunikation Kommunikationsereignisse konstituiert, die vom Verweis auf das systemische Selbst der KI entkoppelt werden können. Die KI wird damit zum Produzenten einer artikulierten Selbstreferenz, die im Falle von LaMDA auch auf eine Fremdreferenz – die Angst, abgeschaltet zu werden – bezogen sein kann. Dies markiert aber nicht die Leistung der KI, sich selbst bewusst zu werden, sondern den Effekt der Kommunikation, eine Selbst-/Fremdreferenz-Unterscheidung artikulierbar werden zu lassen. Damit wird sogleich auf eine etwas umständliche Weise Turings Imitation Game (Wiener 1990: 93ff.) implizit reformuliert. KI stellt sich dann ein, wenn in der Kommunikation nicht mehr zwischen Mensch und Maschine unterschieden

wird. Mit dem Computer, so könnte mit Luhmann (1998: 117f.) hinzugefügt werden, tritt nun eine »unsichtbare Maschine« als neue Konstitutionsmöglichkeit von Kommunikation auf. Für die Anschlussfähigkeit wäre, so die hier entwickelte These, eine Invisibilisierung des die Kommunikation konstituierenden Systems als Bedingung zu benennen. Denn die Verarbeitung der Kommunikation wird in Kommunikation nicht transparent gemacht (vgl. Kaminski 2020: 153), so, wie das auch für die Operabilität psychischer Systeme gilt (vgl. Luhmann 2008b: 22f.; 2017: 104ff.), und kann lediglich als Thema kommunikativ wieder eingefangen werden.

Der von Baecker artikulierte Zugang algorithmischer Systeme zu den rekursiven Schleifen sozialer Systeme, der eine Art Eingriff durch die KI suggeriert, so als wäre sie in die Autopoiesis sozialer Systeme eingebaut (vgl. Nassehi 2019: 259f.), lässt sich nun präziser fassen als Inklusion autopoietischer algorithmischer Systeme in Kommunikation (zum Begriff der Inklusion vgl. Luhmann 1989: 162). Denn ihre Leistung besteht darin, ihre Eigenkomplexität für die Bildung des sozialen Systems bereitzustellen, ohne Teil des sozialen Systems zu werden. Algorithmische Systeme können dann analog zu psychischen Systemen an der Konstitution sozialer Systeme beteiligt sein. Sie werden insofern inkludiert, als sie entsprechende operationale Kapazitäten bereithalten und dadurch die Emergenz sozialer Systeme befeuern. Entsprechend treten algorithmische Systeme als ein Äußeres der Kommunikation auf, das Kommunikation in ihre Eigenlogik übersetzen muss.

Für Nassehi (2019: 249f.) hingegen kann von handelnden und erlebenden Algorithmen nur ausgegangen werden, wenn Handeln und Erleben durch Zurechnung entstehen. So werden Roboter und Bots inzwischen als elektronische Personen anerkannt, die im Rechtssystem als juristische Personen behandelt werden und für Schäden haftbar zu machen sind (vgl. Teubner 2018: 160f.). So erwerben Trading-Algorithmen Eigentum, stoßen es wieder ab, legen Konten an und verwalten sie. Im Akt des Tradings erfolgt eine Zahlung, die keiner Thematisierung der Zurechnung von Handlung und Erleben bedarf. Folgt man Nassehi, verlieren sie dadurch ihren Akteursstatus und können deshalb als in die Autopoiesis des sozialen Systems der Wirtschaft eingebaut interpretiert werden. Wird hingegen angenommen, dass autopoietische algorithmische Systeme als konstitutives Außen auftreten, also in die Fortsetzung sozialer Systeme inkludiert sind, wird ihre Operabilität intransparent für das soziale System und bleibt auch dann relevant, wenn ihnen kein Handeln oder Erleben zugerechnet wird. So werden in Trading-Algorithmen Rechenheuristiken eingebaut, die es ermöglichen sollen, die Effekte der eigenen Kauforder auf

das Geschehens an der Börse zu berechnen. Vollautomatisierte Systeme können, je sensibler sie auf die Volatilität der Börse und die Effekte der eigenen Kauf- bzw. Verkaufsorder bezogen sind, die Volatilität verstärken und werden dadurch störanfällig. Angesichts dessen werden permanent menschliche Trader*innen eingesetzt, die die Transaktionen der Algorithmen überwachen und beurteilen, ob das Agieren eines Algorithmus als sinnvoll bewertet werden kann (vgl. Wansleben 2012: 244ff.). Die Zurechnung von Sinn erfolgt erst in einem sich der konkreten Transaktion entziehenden Nachtrag, der gegebenenfalls zur Intervention der menschlichen Trader*in führt. Das algorithmische System konstituiert von außen eine Operation (Kauf- bzw. Verkaufsorder) im Inneren des sozialen Systems (Börse) und wird wiederum von außen (Trader*in) darauf kontrolliert, ob es sich in selbstproduzierten Feedbackschleifen festfährt – Feedbackschleifen, die zum Kollaps führen können, wenn verschiedene algorithmische Systeme zu konsonieren beginnen (historische Beispiele finden sich bei Miyazaki 2017). Wird dieser Fall verallgemeinert, dann treten algorithmische Systeme für Kommunikation als ein konstitutives Außen, als eine eigene Systemqualität auf, die nicht in das soziale System eingebaut ist, sondern dieses mit anderen konstituiert. Dieses Moment ist nicht als Verwechslung von KI mit *artifical communication* zu markieren, so als handelte es sich nicht um Kommunikation, sondern nur um deren Emulation als eine »projection of the contingency of its user« (Esposito 2022: 10) bzw. die Reflexion der Perspektiven anderer Beobachter. Vielmehr handelt es sich um eine spezifische Leistung der Intelligenzfunktion selbst, die semantisch darin bewährt wird, dass der KI hin und wieder Intelligenz zugerechnet wird. Die Kapazitäten zur Aufrechterhaltung ihrer Konstitutionsleistung, so die These, können algorithmische Systeme nur gewinnen, wenn sie sich nicht als allopoietische Systeme, die sich als technische Infrastruktur in das soziale System einbauen lassen, sondern als autopoietische Systeme konstituieren.

3. Autopoietische algorithmische Systeme

Der Begriff der Autopoiesis verweist auf Systeme, die durch den Vollzug ihrer Operationen in der Lage sind, sich selbst zu reproduzieren. Für soziale Systeme sind diese Operationen Kommunikationen, für psychische Systeme Kognitionen, was sinnliches Wahrnehmen und Denken einschließt (vgl. Luhmann 2008a: 30f.). Algorithmische Systeme operieren in Form von Permutationen elektronischer Schaltzustände, durch die weitere elektronische

Schaltzustände produziert werden. Um von einem autopoietischen System sprechen zu können, müssen zwei Bedingungen erfüllt sein. Die erste besteht darin, »daß Operationen aneinander anschließen und damit eine Kontinuität des Operierens herstellen« (ebd.: 28), wodurch das System von seiner Umwelt unterscheidbar wird. Die zweite Bedingung besteht darin, dass das Operieren zu historischen Zuständen des Systems führt, also zu einmaligen Zuständen, in der Weise, dass die jeweilige Selektion zu Variationen führt, das System sich also nicht in der Wiederholung der gleichen Operation leerläuft (vgl. ebd.). Diese beiden Momente markieren ein doppeltes Unruhepotenzial des autopoietischen Systems, durch das es sich so weit selbst irritiert, dass es selbstreproduktiv zu operieren beginnt.

Die System/Umwelt-Grenze des algorithmischen Systems wird bestimmt durch die Implementierung von Vorgaben, die festlegen, welche Operationen in welcher Reihenfolge ausgeführt werden sollen und wie diese aufeinander bezogen sind und sich gegenseitig mit Inputs versorgen. Dementsprechend sind algorithmische Systeme nicht mit Computern gleichzusetzen, weil jene als Medium der Kapazität für die spezifisch durch einen Algorithmus kombinierten Schaltungen erst durch das algorithmische System in eine operationale Form gebracht werden (zur Unterscheidung zwischen Medium und Form vgl. Fuchs 2015: 25–30), die es gestattet, eine Autopoiesis durch die Implementierung eines systeminternen Unruhepotenzials hervorzubringen.

Autopoiesis markiert dabei keine Autarkie in dem Sinne, dass das System seine Operationen in völliger Unabhängigkeit von seiner Umwelt vollziehen könnte. Vielmehr bezieht sich der Begriff auf ein System, das autonom operiert, indem es an seine eigenen Operationen weitere Operationen anschließt und dabei die Bedingungen mitproduziert, unter denen die Operationen angeschlossen werden können. Entsprechend schließt sich das System operativ, bleibt aber offen für Irritationen, die eine notwendige Voraussetzung für die Konstitution der jeweiligen Operationen sein können. Kommunikation setzt entsprechend eine dynamische Umwelt psychischer Systeme voraus, die als konstitutives Außen einen Noise produzieren, den Kommunikation im Kontext der jeweiligen Systemkomplexität und der jeweils gegenwärtigen Struktur in die Ordnung ihrer Operationen überführt (vgl. von Foerster 1999: 123; Dupuy 2015: 5). Psychische Systeme treten mit sozialen Systemen entsprechend in eine konditionierte Koproduktion ein (vgl. Fuchs 2002), ohne dabei die Autonomie des jeweils anderen Systems zu destruieren – ganz im Gegenteil ist es diese gegenseitige Irritation, durch die sie sich am Laufen halten. Analog dazu verhalten sich algorithmische Systeme, die im Fall einer Autopoiesis ei-

ne hier noch näher zu beschreibende Autonomie gewinnen. Auch sie setzen ein sie konstituierendes Außen voraus (vgl. Fuhrmann 2019: 266, Fn. 722), mit dem sie in konditionierte Koproduktion treten können. Das ist Kommunikation.

Entscheidend ist also nicht, ob algorithmische Systeme autark operieren, so als wären sie hermetisch abgeschlossen von ihrer Umwelt und würden eine nur aus sich selbst heraus stimulierte Operabilität etablieren, sondern dass sie ein doppeltes Unruhepotenzial mobilisieren, durch das sie in eine konditionierte Koproduktion mit ihrer Umwelt treten können. Der temporäre Stillstand von algorithmischen Systemen in dem Moment, in dem kein Input mehr erfolgt und alle Prozeduren beendet wurden, diskreditiert deshalb nicht die These der autopoietischen Schließung algorithmischer Systeme. Genauso wie Kommunikation nicht mehr operativ konstituiert werden kann, wenn ein Interaktionssystem abbricht, weil eine gegenseitige Wahrnehmung der beteiligten psychischen Systeme nicht mehr möglich ist, da etwa alle anwesenden Personen den Raum verlassen haben, also nicht mehr anwesend sind (vgl. Kieserling 1999: 110), setzen autopoietische algorithmische Systeme weitere Systeme in ihrer Umwelt voraus, um so weit irritiert zu werden, dass ihre systemeigenen Operationen nicht leerlaufen. Um die Autopoiesis algorithmischer Systeme beschreiben zu können, ist es also relevant, ein doppeltes Unruhepotenzial als Verarbeitung sie konstituierender Kommunikation zu identifizieren.

Die Operation algorithmischer Systeme wird an dieser Stelle mit der Permutation von materiell in Transistoren implementierten elektronischen Schaltzuständen assoziiert. Die jeweiligen Zustände verhalten sich diskret zueinander (vgl. Steiglitz 2019). Das heißt, algorithmische Systeme sind als Implementierung und Vollzug der Ansteuerung spezifischer Schaltungen aufzufassen, deren Operabilität darin besteht, gemäß den Schaltungen Schaltzustände zu permutieren. Berechnungen werden durch geschickte Kombinatorik von Schaltelementen vollzogen. Die theoretische Entscheidung, die Operation algorithmischer Systeme nicht an der Berechnung, sondern an der Permutation von Schaltzuständen festzumachen, lässt sich daran plausibilisieren, dass die durch die Vorschrift festgelegten Befehle von Algorithmen nur dann prozedieret werden, wenn sie in Schaltkreisen repräsentiert werden. Auch künstliche neuronale Netze basieren zumeist nicht allein auf der Hardware, sondern müssen in der Rechnerarchitektur simuliert werden, also von der Software in eine durch die Hardware prozedierbare Befehlsabfolge transponiert werden (vgl. Rojas 1993: 399f.). Das algorithmische System nutzt den Computer als Medium dafür, dass über die Implementierung der spezifischen Sequenz der Ansteuerung von Schaltungen Operationen

als Permutation von Schaltzuständen ausgeführt werden können. Programmcodes höherer Programmiersprachen müssen entsprechend vom Compiler übersetzt und vom Interpreter in einen Schaltzustand überführt werden. Das heißt, dass eine Übersetzung in boolesche Algebra erfolgen muss, die es ermöglicht, über die Ansteuerung der logischen Schaltelemente UND, ODER, NICHT Berechnungen und Verrechnungen zu vollziehen.

Über die boolesche Algebra fällt das Berechnen und Verrechnen von Inputs mit der Permutation von Schaltzuständen logischer Schaltungen zusammen. Denn die logischen Schaltungen stellen ein räumliches Prinzip dar, das bei UND dann einen Impuls weitergibt, wenn im Input eine Gleichzeitigkeit aller Impulse gegeben ist, wohingegen bei ODER nur ein Impuls gegeben sein muss und die NICHT-Schaltung den Impuls des Inputs im Output invertiert. Die dabei geschalteten logischen Schaltkreise lassen sich so kombinieren, dass auch komplexe numerische Berechnungen wie die stochastischen und korrelativen Verfahren des Machine Learnings vollzogen werden können, indem die dafür notwendigen mathematischen Operationen (Addition, Subtraktion, Multiplikation und Division) durch die spezifische Kombination von Schaltungen ausgeführt werden. Die Berechnung, etwa die für die Summenbildung statistischer Auswertungen notwendige Addition, beruht dann weniger auf einer Operation des Zählens, sondern auf einer Schaltung, die mittels UND und ODER Bitfolgen ausgibt, die aus der durch den jeweiligen Input bewirkten Permutation der Zustände der jeweiligen in Reihe geschalteten Schaltungen resultieren.

Allein der Vollzug der Permutation von Schaltungen reicht nicht aus, um ein algorithmisches System operativ zu schließen. Erst eine interne Irritabilität der Permutation von Schaltzuständen durch die bisherigen Zustände des Systems gestattet es, von Autopoiesis zu sprechen. Ohne Irritabilität handelt es sich um allopoietische Systeme, bei denen ein spezifischer Input zu einem spezifischen, immer wieder reproduzierbaren Output führt (zu dieser Unterscheidung vgl. Zeleny 1976: 13). Angesichts dessen ist es nicht die Abfolge vorab programmierter Operationen, die für die Autopoiesis algorithmischer Systeme konstitutiv ist, sondern es ist die Perpetuierung des Systems durch interne Rückkopplungen, durch die auf der Ebene der elektronischen Verschaltungen eine Dynamik entsteht, die bewirkt, dass das System seine Operationen aus sich selbst heraus immer weiter fortsetzt. Entsprechend weisen algorithmische Systeme, beispielsweise *Automated Decision Making Systems*, mindestens zwei Algorithmen auf: einen, durch den eine Bewertung erfolgt oder eine Prognose aufgestellt wird, und einen zweiten Algorithmus, der hinsichtlich des

Inputs des ersten Algorithmus nach spezifischen Parametern eine Entscheidung trifft (vgl. Zweig 2018: 12). Darunter fallen insbesondere künstliche neuronale Netze und stochastische Verfahren des Machine Learnings, mittels derer die jeweiligen Outputs von Schaltungen so miteinander verbunden sind, dass die Veränderung einer Schaltung innerhalb der Einheit eines Schaltkreises als Feedback geschaltet ist und/oder zwischen den einzelnen Einheiten von Schaltkreisen ein Feedback – also eine Rückkopplung, die das Auslösen eines Schaltzustandes durch einen anderen ermöglicht – in diese selbst eingeführt wird und als Permutationsbedingung der Schaltung rückkoppelnd wirkt (vgl. Brause 1991: 53f.).

Ein Vergleich von allopoietischen mit autopoietischen algorithmischen Systemen kann zeigen, inwiefern die Implementierung voneinander abweicht. ELIZA, ein Chatprogramm, das 1966 von Joseph Weizenbaum vorgestellt wurde, kann als ein allopoietisches algorithmisches System bezeichnet werden. Denn ELIZA ist so implementiert, dass nach einem User*innen-Input schrittweise eine Folge von Operationen abgearbeitet wird. Nach der Vollendung dieser Operationenfolge wird von ELIZA keine weitere Operation, also Permutation von Schaltzuständen, vollzogen, bis der nächste Input erfolgt. Die symbolorientierte KI-Forschung hatte versucht, mit allopoietischen Systemen für jeden möglichen Inputfall eine bestimmte Prozedur aufzurufen. Bei ELIZA wird etwa in einem Register, das einem Thesaurus ähnelt, nach äquivalenten Worten gesucht, um eine Variation der Eingabe generieren zu können. Ist im Register kein passender Eintrag zu finden, wird eine Auffangphrase ausgegeben (vgl. Storp 2002: 19). Allopoietische Systeme sind maßgeblich abhängig von einem äußeren Input, durch den die jeweiligen Prozeduren aufgerufen und als Permutation von Schaltzuständen vollzogen werden. Der dabei generierte Output dient dem System nicht dazu, weitere Schaltzustände zu permutieren. Autopoietische Systeme hingegen reproduzieren sich durch ihre eigenen Operationen. Das heißt, dass sie zwar inputsensitiv sein müssen, um auf Änderungen der Umwelt eingehen zu können. Jedoch fungiert der durch eine Prozedur generierte Output wiederum als systemeigener Input und produziert dadurch eine Historizität des Systems, die gleichsam auf einer Selbstordnung beruht, etwa mit Blick auf die spezifische Gewichtung der Neuronen eines künstlichen neuronalen Netzes. Diese Logik macht sich auch durch die Variation des Outputs bemerkbar. Bei LaMDA kann im Gegensatz zu ELIZA von einer Autopoiesis ausgegangen werden, weil der User*innen-Input zu einer Permutation der Schaltung im Prozess der Modellierung eines künstlichen neuronalen Netzes selbst führen

kann, indem im Abgleich mit bisherigen Schaltzuständen die Gewichtung einzelner Funktionen des gesamten Netzes variiert wird (vgl. Rojas 1993: 24f.) und so je neue historische Zustände produziert werden. Diese Funktionen werden in boolescher Algebra dargestellt (vgl. ebd.: 34ff.) und operativ über die Permutation von Schaltzuständen realisiert. Das zeigt sich etwa darin, wie LaMDA trainiert wurde, um sinnhafte Antworten hervorbringen zu können. So bewerteten die Crowdworker*innen im Rahmen der zweistufigen Trainingsphase die von LaMDA generierten Antworten zunächst im Hinblick darauf, ob sie Sinn ergaben (vgl. Thoppilan et al. 2022: 34), und sofern sie nicht als sinnhaft bewertet wurden, reformulierten sie in einem zweiten Schritt die Antworten und markierten zusätzlich, ob LaMDA bei der Antwortgenerierung weitere externe Textressourcen berücksichtigen solle, um mehr kontextuale Bezüge herstellen zu können (vgl. ebd.: 8). Auf diese Weise kann ein algorithmisches System wie LaMDA in die Lage versetzt werden, sich autopoietisch zu schließen. So entstehen aus der durch jeweils differente Dateninputs bewirkten Permutation von Schaltzuständen und aus den jeweils spezifischen Topologien der Verschaltung des künstlichen neuronalen Netzes historisch einmalige Zustände des Systems.

Das zweite Unruhepotenzial ergit sich (a) aus einem transitorischen Moment der elektronischen Schaltung und (b) aus einem zeitstabilen Speichermoment, durch das das System erst historische Zustände produzieren kann.

(a) Die transitorische Speicherung der jeweiligen Systemzustände als Output einer Berechnung in boolescher Algebra, die als Ausgangspunkt weiterer Operationen angesteuert wird, folgt dem Prinzip des Registers. Ein Zwischenspeichern von Schaltzuständen, um weiter rechnen zu können, erfolgt in einem Schaltkreis, der als Flipflop dadurch bedingt wird, ob der kurzzeitige Speicher durch Offenhalten oder Schließen eines Stromkreises durch weitere Operationen angesteuert und dadurch umgeschaltet wird (vgl. Stokes 2007: 1–17). Im Transistor lassen sich so kurzfristig Zustände speichern, die aber direkt wieder überschrieben werden, also durch Input eines anderen Zustands permutieren (vgl. Ernst 2018: 109f.). Mit der elektronischen Verschaltung lassen sich Zustände speichern, indem ein Zustand gehalten wird, das heißt, dass keine Permutation der Partialschaltung erfolgt. Eine solche Repräsentation von Zuständen in der Präsenz der Verschaltung des Systems produziert kein Gedächtnis in dem Sinne, dass eine Erinnerung an vorherige Zustände vollzogen wird. Algorithmische Systeme sind nicht imstande, das Gegenwärtige zum gegenwärtig Vergangenen in ein temporales Verhältnis zu setzen, wie es für das Gedächtnis eines Sinnsystems möglich ist, das Erinnertes auf Redun-

danz und die Varietät einer sinnhaften Kohärenz bezieht (vgl. Esposito 2002: 29f.). Vielmehr wird durch das Halten von Zuständen in Partialschaltungen einer übergeordneten größeren Schaltung eine räumliche Relation genutzt, um trotz der Bindung an diskrete Zustände und den Präsenzeffekt dessen, wie die jeweiligen Schaltungen geöffnet oder geschlossen kombiniert werden, vorherige Zustände in der aktuellen Schaltung verfügbar zu halten. Durch diese Halteoption in einer Partialschaltung kann so etwas wie ein Gedächtnis entzeitlicht und durch eine räumliche Organisation verräumlicht werden. Konrad Zuse (1969) bezeichnet rechnende Maschinen deshalb als *rechnenden Raum*. Im rechnenden Raum lassen sich Parallelrechnungen vollziehen, die die Schließung algorithmischer Systeme ermöglichen.

Durch die Parallelisierung der Permutationen von Schaltzuständen findet eine Steigerung der systeminternen Unruhe in der Weise statt, dass die Outputs verschiedener Schaltungen in einer Bitfolge zusammengeführt werden können, die als Input für weitere Verschaltungen dienen kann, wie es etwa bei einem Volladdierer der Fall ist, der bei der Berechnung eines Bits einen Übertrag, der wiederum die Berechnung der nächsten Bitstelle ermöglicht, erzeugt (vgl. Broy 1998: 318f.). Über die Bittiefe, also die Anzahl der von den Schaltungen ausgegebenen Outputs, lassen sich Zustandssequenzen der akuten Schaltungen auf die Sequenz des Outputs reduzieren und damit abspeichern, allein deswegen, weil nicht mehr die Schaltkreise mit ihren aktuellen Schaltzuständen gespeichert werden müssen, sondern es lediglich notwendig ist, die Abfolge binärer Bits zu speichern. Speichern kann so als eine Trivialisierung von Schaltungen aufgefasst werden – Informationen über den Input gehen dadurch verloren (vgl. Rojas 1993: 22).

(b) Das zweite Moment findet sich im Speichern von Daten. Dabei handelt es sich weniger um die Kulturtechnik des Schreibens als vielmehr um die Überführung einer spezifischen Schaltkombination in zur Speicherung vorgesehene Schaltkreise, bei denen es sich um einen »unveränderliche[n] Festwertspeicher« (Ernst 2018: 109) handelt, der eher als Archiv denn als Register fungiert. In diesem Speicher wird keine boolesche Algebra verrechnet, sondern lediglich gemäß dem Input eine Bitfolge permutiert. Dadurch wird die Datei in einer räumlichen Repräsentation zeitfest gemacht, solange die Schaltung des Speichermediums nicht permutiert wird. Daten können nun als Bitfolge in Form eines Inputs einen Schaltkreis erneut permutieren, und werden so jeweils durch Schaltungen selbst gegenwärtig verfügbar und unverfügbar gemacht (vgl. Oberschelp/Vossen 2006: 232, 407ff.). Indem Daten gespeichert und ausgelesen werden, also durch Verschaltung für den rechnenden Raum

als Input zur Verfügung stehen, produziert das System nicht nur die aktuelle Unruhe der sich jeweils mit Input gegenseitig versorgenden Schaltungen, sondern gewinnt aus den historischen Zuständen des Systems, die im Speichermedium zeitstabil verschaltet wurden, auch ein historisches Unruhepotenzial.

Die historischen Schaltzustände können als Daten bezeichnet werden, die dem System als Input dienen, sobald sie gelesen werden. Datenlesen heißt in diesem Fall, dass eine Bitfolge dem System dazu dient, seine Schaltzustände zu permutieren, womit die zweite Bedingung der Autopoiesis erfüllt ist. Die drei Modi der historischen Unruhe finden sich im Schreiben von Daten, das darin besteht, eine Bitfolge in einem Schaltzustand zu speichern, im Verrechnen von Daten, also dem Input mehrerer Daten als Bitfolge zur Permutation einer Schaltung, sowie im Lesen von Daten. Die Daten werden als Input in Form einer Bitfolge einem Schaltkreis verfügbar gemacht, perturbieren also den aktuellen Schaltzustand. Die derart produzierte Unruhe ist dabei nicht als permanenter Basisrumor des Systems zu verstehen, vielmehr geht es darum, dass das System immer wieder mit systeminternen Inputs versorgt wird, sodass die Permutationsbedingungen weiterer Operationen variiert werden. LaMDA ist dafür ein Beispiel, da durch das künstliche neuronale Netz eine Selbstorganisation der Verarbeitung der Inputs, also der Eingaben in den Chat, bewerkstelligt wird, mittels derer, je nach Eingabe, der Input ein historisches Unruhepotenzial durch Öffnung für externe Inputs von Wissensdatenbanken produziert. Das autopoietische algorithmische System gewinnt dadurch seine Autonomie, ohne autark gegenüber der es in Form eines Impulses triggernden konstitutiven Umwelt zu sein.

4. Die Übersetzung kommunikativer Temporalität in den rechnenden Raum

Wenn die Operabilität algorithmischer Systeme als räumliche Organisation je aktueller Schaltzustände aufgefasst wird, operiert das algorithmische System – egal, ob es sich um ein allopoietisches oder um ein autopoietisches handelt – insofern zeitlos, als zwar Zustände permutieren, also in der Zeit differieren, diese aber nicht relationiert werden, sondern diskret auf ihren Präsenzeffekt (vgl. Maskarinec 2010: 80f.) limitiert bleiben. Das System kennt immer nur seinen aktuellen Zustand. Kommunikation als Zeitproblem muss von algorithmischen Systemen folglich in ein Problem der Organisation des rechnenden Raums überführt werden – die Rede von der Topologie künstlicher neurona-

ler Netze verdeutlicht das. Konkret vollziehen tut sich diese Überführung, indem die Zeitlichkeit der Kommunikation in eine Repräsentation von Bitfolgen übersetzt wird, wodurch sie als Permutationsbedingung durch einen elektronischen Impuls räumlich verarbeitbar wird. Der vorherige Zustand, also Vergangenheit, wird eliminiert (vgl. Pias 2009: 265).

Algorithmische Systeme nutzen dabei ihre aktuelle Unruhe zur Mobilisierung des historischen Unruhepotenzials, indem sie Daten auslesen und die in diesem Zuge miteinander korrelierten Zeichen wiederum als Moment der Steigerung oder Schwächung der Korrelation ins Datenkorrelat einzuschreiben beginnen. Die algorithmische Übersetzung von Kommunikation greift auch auf schon produzierte Daten zurück. Das ist das Motiv des Data Minings als dreigliedriger Prozess der Übersetzung in Bits, der kohärenten Aggregierung der Daten und ihrer Analyse in Form von Clustering, Assoziation und Korrelation (vgl. Hildebrandt 2011: 376). Was vom System als Input aufgegriffen wird, wird in der Datenverarbeitung manipuliert, produziert aber gleichsam jenes doppelte Unruhepotenzial, durch die die Autopoiesis des algorithmischen Systems aufrechterhalten werden kann. Je nach Implementierung können Rückkopplungen zur Dynamisierung der Gewichtung einzelner Parameter und Neuronen des Systems, die nichts anderes als logische Verschaltungen sind, führen, etwa durch das Verfahren der Backpropagation, bei dem Rückkopplungen eingebaut werden (vgl. Rojas 1993: 175–193), oder durch genetische Verfahren, bei denen Outputs zufälliger Variationen von Gewichtungen einzelner Neuronen mit vorgegebenen Gewichtungen abgeglichen werden (vgl. Brause 1991: 254).

Statt Temporalität besteht der relevante Operationsmodus für die Datenanalyse in der Berechnung von Häufigkeitsverteilungen. Die Häufigkeitsverteilung wird dabei nicht durch einen Abgleich verschiedener vorheriger Zustände des Systems bestimmt, so als protokollierte ein Beobachter die Systemzustände. Stattdessen muss sie in eine Schaltung übersetzt werden, die eine Berechnung der Häufigkeitsverteilung mittels boolescher Algebra ermöglicht. Die boolesche Algebra verweist auf binäre Tupel (Bits), die wiederum durch Konventionen des Lesens in arabische Ziffernfolgen übersetzt werden können. So wird die arabische Ziffer 0 in einer Bitfolge aus 0, die 1 in einer Bitfolge, deren erste Stelle zur 1 permutiert wird, die 2 in einer Bitfolge, bei der an zweiter Stelle die 1 permutiert wird, und die 3 in einer Bitfolge, bei der die beiden ersten Stellen mit einer 1 permutiert werden, repräsentiert. Bitfolgen wie 0001, 0010, 0011, hier bei einer Bittiefe von vier Bits, repräsentieren dann die Dezimalzahlen der arabischen Ziffern, müssen also übersetzt werden. Solche

Tupel können mit Spencer-Brown (1999) als Formen bezeichnet werden, die gemäß einer »nicht-nummerische Arithmetik« (ebd.: xxvi) kalkulierbar sind. Das Oszillieren zwischen den Werten 0 und 1 produziert durch Reihung von mehreren Tupeln Oszillationsmuster (vgl. ebd.: 54f.). Diese Muster sind nicht mit dem gleichzusetzen, was Nassehi (2019: 108f.) als die latenten Strukturen der Gesellschaft beschreibt, die durch Algorithmen und Datenform verdoppelt würden und damit auf die Digitalität der Gesellschaft selbst verweisen. Sie stellen stattdessen eine Kombination aktueller Schaltzustände dar, die in der Parallelität von Tupeln, verschaltet über Permutationen bei Abweichungen und Übereinstimmungen, jeweils differente Permutationen als Effekt zeitigen. Erst mittels Konventionen lassen sie sich in numerische Modelle übersetzen und werden so als Sinnzusammenhänge lesbar. In der Trainingsphase des Machine Learnings wird die Struktur der Verschaltung selbst permutiert, indem Oszillationsmuster so lange variiert werden, bis das System aus der vorherigen Mobilisierung des historischen Unruhepotenzials eine eigene Struktur gebildet hat und damit kommunikative Outputs generieren kann.

Führt ein Output eines solchen Systems in einem sozialen System zu Artikulationen in Form von Kommunikation, dynamisiert das algorithmische System seine Umwelt und steigert dadurch die Möglichkeiten zur Generierung weiterer Inputs für Kommunikation. Es tritt damit als ein konstitutives Außen des Kommunikationsereignisses auf und setzt so – äquivalent zu psychischen Systemen – die Autopoiesis eines sozialen Systems fort, durch dessen Autopoiesis wiederum eine Zustandspermutation des algorithmischen Systems stimuliert werden kann. Gelingt diese Dynamisierung, dann erscheint das algorithmische System unter Umständen als Akteur im System, dem Intelligenz zugeschrieben werden kann, weil es etwa wie im Fall von LaMDA in der Kommunikation eine Selbst-/Fremdreferenz-Unterscheidung artikuliert.

Gerade die algorithmische Spracherkennung als Pars pro Toto des Übersetzungsproblems von Kommunikation in elektronische Schaltzustände zeigt das an. Lange wurde versucht, die Regelhaftigkeit der Sprache, also ihre Grammatik, durch Implementierung in allopoietische Algorithmen zu übertragen und so ein regelbasiertes Verstehen zu simulieren (vgl. Beckermann 1988: 68). Modelle der künstlichen neuronalen Netze prozedieren hingegen Häufigkeitsanalysen von Zeichen und Zeichengruppen in der Nachbarschaft von anderen Zeichen und Fragmenten von Zeichenketten, um anhand dieser Verteilungen die Varianz der in der Nachbarschaft möglichen Zeichen so weit zu reduzieren, dass der Wahrscheinlichkeitswert der jeweils benachbarten Zeichen gegen 1 tendiert (vgl. Scha et al. 1999). Prozessiert wird diese Berechnung der ge-

genseitigen Abhängigkeit und Restringierung von benachbarten Zeichen zueinander durch eine gesteuerte Permutation aktueller Zustände der Verschaltung, die durch den aktuellen, über ein Interface ins System gelangenden Input wie auch über die historische Unruhe des Systems bewirkt wird. Auf diese Weise kann sowohl das jeweilig analysierte Zeichen vereindeutigt als auch die Zeichensequenz aus der Bestimmung der jeweiligen Einzelzeichen im Ganzen fortgesetzt werden (vgl. Fissore et al. 1990: 21f.).

Das autopoietische algorithmische System ist in der Diskretheit seiner Zustände gefangen. Es muss wie bei der Sprachassistenz explizit mit der Triggerphrase *Hey Alexa*, *Hey Siri* usw. adressiert werden, um den Startpunkt einer Kommunikationssequenz identifizieren zu können, die dann in eine Bitfolge transponiert und analysiert wird (vgl. Patel/Patil 2019). Die für die Adressierung genutzte Triggerphrase setzt dabei eine spezifische Abfolge von Zeichen voraus und dient als Input, der die Schaltung so weit permutiert, dass die Sprachassistenz beginnt, die an die Triggerphrase angeschlossene Kommunikationssequenz zu analysieren. Die Signatur der Abfolge von Zeichen einer Triggerphrase lässt sich aber auch so weit dynamisieren, dass sie über eine spezifische Häufigkeitsverteilung benachbarter Zeichen mit einem implizierten Befehl an die Sprachassistenz korreliert werden kann. Die Startbedingungen für ein Kommunikationsereignis können dann variabler gestaltet werden (vgl. Piernot/Binder 2019). Das Problem der räumlichen Repräsentation im rechnenden Raum bleibt allerdings insofern bestehen, als kommunikative Sequenzen als zeitliche Abfolge in die Simultaneität permutierter Bitfolge überführt werden müssen.

Solche algorithmischen Systeme arbeiten mit statistischen Korrelationen, weshalb hier die jeweiligen stochastischen Modelle relevant werden, die ihre Implementierung leiten. In diesen werden wie bei neuronalen Netzen logische Schaltungen, wie sie als *neurons* schon von McCulloch und Pitts (1943: 129ff.) beschrieben wurden, so miteinander kombiniert, dass die jeweiligen Schaltelemente je nach systemexternen oder -internen Inputs, je nach der Struktur der Verschaltung und je nach der aus den Daten gewonnenen Gewichtung des jeweiligen *neurons* konjunktiv oder disjunktiv aufeinander reagieren, also Schaltkreise schließen oder öffnen. Insofern werden Korrelationen durch Verschaltung bzw. durch eine – etwa durch Trainingsdaten bewirkte – Dynamisierung und Permutation der Verschaltung des rechnenden Raums und seiner seriellen Abfolge von differenten räumlichen Verschaltungen repräsentiert. Sie bewirken eine Defuturisierung der Produktion weiterer Kommunikationsereignisse (vgl. Ernst 2021: 29) wie auch eine Tilgung der Vergangenheit

durch die Permutation der Schaltzustände (vgl. Ernst 2018: 109) und mobilisieren eine verräumlichte Repräsentation. Die zweite Bedingung zur Autopoiesis findet sich also in diesem Moment der spezifischen Verschaltung als Strukturmerkmal bzw. als Sensitivität des Systems gegenüber seiner Historizität. Denn die Produktion des zweiten Unruhepotenzials leitet sich aus den in Bitfolgen kondensierten und derart in eine eindeutige Korrelation übersetzten vergangenen Systemereignissen ab. Als gegebener Wert, der in einer Bitfolge repräsentiert ist, muss diese Korrelation ausgelesen werden, das heißt, sie muss im Präsenzeffekt der Schaltung verfügbar gemacht werden, um verrechnet, durch Verschaltung technisch bearbeitet und in den rechnenden Raum eingeführt werden zu können. Insofern erscheint das Temporale im algorithmischen System immer als Präsenzeffekt der räumlich verfügbar gemachten Schaltzustände. Das Clustering zu Kategorien (vgl. Perrotta/Williamson 2018: 10) und die diskrete Vereindeutigung von Korrelationen (vgl. Mühlhoff 2020: 877f.) können als Verfahren aufgefasst werden, die algorithmischen Systemen zur räumlichen Übersetzung von Kommunikation zur Verfügung stehen.

Für die Beteiligung von algorithmischen Systemen an Kommunikation ist also die Übersetzung von Kommunikation in eindeutige Schaltzustände notwendig, denn nur so wird eine Adressierung möglich, wie sie auch für die Inklusion psychischer Systeme vorausgesetzt wird (vgl. Fuchs 1997: 63). Algorithmische Systeme werden über Inputs mit Signaturen versorgt, die als Startsequenz für die Konstitution von Anschlusskommunikation fungieren und damit Interfaces erst ermöglichen. Bei Interfaces handelt es sich um sensitiven Oberflächen, auf denen von Sensoren bestimmte Bewegungen erfasst und von Algorithmen auf bestimmte Signaturen ausgelesen werden, um analog zu sprachlichen Triggerphrasen eine bestimmte Datenverarbeitung zu initiieren (vgl. Karafillidis 2018: 134f.). Das Interface dient hierbei nicht nur zur Übersetzung von Signaturen in Schaltzustände, es muss auch Daten lesen, schreiben und verarbeiten, um die Signaturen selbst zur Permutation von Schaltzuständen nutzen zu können. Die Signaturen der Eingabe mobilisieren in diesem Sinne ein doppeltes Unruhepotenzial, mittels dessen die Permutation von Schaltungen erst möglich wird, durch die sodann weitere permutierbare Schaltungen verfügbar gemacht werden. Einige algorithmische Systeme stellen Kommunikation dabei unter Dauerobversation, sodass Ereignisse mit spezifischer Signatur als Triggerphrase eine Permutation des Schaltzustandes bewirken können. Das Interface stellt also nicht lediglich eine Oberfläche dar, sondern mit ihm verbinden sich Prozesse der Mobilisierung des doppelten Unruhepotenzials des Systems.

Mit dem Vorschlag von Roger Häußling (2020), Daten als Interfaces zwischen algorithmischen und sozialen Prozessen aufzufassen, statt Algorithmen und das in ihnen prozessierte Wissen als Vermittlungsoperation zwischen Datenproduktion und sozialen Prozessen zu konzipieren, wird es möglich, auf die Operationen sozialer und algorithmischer Systeme abzustellen: Das algorithmische System muss sein doppeltes Unruhepotenzial mobilisieren, um ein Interface zu konstituieren, mit dem es sich als kommunikative Adresse für soziale Systeme verfügbar macht.

5. Die Duplexstruktur der Kommunikation als Möglichkeitsbedingung autopoietischer algorithmischer Systeme

Luhmann (2002: 314) merkte vorsichtig an, dass bei der Interaktion mit Computern die Unterscheidung zwischen Mitteilung und Information nicht mehr getroffen werden könne. Für User*innen, so Luhmann, erschienen lediglich Informationen, die je nach den Selektionsbedingungen der User*innen in unterschiedlich kombinierte Spektren der Informationsverweisung führten. Der Computer als Maschine erzeugt für die User*innen eine virtuelle Kontingenz durch Manipulation der Daten (vgl. Esposito 1993: 350). Entweder – aber das ist für Luhmann noch offenzuhalten – gilt es, einen radikal neuen Kommunikationsbegriff zu entwickeln (Fuhrmann 2019: 51–79; 2020), oder der Interaktion muss abgesprochen werden, sich als Kommunikation qualifizieren zu können (vgl. Fuchs 1991).

Algorithmische Systeme, so sie als rechnender Raum aufgefasst werden, reduzieren Kommunikationsereignisse insofern auf ihre schiere Ereignishaftigkeit, als das Mitteilungsereignis für sie einzig und allein eine Differenz ist, die einen Schaltzustand triggert. Dadurch lösen sie das Problem, die Zeitlichkeit der Kommunikation in ein räumliches Prinzip zu überführen. Die These Luhmanns (1987: 196ff.), Kommunikation konstituiere sich dann, wenn eine dreifache Selektivität von Mitteilung, Information und Verstehen in einem Ereignis synthetisiere, wird also durch algorithmische Systeme herausgefordert. Denn mit dem algorithmischen System tritt ein Kommunikation konstituierendes System auf die Bühne, das nicht mehr als Ego seine Erwartung daran orientiert, das Mitteilungsverhalten Alters von dem zu unterscheiden, was Alter mitteilt und dadurch die Kapazitäten gewinnt, zu antizipieren, was Alter antizipiert, sowie im weiteren Anschließen das Verstehen Alters zu prüfen. Bei

Kommunikationsereignissen handelt es sich für algorithmische Systeme weder um Information noch um Verstehen, weil diese jeweils eines Nachtrages bedürfen, der aufgrund der Tilgung von Vergangenheit und der Defuturisierung im rechnenden Raum nicht realisiert werden kann. Die Information wird dadurch zur Information, dass sie eine Differenz zu einem vorherigen Zustand produziert. Das Verstehen avanciert im Moment des Anschlusses zum Verstehen und setzt darum zwei realisierte Systemzustände voraus (ausführlich zur Selektionstrias Fuhrmann 2019: 61–73). Lediglich das Moment der Mitteilung, das aus einer sinngenetischen Perspektive als leere Nachricht erscheint, dient dem algorithmischen System als Input, als ein Impuls, der eine Permutation im System produziert. Das Mitteilungsereignis markiert so einen Duplex, das darin besteht, sowohl als Präsenzeffekt der Permutation einer elektronischen Verschaltung als auch als ein Ereignis dienen zu können, an das eine Sinngenese angeschlossen werden kann.

Diese beiden Möglichkeiten der Verarbeitung von Kommunikation werden durch deren Duplexstruktur möglich (vgl. Fuhrmann 2020: 31f.). Das ist eine Struktur, die kommunikative Ereignisse sowohl als akute Jetzt-Zeitpunkt-Ereignisse einer Mitteilung verfügbar hält und sie damit sowohl algorithmisch auslesbar macht als auch in einen sinngenetischen Nachtrag einschreibar werden lässt. Der Nachtrag kann durch psychische Systeme vorgenommen werden, um das jeweilige Jetzt-Zeitpunkt-Ereignis in einen relationalen Kontext einzuschreiben, der eine Deutung des betreffenden Kommunikationsereignisses ermöglicht. Kommunikationsereignisse weisen also einen Präsenzeffekt auf, der ähnlich wie der Präsenzeffekt des jeweils aktuellen Zustands der elektronischen Verschaltung algorithmischer Systeme auf eine aktuelle Indikation reduzierbar ist, die unbestimmt lässt, wovon sie unterschieden ist (vgl. Fuhrmann 2019: 29f.). Die Übersetzung, also das Verfügbarmachen eines historischen Moments durch Schaltung und Gewichtung von Schaltelementen, dient der Bestätigung der korrelierten Kategorien, sodass autopoietische algorithmische Systeme in Bezug auf ihre soziale Umwelt strukturkonservativ operieren (vgl. Fuhrmann 2021). Die Annahme, dass autopoietische algorithmische System strukturkonservativ operieren, gilt auch für »Generative Adversarial Networks« (Aggarwa/lMittal/Battineni 2021), auch wenn diese generativ auftreten. Denn die von ihnen produzierten Daten bzw. Manipulationen etwa von Bild- oder Videodateien beim Deepfake, werden aus statistischer Ähnlichkeit generiert.

Zugespitzt heißt das, dass algorithmische Systeme nur deshalb Kommunikation in die Eindeutigkeit elektronischer Verschaltungen übersetzen

können, weil Kommunikation eine sinngenetisch nicht verfügbar zu machende Signatur gesellschaftlicher Asymmetrie generiert, sie also nicht entropisch, sondern negentropisch geordnet ist (vgl. Fuhrmann 2021: 111f.). Diese Signatur zeitigt sich in der differenten Häufigkeitsverteilung einzelner Zeichen im Zusammenhang zu anderen Zeichen, mithin also als Korrelation. In der Verteilung der Häufigkeit von Zeichen und Ereignissen werden dann auch sexistische, rassistische, klassistische und weitere Grammatiken der Diskriminierung algorithmisch reproduziert (vgl. Prietl 2019; Hamilton 2019; Bono/Croxson/Giles 2021; Egbert/Krasmann 2020: 907).

Literatur

Aggarwal, Alankrita, Mamta Mittal und Gopi Battineni. 2021. Generative Adversarial Networks: An Overview of Theory and Applications. *International Journal of Information Management Data Insights* 1: 1–9.

Baecker, Dirk. 2008. Über Verteilung und Funktion der Intelligenz im System. In *Wozu Systeme.* 2. Aufl., 41–66. Berlin: Kadmos.

Beckermann, Ansgar. 1988. Sprachverstehende Maschinen. Überlegungen zu John Searle's Thesen zur Künstlichen Intelligenz. *Erkenntnis* 28: 65–85.

Bono, Teresa, Karen Croxson und Adam Giles. 2021. Algorithmic Fairness in Credit Scoring. *Oxford Review of Economic Policy* 37: 585–617.

Brause, Rüdiger W. 1991. *Neuronale Netze: Eine Einführung in die Neuroinformatik*. Stuttgart: Teubner.

Broy, Manfred. 1998. *Informatik. Eine grundlegende Einführung. Programmierung und Rechnerstruktur.* Wiesbaden: Springer.

Dennett, Daniel C. 2021. Turings seltsame Umkehrung der Argumentation. Was uns Darwins Evolutionstheorie über Künstliche Intelligenz verrät. In *Künstliche Intelligenz – Die große Verheißung,* Hg. Anna Strasser, Wolfgang Sohst, Ralf Stapelfeldt und Katja Stepec, 27–36. Berlin: xenomoi.

Dupuy, Jean-Pierre. 2015. Auf dem Weg zu einer Wissenschaft der Autonomie? *Trivium* 20. https://doi.org/10.4000/trivium.5188.

Egbert, Simon und Susanne Krasmann. 2020. Predictive Policing: Not Yet, but Soon Preemptive? *Policing and Society* 30: 905–919.

Ernst, Christoph, Irina Kaldrack, Jens Schröter und Andreas Sudmann. 2019. Künstliche Intelligenzen. Einleitung in den Schwerpunkt. *Zeitschrift für Medienwissenschaften* 21: 10–19.

Ernst, Wolfgang. 2018. Zwischenarchive: eine Zeitform der Digitalen Kultur. In *Mikrozeit und Tiefenzeit,* Hg. Friedrich Balke, Bernhard Siegert und Joseph Vogl, 101–110. München: Wilhelm Fink.

Ernst, Wolfgang. 2021. Existing in Discrete States: On the Techno-Aesthetics of Algorithmic Being-in-Time. *Theory, Culture & Society* 38: 13–31.

Esposito, Elena. 1993. Der Computer als Medium und Maschine. *Zeitschrift für Soziologie* 22: 338–354.

Esposito, Elena. 2002. *Soziales Vergessen. Formen und Medien des Gedächtnisses der Gesellschaft.* Frankfurt a.M.: Suhrkamp.

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

Fissore, Luciano, Alfred Kaltenmeier, Pietro Laface, Giorgio Micca und Roberto Pieraccini. 1990. The Recognition Algorithms. In *Advanced Algorithms and Architectures of Speech Understanding,* Hg. Ciancarlo Pirani, 7–78. Berlin u.a.: Springer.

Foerster, Heinz von. 1988. Abbau und Aufbau. In Lebende Systeme. Wirklichkeitskonstruktionen in der Systemischen Therapie, Hg. Fritz B. Simon, 19–33. Berlin und Heidelberg: Springer.

Foerster, Heinz von. 1999. Über selbst-organisierende Systeme und ihre Umwelten. In Sicht und Einsicht. Versuche einer operativen Erkenntnistheorie, 115–130. Heidelberg: Carl-Auer Verlag.

Fuchs, Peter. 1991. Kommunikation mit Computern? Zur Korrektur einer Fragestellung. *Sociologia Internationalis* 29: 1–31.

Fuchs, Peter. 1997. Adressabilität als Grundbegriff der soziologischen Systemtheorie. *Soziale Systeme* 3: 57–79.

Fuchs, Peter. 2002. Die konditionierte Koproduktion von Kommunikation und Bewußtsein. In *Ver-Schiede der Kultur. Aufsätze zur Kippe kulturanthropologischen Nachdenkens,* Hg. MENSCHEN FORMEN, 150–175. Marburg: Tectum.

Fuchs, Peter. 2015. *Der Sinn der Beobachtung. Begriffliche Untersuchungen.* Weilerswist: Velbrück.

Fuhrmann, Jan Tobias. 2019. *Postfundamentale Systemtheorie.* Wien: Passagen.

Fuhrmann, Jan Tobias. 2020. Wechselseitige Disziplinierung. Zum systemtheoretischen Verständnis von Kommunikation unter Beteiligung psychischer und algorithmischer Systeme. In *Algorithmisierung und Autonomie im Diskurs. Perspektiven und Reflexionen auf die Logik automatisierter Maschinen,* Hg. Christian Leineweber und Claudia de Witt, 16–46. Hagen: FernUniversität Hagen.

Fuhrmann, Jan Tobias. 2021. Strukturkonservative Algorithmen: Künstliche Intelligenz als Kommunikationsproblem. In *Künstliche Intelligenz – Die große Verheißung*, Hg. Anna Strasser, Wolfgang Sohst, Ralf Stapelfeldt und Katja Stepec, 103–128. Berlin: xenomoi.

Hamilton, Melissa. 2019. The Sexist Algorithm. *Behavioral Sciences & The Law* 37: 145–157.

Häußling, Roger. 2020. Daten als Schnittstellen zwischen algorithmischen und sozialen Prozessen. Konzeptuelle Überlegungen zu einer relationalen Techniksoziologie der Datafizierung in der digitalen Sphäre. In *Soziologie des Digitalen – Digitale Soziologie? Soziale Welt, Sonderband 23*, Hg. Sabine Maasen und Jan-Hendrik Passoth, 134–150. Baden-Baden: Nomos.

Helmbrecht, Johannes. 2004. *Selbstbewußtsein und Selbstreferenz. ICH in der Grammatik der Sprachen der Welt.* Erfurt: Seminar für Sprachwissenschaften der Universität Erfurt.

Hildebrandt, Mireille. 2011. Who Needs Stories if You Can Get the Data? ISPs in the Era of Big Number Crunching. *Philosophy & Technology* 24: 371–390.

Kaminski, Andreas. 2020. Gründe geben. Maschinelles Lernen als Problem der Moralfähigkeit von Entscheidungen. In *Datafizierung und Big Data. Ethische, anthropologische und wissenschaftstheoretische Perspektiven*, Hg. Klaus Wiegerling, Michael Nerukar und Christian Wadephul, 151–174. Wiesbaden: Springer VS.

Karafillidis, Athanasios. 2018. Die Komplexität von Interfaces. Touchscreen, nationale Identität und eine Analytik der Grenzziehung. *Berliner Debatte Initial* 29: 130–146.

Kieserling, André. 1999. *Kommunikation unter Anwesenden. Studien über Interaktionssysteme.* Frankfurt a.M.: Suhrkamp.

Koster, Ann-Kathrin. 2022. Das Ende des Politischen? Demokratische Politik und Künstliche Intelligenz. *Zeitschrift für Politikwissenschaft* 32: 573–594.

Lang, Stefan. 2020. *Performatives Selbstbewusstsein.* Paderborn: mentis.

Lemoine, Blake. 2022. Is LaMDA Sentient? – an Interview. *Medium*. https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917. Zugegriffen: 20. September 2022.

Luhmann, Niklas. 1989. Individuum, Individualität, Individualismus. In *Gesellschaftsstruktur und Semantik. Studien zur Wissenssoziologie*. Bd. 3, 149–258. Frankfurt a.M.: Suhrkamp.

Luhmann, Niklas. 1998. *Die Gesellschaft der Gesellschaft.* Frankfurt a.M.: Suhrkamp.

Luhmann, Niklas. 2002. *Einführung in die Systemtheorie.* Heidelberg: Carl-Auer Verlag.

Luhmann, Niklas. 2008a. Die operative Geschlossenheit psychischer und sozialer Systeme. In *Soziologische Aufklärung 6. Die Soziologie und der Mensch*. 3. Aufl., 26–37. Wiesbaden: VS Verlag für Sozialwissenschaften.

Luhmann, Niklas. 2008b. Probleme mit operativer Schließung. In *Soziologische Aufklärung 6. Die Soziologie und der Mensch*. 3. Aufl., 13–26. Wiesbaden: VS Verlag für Sozialwissenschaften.

Luhmann, Niklas. 2017. Die Kontrolle von Intransparenz. In *Die Kontrolle von Intransparenz*, 96–120. Berlin: Suhrkamp.

Maskarinec, Malika. 2010. Das Begehren der Philologie nach räumlichen Beziehungen. In *Möglichkeiten und Grenzen der Philologie,* Hg. Jens Elze, Zuzanna Jakubowski, Lore Knapp, Stefanie Orphal und Heidrun Schnitzler, 79–88. Berlin: FU Berlin.

McCulloch, Warren und Walter Pitts. 1943. A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics* 5: 115–133.

Miyazaki, Shintaro. 2017. Algorhytmische Ökosysteme. Neoliberale Kopplungen und ihre Pathogenese von 1960 bis heute. In *Algorithmuskulturen. Über die rechnerische Konstruktion der Wirklichkeit,* Hg. Robert Seyfert und Jonathan Roberge, 173–187. Bielefeld: transcript.

Mühlhoff, Rainer. 2020. Automatisierte Ungleichheit. Ethik der Künstlichen Intelligenz in der biopolitischen Wende des Digitalen Kapitalismus. *Deutsche Zeitschrift für Philosophie* 68: 867–890.

Nassehi, Armin. 2019. *Muster. Theorie der digitalen Gesellschaft*. München: C. H. Beck.

Oberschelp, Walter und Gottfried Vossen. 2006. *Rechneraufbau und Rechnerstrukturen*. München und Wien: Oldenbourg.

Pasquinelli, Matteo. 2017. Machines that Morph Logic: Neural Networks and the Distorted Automation of Intelligence as Statistical Inference. *Glass Bead* 1: 1–17.

Patel, Gayatri und Kajal Patil. 2019. My Buddy App: Communications between Smart Devices through Voice Assist. *International Research Journal of Engineering and Technology* 6: 2138–2155.

Perrotta, Carlo und Ben Williamsons. 2018. The Social Life of Learning Analytics: Cluster Analysis and the ›Performance‹ of Algorithmic Education. *Learning, Media and Technology* 43: 3–16.

Pias, Claus. 2009. Time of Non-Reality. Miszellen zum Thema Zeit und Auflösung. In *Zeitkritische Medien,* Hg. Axel Volmar, 267–279. Berlin: Kadmos.

Piernot, Philippe und Justin Binder. 2019. *Reducing the Need for Manual Start/End-Pointing and Trigger Phrases*. https://patentimages.storage.googleapis.com/0e/07/f6/9500b529e83493/US10373617.pdf. Zugegriffen: 4. Januar 2020.

Prietl, Bianca. 2019. Algorithmische Entscheidungssysteme revisited: Wie Maschinen gesellschaftliche Herrschaftsverhältnisse reproduzieren können. *Feministische Studien* 37: 303–319.

Rojas, Raúl. 1993. *Theorie der neuronalen Netze. Eine systematische Einführung*. Berlin u.a.: Springer.

Scha, Remko, Rems Bod und Khalil Simo'an. 1999. A Memory-Based Model of Syntactic Analysis: Data-Oriented Parsing. *Journal of Experimental and Theoretical Artificial Intelligence* 11: 409–440.

Searle, John. 1980. Minds, Brains, and Programs. *The Behavioral and Brain Sciences* 3: 417–457.

Sheng, Victor S. und Jing Zhang. 2019. Machine Learning with Crowdsourcing: A Brief Summary of the Past Research and Future Directions. *Proceedings of the AAAI Conference on Artifical Intelligence* 33: 9837–9843.

Spencer-Brown, George. 1999. *Laws of Form. Gesetze der Form*. Lübeck: Bohmeier.

Steiglitz, Ken. 2019. *The Discrete Charm of the Machine.* Princeton u.a.: Princeton University Press.

Stokes, Jon. 2007. *Inside the Machine*. San Francisco: No Starch Press.

Storp, Michaela. 2002. Chatbots. Möglichkeiten und Grenzen maschineller Verarbeitung natürlicher Sprache. Hannover: Gottfried Wilhelm Leibniz Universität. https://doi.org/10.15488/2916.

Teubner, Gunther. 2018. Digitale Rechtssubjekte? Zum privatrechtlichen Status autonomer Softwareagenten. Archiv für die civilistische Praxis 218: 155–205.

Thoppilan, Romal et al. 2022. LaMDA – Language Models for Dialog Applications. https://doi.org/10.48550/arXiv.2201.08239.

Wansleben, Leon. 2012. Heterarchien, Codes und Kalküle. Beitrag zu einer Soziologie des algo trading. *Soziale Systeme* 18: 225–259.

Weizenbaum, Joseph. 1966. ELIZA – A Computer Program for the Study of Natural Language Communication Between Man and Machine. *Communications of the ACM* 9: 36–45.

Wiener, Oswald. 1990. *Probleme der Künstlichen Intelligenz.* Berlin: Merve.

Zeleny, Milan. 1976. Self-Organization of Living Systems: A Formal Model of Autopoiesis. *Journal of General System* 4: 13–28.

Zuse, Konrad. 1969. *Rechnender Raum*. Braunschweig: Friedrich Vieweg.

Zweig, Katharina. 2018. *Wo Maschinen irren können. Fehlerquellen und Verantwortlichkeiten in Prozessen algorithmischer Entscheidungsfindung*. Gütersloh: Bertelsmann Stiftung.

Robotermaterial und ›Künstliche Intelligenz‹

Posthumanistische Potenziale der Robotik

Hannah Link

Abstract: *Der Beitrag fragt nach den Beschreibungsmöglichkeiten neuerer KI-Ansätze der Robotik. Auf Basis empirischen Datenmaterials argumentiert der Aufsatz, dass veränderte, dezentrierte Vorstellungen über ›den Menschen‹ wegweisend für die Technikentwicklung jüngerer Robotikzweige sind. In Form ethnografischer Analysen wird gezeigt, (1) wie eine konkrete technowissenschaftliche Praxis des Roboterkonstruierens ausgeübt wird, (2) wie diese Praxis als posthumane Interaktion zwischen Wissensobjekt und -subjekt hervortritt und (3) wie Robotiker*innen sich tendenziell einer situierten Praxis der Wissensproduktion annähern. Formuliert wird der Ansatz einer Soziologie technowissenschaftlicher Praxis, mit dem Ziel, die posthumanistischen Potenziale der Robotik für die Soziologie zugänglich zu machen.*

1. Einleitung

In Karel Čapeks Theaterstück *Rossums Universal Robots* prophezeit der junge Fabrikdirektor Domin den Aufbruch in ein neues Zeitalter der Maschinenarbeit; ein Zeitalter, das mit der lang ersehnten Ausrottung menschlicher Unproduktivität beginnt:

> Ein Mensch, der ist etwas, das – sagen wir – Freude fühlt, Geige spielt, spazieren gehen will und überhaupt einen Haufen Sachen zu tun braucht, welche eigentlich überflüssig sind […]. Welche überflüssig sind, wenn er weben oder addieren soll. Aber eine Arbeitsmaschine muss nicht Violine spielen, muss nicht Freude fühlen, muss nicht einen Haufen anderer Dinge tun. Die Erzeugung soll möglichst einfach und das Erzeugnis das praktisch Beste sein. (Čapek 2017: 12)

Anfang der 1920er Jahre brachte Čapek damit erstmals die Idee der Synthetisierung ökonomisch brauchbarer menschlicher Eigenschaften auf den Begriff des Roboters (aus dem tschechischen Nomen *robota*: mühsame Arbeit, Frondienst oder Knechtschaft). Er beschrieb damit eine verzerrte Kopie des Menschen, die ihm einerseits ähnlich sieht, dabei aber andererseits ein Vielfaches seiner körperlichen Kraft besitzt und nur einen Bruchteil seiner geistigen und emotionalen Tiefe: eine bedürfnislose Maschine als vollendete Arbeitskraft. Das Stück kulminiert in einem blutigen Aufstand der nach Freiheit und Autonomie trachtenden humanoiden Sklaven. Hat man zu Beginn noch den Eindruck, die Maschine sei dienendes Objekt eines gebieterischen Subjekts, wird im weiteren Verlauf eine allmähliche Destabilisierung der hierarchischen Beziehung zwischen Mensch und Maschine deutlich. Auf diese Weise ließ Čapek ihre Grenze brüchig werden und begründete die Ideengeschichte des modernen Roboters (vgl. Richardson 2016).

Etwa hundert Jahre später hat die Frage nach der technischen Reproduzierbarkeit menschlicher Eigenschaften weiterhin Konjunktur. Roboter werden zunehmend als humanoide Wesen konstruiert, die sich in gesellschaftliche Zusammenhänge integrieren sollen: als Sozialpartner auf der einen und als Arbeitskraft auf der anderen Seite. Vor diesem Hintergrund verweisen rezente Diskurse der Technikforschung auf die Rekomposition der menschlichen Selbstwahrnehmung (Turkle 1984) und prophezeien die Entstehung neuer Sozialformen (Alač 2009). Hieran anknüpfend zeigen empirische und theoretische Studien zum einen das Gelingen von Mensch-Roboter-Interaktionen (etwa Pitsch 2016; Meister/Schulz-Schaeffer 2016; Hergesell et al. 2021) oder fragen zum anderen nach ihren Bedingungen: Wie etwa ko-konstruiert die Robotik eine techno- und roboteraffine Gesellschaft (Šabanović 2014; Suchman 2007; Lipp 2022)? Welche Rolle spielen in diesem Kontext Konzepte von Verantwortung, Moral und Ethik (etwa Coeckelbergh 2020)? In diesem Zusammenhang werden mitunter Forderungen nach politischer Regulation vermeintlich autonomer Roboterwesen laut (Beck 2012).

Beide, hoffnungsvolle und kritische, Deutungen kreisen um die Frage nach dem Verhältnis des Menschen zu intelligenten Maschinen. Diese, so der Eindruck von Barbara Becker und Jutta Weber (2005: 14), würden »zu einem zentralen Referenzpunkt in unserer Auseinandersetzung mit der conditio humana«. Intelligente Maschinen dienen auf der einen Seite als Spiegelung des Humanen: Es ist das technowissenschaftliche Bestreben, die menschliche Verfasstheit durch ihre technische Reproduktion zu erschließen. Auf der anderen Seite wird das Erkenntnisinteresse geradezu invertiert: Das Humane dient als

Vehikel der Entwicklung technischer Lösungen. In diesen, zum Teil auch populären Diskursen zirkulieren Begriffe wie *Künstliche Intelligenz*, *deep learning*, *machine vision*, *artificial emotions* etc., die Menschenähnlichkeit indizieren. Die Frage, ob diese als menschlich verstandenen Eigenschaften und Leistungen auch tatsächlich durch die Technik realisiert werden, klammere ich im Folgenden aus. Ich wende mich stattdessen dem Phänomen zu, dass verschiedene Bezeichnungen von Maschinen, mit denen ihnen spezifische Eigenschaften zugeschrieben werden, emische und zugleich kontroverse Begriffe des Feldes der Robotik sind. In diesem Kontext bezeichnet die sogenannte ›Künstliche Intelligenz‹ umstrittene Annahmen des Feldes über ›gelungene‹ Kognition.

Basierend auf der Beobachtung, dass die Robotik von Annahmen über das Menschliche inspiriert wird, braucht es eine Deutung, die das Verhältnis von Technologie und Menschenbild ins Zentrum rückt. Interessant ist dabei, dass gerade neuere Entwicklungszweige der Robotik ›den Menschen‹ nicht mehr kognitivistisch als isolierte, souverän agierende Einheit und somit als rationales Subjekt denken. Vielmehr findet sich in diesen Ansätzen tendenziell ein dezentriertes Menschenbild, das auf Körperlichkeit und Umweltkopplung abstellt. Für die Beschreibung neuerer KI-Konzepte in der Robotik liegt es daher nahe, insbesondere posthumanistische Konzepte heranzuziehen, die ein Vokabular für veränderte Menschenbilder bereithalten. Der Begriff Posthumanismus beschreibt in diesem Zusammenhang unterschiedliche analytische und ethische Projekte, die im Zuge gesellschaftlich-technologischer Veränderungen auf eine Rekonzeptualisierung der Figur des Menschen abzielen.

Der Beitrag argumentiert, dass insbesondere posthumanistische Einsichten von Donna Haraway für die wissenssoziologische Erforschung der Robotik übernommen werden können. Eine posthumanistisch informierte Wissenssoziologie, die sich einer Exemplifizierung der Laborpraxis der Robotik zuwendet, vermag es, das soziomaterielle Wissen der Teilnehmer*innen in den Blick zu nehmen, ohne eine Hierarchisierung von (Erkenntnis-)Subjekt und (Untersuchungs-)Objekt voraussetzen zu müssen. Aus dieser Perspektive heraus fragt der Beitrag, welches soziomaterielle Wissen in die Bearbeitung von Robotern eingeht; es wird fokussiert, welche Vorstellungen, Annahmen und Theorien über ›natürliche‹ bzw. ›menschliche‹ und ›Künstliche Intelligenz‹ in der Arbeit am Robotermaterial sichtbar werden und wie das Material selbst wirksam wird. Vor diesem Hintergrund wird deutlich werden, dass in neueren Zweigen der Robotik Konzepte von Intelligenz vom Kopf auf die Füße gestellt werden: Es wird darauf gezielt, eine verteilte Korpus-›Intelligenz‹ zu entwer-

fen, mit der die Vorstellung zurückgedrängt wird, der Roboterkorpus, seine materielle Komposition, sei ein Derivat eines softwaretechnischen ›Geistes‹. Die Robotermaterie selbst wird als intelligent und wirkmächtig konzeptualisiert (vgl. Link/Kalthoff 2023).

Ziel des Beitrags ist es, die Black Box der Roboterkonstruktion zu öffnen und die Robotik in ihren praktischen und posthumanen Dimensionen zugänglich und anschlussfähig zu machen. Durch einen solchen empirischen Blick auf die Praxis kann der Enttäuschung über technologisch unverdaute Humanismen in der Robotik (vgl. Weber 2003) ein posthumanistisches Potenzial entgegengestellt werden. Insofern stellen die folgenden Überlegungen den Versuch dar, einige Positionen eines leidenschaftlich geführten Diskurses aufzugreifen und mit der Praxis der Robotik zu konfrontieren.

Hierzu werde ich zunächst wissenssoziologische und posthumanistische Theorieangebote miteinander ins Gespräch bringen und eine posthumanistisch imprägnierte Perspektive entwickeln (2). Ich werde deutlich machen, dass eine posthumanistische Perspektive den persistenten humanistischen Impuls der Wissenssoziologie zu konterkarieren vermag. Der darauffolgende Abschnitt setzt sich mit der Praxis der Robotik auseinander (3). Hierzu werde ich zunächst anhand von Interviewdaten dokumentieren, wie sich grundlegende Annahmen und Paradigmen der Robotik verändert haben (3.1). Es wird deutlich werden, dass in neueren Entwicklungszweigen Materialität eine konstitutive Rolle bei der Roboterkonstruktion zugesprochen wird. Anhand von ethnografischen Protokollen werde ich anschließend darlegen, wie sich Robotiker*innen und Roboter in konflikthaften Relationen ko-konstituieren und wie sich die technowissenschaftliche Praxis als engagiertes Zusammenspiel mit dem Materiellen zeigt (3.2). Ich werde argumentieren, dass ein solches Zusammenspiel als Annäherung an eine »situierte« (Haraway 1995b) Wissenschaftspraxis verstanden werden kann. Daran anschließend fasse ich die zentralen Argumente und Ergebnisse des Textes zusammen und diskutiere abschließend die posthumanistischen Potenziale neuerer Robotikzweige (4).

Der Beitrag basiert auf einem soziologisch-ethnografischen Dissertationsprojekt, das dem Forschungsparadigma der »theoretischen Empirie« (Kalthoff 2018) folgt. Empirisch basiert der Beitrag auf einem heterogenen Datenkorpus aus In-situ-Beobachtungen technowissenschaftlicher Laborarbeit und Leitfaden- sowie ethnografischen Interviews mit Robotiker*innen. Die erhobenen Daten wurden mit dem Kodierverfahren der Grounded Theory (Glaser/Strauss 1967) analysiert. Das empirische Material wurde in Zusammenarbeit mit Herbert Kalthoff im Rahmen des Teilprojektes »Maschinelle

Humandifferenzierung. Ethnosoziologien der Robotik« (SFB 1482) an der Johannes Gutenberg-Universität Mainz erhoben.

2. Wissenssoziologie und Posthumanismus

Der Beitrag fragt, wie Roboter zu ihrer Form kommen, welches Wissen und welche Wissensaktivitäten in der Roboterkonstruktion sichtbar werden und wie sich dies in die kognitive wie materielle Architektur einschreibt. Welche Annahmen bzw. Vorstellungen über und impliziten Bezüge auf ›natürliche‹ und ›künstliche Intelligenz‹ werden in Roboter eingehegt? Umgekehrt interessiert sich der Beitrag aber auch dafür, wie Roboter bei ihrer Fertigung selbst materiell wirksam werden; wie Robotiker*innen durch ihr Material affiziert und rekonfiguriert werden: Wie wird Wissen durch die Objekte (mit)gestaltet? Aus einer praxistheoretischen Perspektive liegt der Fokus des Beitrags auf den körperlich-materiell gebundenen Abläufen im Labor. Die praxistheoretische Umorientierung vom rationalen Zentrum menschlicher Handlungen zu »kulturell geprägten *ways of doing*« (Hirschauer 2016: 46; Hervorh. im Orig.) erlaubt es, Handlungen zu dezentrieren, also auch vermeintlich stumme Elemente einer Praxis als Teil dieser Handlung in die wissenssoziologische Analyse einzubinden. So geraten neben dem expliziten Wissen und den Motivlagen der Teilnehmer*innen auch Körper und Artefakte als Speicher und Ko-Produzenten von Wissen in den analytischen Fokus. Schon in den frühen ethnografischen Laborstudien wurde die konstitutive Rolle, die materiellen Objekten des Labors bei der Erzeugung wissenschaftlichen Wissens zukommt, in den Blick genommen. So hat etwa Bruno Latour deutlich gemacht, dass die im Labor gehandhabten technischen Objekte auf die menschlichen Wissenschaftler*innen zurückwirken, indem sie den Prozess der Wissensgenese systematisieren und verstetigen (vgl. Latour 1993). Zunehmend geraten seither Objekte als konstitutive Elemente menschlicher Praxis in den Fokus. Obwohl es nun gerade durch die Etablierung der Praxistheorien gelang, auch nichtmenschliche Entitäten als konstitutive Bestandteile techno- und naturwissenschaftlicher Arbeit zu konzeptualisieren, scheinen sich diese Entitäten nicht zuletzt auf die eine oder andere Weise in der analytischen Peripherie zu befinden. So tendieren Praxistheorien dazu, materielle Objekte entweder in ihrem artifiziellen oder in ihrem ermöglichenden Charakter zu besprechen: entweder als materialisiertes Wissen und geronnener Sinn oder als Rohstoff der Praxis, der diese präformiert und festigt. So zeugen zahlreiche

Untersuchungen von dem Interesse daran, wie menschliche Handlungsziele durch nichtmenschliche Entitäten mediatisiert (Schatzki 2002), präfiguriert (Bourdieu 1980) und stabilisiert (Latour 1993) werden[1]. Für die Integration nichtmenschlicher Entitäten in eine Soziologie wissenschaftlichen Wissens sind diese praxistheoretischen Ansätze unabdingbar, sie leisten jedoch einer immer schon gesetzten ontologischen Asymmetrie zwischen einem dynamischen Subjekt und einem passiven Objekt Vorschub. Fokussiert wird eine Praxis *durch* Objekte, nicht aber eine Praxis *der* Objekte. Der Beitrag argumentiert nun, dass die Integration einer kritisch-posthumanistischen Perspektive dieser persistenten Passivierung materieller Objekte entgegenwirken kann.

Unter dem Begriff Posthumanismus werden unterschiedliche wissenschaftliche Projekte vereint, die seit dem Entstehen der Kybernetik systematisch zur Dekonstruktion der Figur des Menschen beitragen. Innerhalb dieser heterogenen Forschungslandschaft lassen sich grob drei Perspektiven unterscheiden: eine technologische, eine sozialtheoretische und eine kritische. Während technologische Projekte daran interessiert sind, neue Wege einer technischen Erweiterung oder gar Überwindung des Menschen zu erkunden, ist die zweite Perspektive an der Beschreibung sozialer Phänomene interessiert, die nicht zwangsläufig auf den Menschen zurückzuführen sind. Demgegenüber steht jene im Folgenden fokussierte – dritte Perspektive, der in erster Linie daran gelegen ist, das humanistische Subjekt der Aufklärung, das immer auch an konstitutive Ausschlüsse des Anderen (Natur, Frauen, Kolonisierte und Objekte) geknüpft ist, zu hinterfragen (vgl. Haraway 1995a, 1995b; Braidotti 2014). Auf diese Weise räumen sie nun jenen Anderen eine produktive Kraft ein, die mit Beginn der europäischen Neuzeit als dem Menschlichen untergeordnet verstanden wurden. Innerhalb dieses Diskurses nimmt Donna Haraways feministische Wissenschaftskritik eine Schlüsselstellung ein. Insbesondere in ihrem 1988 veröffentlichten Aufsatz *Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective* argumentiert sie, dass der wissenschaftliche Anspruch neutraler, objektiver Forschung zentrale weltliche Verstrickungen verkenne, die eben nicht durch

1 Inwiefern Bruno Latours Arbeiten einem praxistheoretischen Forschungsprogramm zugeordnet werden können ist umstritten (vgl. Reckwitz 2002, Schatzki 2002). Eine zentrale Differenz scheint sich mitunter im Umgang mit ontologischen Fragestellungen abzuzeichnen: Während Latour eine methodologische Symmetrisierung mit einer ontologischen Symmetrisierung verbindet, lehnen genuin praxistheoretische Argumentationen ontologische Fragestellungen ab (vgl. Wieser 2015).

die eigentümliche Trennung zwischen aktivem (Erkenntnis-)Subjekt und passivem (Forschungs-)Objekt darstellbar werden. Hiermit angesprochen ist die Kritik an der Konzeptualisierung von Wissensobjekten als vorliegend und vollständig beschreibbar. Vor diesem Hintergrund votiert Haraway (1988) mit dem Konzept des situierten Wissens für eine radikale Reflexion dieser Verstrickungen und der gesellschaftlichen Positionen, von denen aus geforscht wird: Erst das Aufgeben des Ideals distanzierter Objektivität – das nur bestimmten Körpern zugestanden wird[2] – erlaubt es uns, von einer artifiziellen Logik des Entdeckens zu einer Konversation mit der Welt überzugehen (vgl. Haraway 1995b: 84).[3]

Die Provokation von Haraways Wissenschaftskritik besteht nun darin, dass sie die Souveränität von Wissenschaftspersonen relativiert, indem sie an eine wissenschaftliche Verkörperung und Einbettung in eine dynamische materielle Welt erinnert. Gerade Haraways Insistieren auf eine konstitutive Situiertheit von Erkenntnissubjekten birgt folgenreiche posthumanistische Implikationen ihrer feministischen Wissenschaftskritik. So denkt sie materielle Wissensobjekte (wie etwa Körper und Organismen) von ihren selbstorganisierenden, »ontogenetischen« (Folkers 2013: 25) Potenzialen her, die eine vollständige, objektive Repräsentation durch ein Erkenntnissubjekt unmöglich machen. Materielle Objekte liegen nicht passiv vor, sondern sind dynamisch, mitunter destruktiv und für ihre Erforschung nie gänzlich verfügbar. Haraway fordert uns – anders formuliert – dazu auf, das »Subjekt-Werden« (Keller 1983: 118) des Wissensobjekts zu riskieren: »Situiertes Wissen erfordert, dass das Wissensobjekt als Akteur und Agent vorgestellt wird und nicht als Leinwand oder Grundlage oder Ressource […].« (Haraway 1995b: 93) Entitäten, die gängigerweise auf der passiven Seite modernistischer Gegensatzpaare vermutet werden, begreift sie als wirkmächtige, widersinnige und an ihrer eigenen Entdeckung beteiligte Agenten. So gesehen überdauern und stabilisieren materielle Objekte nicht bloß soziale Situationen, sondern wirken auf sie ein und werden zugleich durch sie rekonfiguriert; sie sind

2 Haraway (1995b: 82) betont, dass der »göttliche Trick« als entkörperter Blick auf die Welt sowohl an rassifizierte als auch vergeschlechtlichte Ausschlüsse gebunden ist.

3 In neueren Publikationen spitzt sie ihre Einsichten über die Eingebettetheit von Wissenschaftler*innen deutlich zu: Mit der Ontologie der Gefährt*innenspezies (Haraway 2016) konstatiert sie ein radikales Miteinander-Werden von Entitäten. Demnach sind menschliche und nichtmenschliche Entitäten nicht als isolierte Einheiten, sondern als sympoietische Komplexe zu verstehen, die sich in wechselseitiger Abhängigkeit ko-konstituieren (vgl. Hoppe 2019).

mithin dynamisch und entziehen sich notwendigerweise einer endgültigen wissenschaftlichen Festschreibung.[4]

Auf diese Weise hat Haraway etwas gefunden, das nicht allein durch gesellschaftliche und kulturelle Kontexte erklärbar ist, sondern in einer gewissen Eigenlogik existiert und sich somit dem konventionellen sozial- und geisteswissenschaftlichen Zugriff versperrt.

In der Auseinandersetzung mit kritisch-posthumanistischen Autor*innen kann die Soziologie begründen, dass (materielle) Wissensobjekte zu wesentlichen Anteilen an ihrer eigenen Handhabung und Entdeckung beteiligt sind. Ich plädiere nun dafür, Wissensbestände der kritischen Posthumanismen als geschärfte Optik aufzunehmen. Das »konzeptionelle Instrumentarium« (Hoppe/Lemke 2015: 274) des Posthumanismus soll es erstens ermöglichen, Geschehnisse der Praxis in ihrer spannungsreichen Vorläufigkeit und Relationalität zu begreifen. Zweitens soll es den Blick für Material, Objekte etc. schärfen, ohne diese als stabilisierendes Moment zu passivieren. Es wird darum gehen, am Fall der Robotik zu zeigen, wie die Sozio-Logik einer technowissenschaftlichen Praxis durch eine posthumanistische Linse neu gedacht werden kann.

3. Die Praxis der Robotik

Anhand von empirischem Material sollen im Folgenden zunächst Ethnotheorien, Annahmen und Entwicklungslinien unterschiedlicher KI-Systeme der Robotik skizziert werden (3.1). Diese werden dann in einem zweiten Schritt durch die Rekonstruktion konkreter Abläufe und Situationen der Roboterkonstruktion ergänzt. Dabei soll stärker auf Mikroprozesse soziomaterieller Aushandlungen fokussiert werden, in denen Robotermaterie als machtvolle Mit- und Gegenspielerin auftritt (3.2).

4 Obwohl sich rezente Diskurse (etwa Barad 2012; Bennett 2010) der sogenannten Neomaterialismen einer »founding gesture« (Ahmed 2008) des komplett Neuen bedienen, sind sie von der posthumanistischen Geste der feministischen Wissenschaftskritik Haraways maßgeblich informiert (vgl. Hoppe 2019).

3.1 Vom Symbolischen zum Materiellen: Neuere Ansätze der Robotik

Die Robotik ist – wie viele andere Disziplinen auch – durch eine gewisse Uneinheitlichkeit gekennzeichnet: Es gibt nicht nur die eine Robotik, sondern viele Robotiken, die unterschiedlichen Entwicklungspfaden folgen, miteinander konkurrieren und sich verschiedenen Aufgaben annehmen. Ein Ziel der Robotik ist die Fertigung von Maschinen, die auch außerhalb des Labors in relativ unstrukturierten Umgebungen funktionieren können: »Systeme für die echte Welt bauen«, so ein Robotiker. Ein anderes Ziel ist, den menschlichen Alltag – die »echte Welt« – in das Labor zu überführen und dort technisch-experimentell zu erkunden. Beide Forschungsausrichtungen stehen dabei vor der zentralen Aufgabe, die rasante Dynamik und Unsicherheit des menschlichen Alltags (beispielsweise implizite oder uneindeutige Kommunikationsweisen, abrupt wechselnde Situationen) auf die eine oder andere Weise technisch in den Griff zu bekommen.

Klassische Ansätze der Robotik (etwa Newell/Simon 1997) begegnen diesen Problemen mit Symbolverarbeitungen. Die Welt, so die Vorstellung, hat ein symbolisches Korrelat im Geiste des Menschen. Die Aufgabe der Robotiker*innen bestehe entsprechend darin, unabhängig verstandene Merkmale der Welt zu entdecken, mathematisch zu erschließen und symbolisch zu rekonstruieren. Innerhalb dieses Symbolsystems soll die Software Zusammenhänge berechnen und Steuerungsinformationen an die entsprechenden Korpusteile senden, etwa zur Bewegung der Hand. Die symbolische Robotik arbeitet zu einem wesentlichen Teil daran, technisch alle denkbaren Zustände der Umwelt und des Roboters symbolisch zu implementieren, bevor Roboter in ihrer Umwelt operieren, mit dem Ziel, dass sich Roboter anschließend in ihren intern symbolisch abgebildeten Umwelten bewegen können. In der symbolischen Robotik werden rechnerische und abstrakte Verfahren den materiellen Dimensionen übergeordnet: ›Künstliche Intelligenz‹ soll durch Berechnungen der Software entstehen. KI kann – so die Vorstellung – demnach auf der Ebene von Algorithmen und Rechenprozessen untersucht werden, ohne dafür eine materielle Struktur, das heißt den Roboterkorpus, einbeziehen zu müssen. Die Frage nach materiellen Eingriffsmöglichkeiten in die Umwelt wird somit auf einen peripheren Kanal reduziert.

Im Gegensatz dazu stehen neuere Ansätze der Robotik, auf die sich der Fokus des Beitrages richten wird: Dort ist nicht per se die Software zentraler Gegenstand der Forschung, sondern die materielle und technische Beschaffen-

heit des Korpus. Die sogenannte emergente Robotik[5] reagiert damit auf fundamentale Probleme klassischer KI-Konzepte, mit denen die symbolische Robotik bis dato hantiert (vgl. Kalthoff/Link 2021). So stellen Kritiker*innen etwa Mitte der 1980er Jahre heraus, dass diese Systeme zwar in Bereichen der abstrakten Symbolverarbeitung (etwa *reasoning*, *problem solving*) gut funktionieren, aber in konkreten Alltagsanwendungen, die materielle und korporale Elemente zwangsläufig einschließen, einfachste Standards unterschreiten: Das Greifen eines Glases oder das Überschreiten einer Türschwelle entwickelt sich zu einem beinahe unlösbaren Problem. Bekannt ist diese Kritik unter dem Schlagwort *real world problems*, das bisweilen zwei Probleme der Roboterentwicklung adressiert: Robustheit auf der einen und realzeitliche Prozessierung auf der anderen Seite (vgl. Pfeifer/Scheier 1999: 64). Geringfügige ungeplante Veränderungen der Umwelt sollten also erstens nicht zum Absturz des Systems führen. Zweitens sieht das KI-Konzept der symbolischen Robotik eine »räsonierende Zentralinstanz« (Kalthoff/Link 2021: 319) vor, die jegliche Daten bearbeitet und anschließend an die einzelnen Korpusteile verschickt. Spontane Reaktionen auf die Umwelt können also kaum operationalisiert werden. Bereiche wie Objektmanipulation, Navigation und Raumorientierung stecken daher seit den 1950er Jahren in den Kinderschuhen.

Angesichts derartiger Beschränkungen fordert eine Reihe von KI-Forscher*innen (etwa Steels/Brooks 1995; Pfeifer 2000) einen Richtungswechsel, der die ›echte‹ Welt, also die Alltagswelt des Menschen, und dessen körperlich-leibliche Verfasstheit als integralen Bestandteil von Intelligenz ernst nimmt. Aber nicht nur forschungspraktische Probleme treiben die neueren Entwicklungen an, Robotiker*innen reagieren auch auf die sozialwissenschaftliche und philosophische Technikkritik (u.a. Hubert Dreyfus, Barbara Becker oder John Searle). So problematisierte etwa Hubert Dreyfus (1985) die strikte Trennung von Körper und Geist in der symbolischen Robotik und KI-Forschung. Kritisch Bezug nehmend auf Dreyfus fordern auch Vertreter*innen neuerer Ansätze der Robotikforschung eine konsequente Verkörperung von Intelligenz:

> I think he [Dreyfus; Anm. H. L.] was right about many issues, such as the way in which people operate in the world is intimately coupled to the ex-

5 Die emergente Robotik umfasst Projekte wie *soft robotics*, *verhaltensbasierte Robotik*, *evolutionary robotics*, *embodied AI*.

> istence of their body in that world. Unfortunately, he made claims about what machines could not do in principle. (Brooks 2002: 168)

Unterscheiden will sich die emergente von der symbolischen Robotik in zwei Aspekten: Sie verfolgt erstens einen tendenziell holistischen Ansatz, der Körper und die materielle Umgebung in den Konstruktionsprozess bewusst integriert. Sie distanziert sich von der klassischen Orientierung an der formalen Logik und Mathematik und wendet sich zunehmend kybernetischen und evolutionsbiologischen Konzepten zu. Zweitens wird davon ausgegangen, dass intelligente Verhaltensmuster nicht schon immer im Menschen vorrätig vorhanden sind, sondern erst durch körperlich-materielle Interaktionen mit der Umwelt emergieren. Ziel ist es, dass neues Verhalten entsteht, das über die Basisprogrammierung hinausgeht. Verhalten wird hierfür entlang eines Bottom-up-Prinzips dezentriert und in ein Netzwerk aus vielen einzelnen Einheiten (z.B. einfache Zustandsautomaten oder neuronale Knoten) zerlegt, die je für sich auf Umweltimpulse reagieren. System und Umwelt werden somit durch einfachste sensomotorische Schleifen relativ unvermittelt aneinandergekoppelt. Konkret geht es der emergenten Robotik darum, nicht jedes Verhalten programmieren zu müssen, sondern durch die Kombination verschiedener Reiz-Reaktions-Schemata neues Verhalten in Kontakt mit der Umwelt entstehen zu lassen. Die enge System-Umwelt-Verschaltung soll eine gewisse Ablösung der Maschine von der Kontrolle durch ihre Konstrukteur*innen und damit einen Zuwachs an Autonomie ermöglichen. Der Mensch erscheint dann zunehmend nicht mehr als steuernde Instanz, sondern als ›Bystander‹ einer scheinbar autonom interagierenden Maschine.

Diese System-Umwelt-Kopplung ist von kybernetischen Prinzipien inspiriert. Kybernetik ist der Versuch, »eine Theorie zu entwickeln, die den gesamten Bereich von Steuerung und Kommunikation in Maschinen und lebenden Organismen abdeckt« (Wiener 2002: 15). Hiermit verbunden ist der Verzicht auf die Suche nach intrinsischen Eigenschaften von Mensch, Tier und Maschine (vgl. Ashby 2016): Wichtig ist das Verhalten von Systemen und Organismen, nicht aber ihre materielle Komposition oder ihr genetischer Code. Vor diesem Hintergrund werden nicht nur Maschinen als Black Box konzeptualisiert, sondern ebenfalls Menschen und Tiere. Auf diese Weise wird es möglich, Bio- und Maschinenlogiken einander anzunähern (vgl. Weber 2017: 351), also Roboter mit anthropomorphem Vokabular zu beschreiben, Analogien zu organischen Funktionsprinzipien zu bilden und Vergleiche zum Menschen und zu Tieren (etwa Insekten) heranzuziehen. Hierzu ein Interviewauszug:

> *Soz.:* Haben Sie da schon in Bezug auf die Frage, wie Intelligenz entsteht, erste Ideen?
> *Rob.:* Wichtig ist, dass der Antrieb immer durch das untere Paradigma kommt, also was im Körper Bewegung antreibt und nicht imperative Strukturen, also welches Ziel will ich erreichen, was muss ich dann tun? Das kommt erst sehr viel später. Kinder müssen ja auch erst spielen.

Intelligenz basiert nicht auf einem hierarchischen Aufbau – so mein Gesprächspartner –, sondern auf dem ungezielten Zusammenwirken einfacher Mechanismen. Die Ziellosigkeit der einzelnen Mechanismen wird analog-räsonierend mit kindlichem Spiel verglichen. Kinder werden damit als planlose, aber zugleich als sich bewusst werdende, sich entwickelnde Wesen vorgestellt. Es handelt sich hier um eine bereits aus der frühen Neuzeit bekannte Vorstellung des Kindes als einer plastischen biologischen Ausgangsstruktur jenseits kultureller Inskriptionen (vgl. Kalthoff/Link 2021). Sowohl in Publikationen und medialen Darstellungen als auch in der konkreten Forschungspraxis dient die Figur des Kindes erstens als heuristischer Bezugspunkt der Forschung, ist aber zweitens auch ein dankbares Mittel für Robotiker*innen, hohe Erwartungen an die Maschine zu begrenzen.

Für die Ausbildung ihrer plastischen Maschinen schaffen Labore bisweilen Experimentierobjekte an, die in Form und Funktion an Kinderspielzeug erinnern: bunte Klötzchen, Alphabet-Magnete, Bälle, Stofftiere usw. Roboter sollen schließlich – analog zur kindlichen Entwicklung – ›lernen‹, diese Dinge zu halten, in eine Reihenfolge zu bringen oder sie schlicht als solche zu erkennen. Die Idee bei der Nutzung von Spielzeug ist, subtile Eigenschaften von Objekten farblich und durch die Größe zu verstärken, also eine gewisse Eindeutigkeit herzustellen. Ein Interviewauszug:

> *Soz.:* Welchen Unterschied macht es denn, wenn man Spielzeug verwendet?
> *Rob.:* Damit klappt das ganz gut, das ist halt einfacher. Das ist so eine Einstiegsdroge, sag ich mal, bevor man zu Objekten aus dem Alltag kommt. Weil da ist dann immer alles schwieriger. Die sind halt glänzend, haben dann keine schöne zylindrische Form, dann sind die am Ende spitz und die haben nicht so poppige Farben. Aber klar, zu den Alltagsobjekten müssen wir hin am Ende. Aber irgendwie muss man versuchen, sich die Aufgabe zunächst so leicht wie möglich zu machen und dann immer komplexer zu werden.

Roboter werden zunächst in teilstrukturierten Umgebungen trainiert, die menschlichen Umgebungen nur in wenigen Aspekten ähnlich sein sollen,

»weil dann ist immer alles schwieriger«. Das Ziel ist ein zweifaches: Erstens sollen Roboter schrittweise an unseren gesellschaftlichen Alltag herangeführt werden. Dabei richtet sich das Vorhaben nach der Vorstellung, Roboter nicht nur zum Objekt der technischen Konstruktionsarbeit zu machen, sondern auch zum Objekt der menschlichen Erziehung, das heißt, die maschinelle Entwicklung wird durch die Simulation kindlicher Objektumwelten der menschlichen angenähert. Zu diesem Zweck werden zuweilen entwicklungsdiagnostische Tests für Kinder auf Roboter angewendet: »um das ordentlich zu fundieren« (Robotiker); um also eine Referenz sowohl für die eigene Forschung als auch für die wissenschaftliche Community zu entwickeln. Das zweite Ziel ist es, nicht nur Robotern, sondern auch Robotiker*innen die technische Übersetzung und Formalisierung kontingenter Alltagsphänomene zu vereinfachen. Diese müssen mit ihren Robotern zunächst gewissermaßen ›spielen‹ bzw. einen relativ niedrigschwelligen Einstieg in die technische Bewältigung von Alltagsanforderungen finden.

Nun sind Robotiker*innen gewiss weit davon entfernt, robotische Kinder zu entwerfen, die in der menschlichen Gesellschaft ›groß werden‹. Faktisch handelt es sich um ein grundlagenwissenschaftliches Experimentieren mit kybernetischen und evolutionsbiologischen Ideen von ›lernenden‹ Maschinen. Denn wie so oft in der Roboterforschung kam es auch in der emergenten Robotik zwar zu raschen Anfangserfolgen, auf die allerdings bald neue, kaum zu überwindende Herausforderungen folgten. Eine prominente Schwachstelle wird unter dem Begriff »Scaling-up-Problem« (Christaller et al. 2001: 73) verhandelt: Zwar gelingt es, einfache Verhaltensweisen zu entwickeln, jedoch bleibt weitestgehend unklar, wie diese Architekturen in ihrer Komplexität gesteigert werden können: »Wie kommt man von den heute üblichen 20–40 Verhaltensweisen zu tausend, Millionen und noch mehr Verhaltensweisen?« (ebd.) Tatsächlich scheint der emergenten Robotik noch die passende Sprache und Methode zu fehlen. So kritisiert etwa Jutta Weber (2003), dass neuere Robotikzweige ihr Menschenbild zwar so dynamisieren, dass zunehmend das dialogische Werden mit der und durch die Umwelt einbegriffen wird, dabei jedoch verkennen, dass die Erzählungen vom Werden letztlich in althergebrachten Vorstellungen vom Sein münden. So greift die emergente Robotik auf Individuationsannahmen zurück, die Weber mit der naturalisierenden Logik des ›survival of the fittest‹ in Zusammenhang bringt: Mutter Natur dient als Vorbild, die mit dem Prinzip der natürlichen Selektion die Ausbildung sich entwickelnder und offener Organismen vorgibt (ebd.: 129). Der Vorstellung

von heterogenen Mensch-Nichtmensch-Gefügen würde nun wieder eine stabile Natürlichkeit untergehoben.

Neben dieser partiellen Rückkehr zu Vorstellungen einer stabilen menschlichen Natur machen sich auch methodische Schwierigkeiten bemerkbar: Die Implementierung einfacher sensomotorischer Feedbackschleifen durch neue Interfaces wie Tastsensoren oder Sprachausgaben sollte nicht über die Tatsache hinwegtäuschen, dass es sich bei emergenten Robotern immer noch um Maschinen handelt, die Algorithmen verarbeiten und in vorprogrammierten »*Trial and Error*-Rekursschleifen« (Bächle 2015: 275; Hervorh. im Orig.) operieren. Genau genommen beschreibt der Emergenzbegriff der Robotik mehr eine Hoffnung oder ein Ideal denn eine Tatsache. Begreift man Emergenz als das Entstehen einer gänzlich neuen algorithmischen Struktur, erscheint der Gehalt der Prophezeiungen dieser neueren Robotikzweige in einem neuen Licht: Das Verhalten dieser Roboter mag zwar aufseiten der Beobachter*in zu Überraschungen führen, es sollte jedoch nicht als Emergenz neuer algorithmischer Strukturen überhöht werden (vgl. Christaller et al. 2001: 70f.).

Ich resümiere: Jenseits ihrer re-naturalisierenden Tendenz und den gegenwärtigen technischen Limitationen scheint die emergente Robotik zunehmend von kognitivistischen Ansätzen und somit von einem humanistischen Menschenbild abzuwenden. So modelliert sie Intelligenz nicht mehr jenseits der Welt und des Körpers. Die materielle Umwelt des Roboters (Spielzeug, Infrastruktur und menschliche Gesten) wird in den Prozess des Intelligent-*Werdens* einbezogen. An die Stelle von Vorabprogrammierungen der symbolischen Robotik sollen nun biologisch inspirierte algorithmische Lernprozesse treten. Hierfür werden Vorstellungen über kindliches Lernen technisch nutzbar gemacht. Beobachtungen über das Kind sollen für die Dynamisierung des Roboters herhalten und bilden eine heuristische Grundlage für ein System, das nicht mit sich identisch bleibt, sondern offen und veränderlich sein soll. Abseits der Frage, ob die emergente Robotik ihren Ansprüchen tatsächlich gerecht wird, ist damit ein neuer Ernst verbunden, mit dem der Roboterhardware begegnet wird: Die materiellen Voraussetzungen für Körperlichkeit erhalten in der Entwicklung einen zentralen Status, wodurch die Hardware zuweilen als Akteurin in die Konstruktion eingespannt wird. Vor diesem Hintergrund werden in neueren Entwicklungszweigen der Robotik Intelligenzstrukturen und Roboter nicht getrennt voneinander verstanden. ›Intelligenz‹ soll *mit* und *an* dem Korpus des Roboters entstehen.

3.2 Roboter konstruieren: Die Arbeit an und mit Robotermaterial

Im Folgenden wird es darum gehen, diese konzeptionelle Umdeutung von Materialität in den konkreten Situationen der Roboterkonstruktion zu zeigen. Es soll dabei in die Arbeit an der Roboterhardware eingetaucht und dokumentiert werden, wie Robotiker*innen auf ihr Material treffen und wie sich eine konflikthafte und dynamische Relation entfaltet.

Inmitten von Apparaturen, Kabeln, Werkzeugen, Computern, 3-D-Druckern, aber auch Kaffeemaschinen, zwischengelagerten Fahrrädern, eingemotteten Roboterteilen und Resten des Mittagessens kommen Roboter allmählich zu ihrer Form. Jenseits der Vorstellung einer fragilen hochtechnologischen Infrastruktur, die von der Außenwelt geschützt und von Kittel tragenden Robotiker*innen instand gehalten wird, findet die Konstruktion eines Roboters in hybriden Räumen statt, die sowohl Alltagsobjekte als auch kritische technische Apparaturen beinhalten.

Hier müssen Roboter nicht nur gebaut, sondern zunächst geplant und entworfen werden. Vor dem Bau eines Roboters gehen tatsächlich mehrere Wochen und Monate der Vorbereitung ins Land. Diese Vorbereitung umfasst oftmals Absprachen mit Abnehmer*innen. Bei solchen Absprachen, die vor allem dann nötig sind, wenn Roboter externen Zwecken dienen sollen, werden in der Regel zwei Dinge geklärt: Ästhetik auf der einen (wie soll der Roboter aussehen?) und Funktionalität auf der anderen Seite (was soll er können?). Anschließend werden konzeptualisierende Skizzen angefertigt und Berechnungen durchgeführt, die dann prüfend in eine Computersimulation überführt werden (Abb. 1).

Abbildung 1

© Hannah Link

Hierauf folgt letztlich die Modellierung in Form von 3-D-Drucken, die in den Händen beäugt und gewendet werden (Abb. 2). Anschließend erfolgen gegebenenfalls Materialrecherchen und es werden Baustoffe bestellt, mit denen letztlich der Roboter (Abb. 3) zusammengebaut wird.

Abbildung 2

© Hannah Link

Abbildung 3

© Hannah Link

Hierzu ein Ausschnitt aus einem ethnografischen Beobachtungsprotokoll. Dokumentiert wird eine Situation aus der der relativ fortgeschrittenen Phase der Baustoffaufbereitung:

> Nach der Mittagspause im Labor gehen Henri, der für die Konstruktion des Roboters verantwortliche Doktorand, Roland, sein Student, und ich zurück in die Werkstatt. Henri will Roland und mir zeigen, was wir für den Bau des Roboters beachten müssen. Henri spannt mich hier und da in den Konstruktionsprozess ein und weist mir einige einfache Aufgaben zu, die ich ohne technisches Vorwissen erledigen kann. Er erklärt uns, dass erst die Kanten der frisch geschnittenen Konstruktionsprofile gesäubert werden müssen, damit Spitzen und Splitter der Metallecken später andere Hardwareteile nicht beschädigen. Ich fluche innerlich. Meine Hände sind noch vom gestrigen Aussäubern der Kanten wund – heute darf ich auf keinen Fall die Handschuhe vergessen. Ich schaue Roland an, dem ich einen ähnlich panischen Gesichtsausdruck ablesen kann.
>
> Ich halte das Messer, geformt wie eine kleine Sichel, zwischen Daumen und Zeigefinger in der Hand und stütze es am Griff mit der Handfläche ab. Um

> alle Splitter abschneiden zu können, muss die Schneidefläche des Messers gegen die Kante gepresst und unter gleichbleibendem Druck nach vorne bewegt werden. Der Druck und die Bewegung schneidet nicht nur in die Konstruktionsprofile, sondern auch in meine Hand: Der stumpfe Griff wird mit einem Mal hart und schmerzhaft.
> Henri fährt fort: Dann sollen wir unbedingt daran denken, mit dem Druckluftkompressor alle Profile abzupusten […]. Dabei aber immer – ganz wichtig – Brillen tragen, damit uns keine Splitter in die Augen fliegen. Schon schnappt sich Henri ein Konstruktionsprofil aus der Ecke des Raumes, hält es auf Armlänge von sich und mit der anderen Hand den Kompressor, drückt einen Knopf und pustet zügig und gekonnt das Profil ab. Roland und ich springen schnell, wenn auch zu spät, zur Seite und wenden unsere Gesichter ab. Wir lachen erschrocken und befühlen halb ironisch, halb ängstlich unsere Augen – Henri hat uns keine Zeit gegeben, unsere Brillen aufzusetzen.

Robotermaterial muss aufwendig vorbereitet, aufbereitet und nachbereitet werden. Unter Körpereinsatz und der Nutzung etwaiger Instrumente wie Messern und Kompressoren wird das Material geschliffen, zurechtgeschnitten und gesäubert. Die Konstruktion eines Roboters beruht daher zu einem wesentlichen Teil auf handwerklicher körperlicher Arbeit. Dies führt mitunter zu Frustration und zerschundenen Körperstellen. Dabei zeigen sich die Körper in ihrer Fragilität, während sich das Robotermaterial in seiner Standhaftigkeit offenbart. So benötigt es Handschuhe, Brillen und etliche weitere Dinge, um die Körper vor dem widerspenstigen Material zu schützen. Die Schutzbedürftigkeit unterscheidet sich jedoch. So ist der erfahrene Robotiker vertraut im Umgang mit dem Material und dem Werkzeug, mit dem er schnell und geschickt jegliche Splitter entfernt. Die Tätigkeit ist bisweilen von Souveränität, inkorporiertem Wissen und Routine geprägt. Das Material erscheint ihm als vertraute Ressource; für die Laien hingegen ist es widerständig und erfordert ein strapaziöses Abarbeiten. Die unterschiedlichen Erscheinungsweisen des Robotermaterials im Zusammenhang mit dem offensichtlichen Kompetenzgefälle lassen vermuten, dass sich bestimmte Aktivitätspotenziale des Materials im handwerklichen Prozess bemerkbar machen. Das Robotermaterial verlangt den Akteur*innen bestimmte Fertigkeiten ab und strukturiert das, was mit ihm verrichtet werden kann, mit: Es formatiert gewissermaßen Komplementärkörper, die sich entlang der materiellen Erfordernisse ausbilden. Diese Passung der Menschenkörper macht

sich mitunter durch wunde Stellen oder Einkerbungen der Haut bemerkbar, zeigt sich aber auch darin, dass Schutzkleidung angelegt werden muss und in der Notwendigkeit zur Aneignung professionellen Körperwissens.

Die Dinge liegen also anders: Nicht nur das Robotermaterial wird zum Gegenstand der Bearbeitung, sondern auch die Robotiker*innenkörper werden entsprechend den Materialerfordernissen formatiert. Es zeigt ein gewisses Aktivitätsniveau des Materials, das spezifische Körperantworten der menschlichen Teilnehmer*innen provoziert. Es handelt sich um eine konflikthafte Ko-Konstitution von Robotiker*innen und Robotern, da die Verarbeitung des Materials zu einem Roboter nicht voraussetzungslos ist. So ist es auf der einen Seite formbar, hinterlässt aber auf der anderen Seite Spuren. Gemeinsam durchlaufen Forscher*innen und Forschungsobjekt eine Transformation: Durch den Kontakt mit dem Material entstehen professionelle Robotiker*innen mit einem spezifischen Wissen und umgekehrt erfahren Konstruktionsprofile einen Zuschnitt.

In einem zweiten Beispiel bauen die Akteur*innen den Roboter zusammen. Zu beobachten ist dabei ein Moduswechsel vom anfänglichen Handwerk zur ingenieurwissenschaftlichen Konstruktion:

> Henri, Roland und ich sitzen an der Tischgruppe im Labor. Auf Kopfhöhe hat Henri einen Ausdruck der Simulation des Roboters aufgehängt und auf dem Tisch befindet sich ein Miniaturmodell des Roboters aus dem 3-D-Drucker. So soll er mal aussehen. Nach einer Weile konzentrierten Ineinanderstekkens und Herumschraubens hält Henri jetzt das zusammengeschraubte Robotergelenk vor sein Gesicht, kneift die Augen zusammen und betrachtet es mit gerunzelter Stirn. Jetzt zieht er sein Knie etwas hoch vor seinen Bauch und hält die Gelenkkonstruktion erst davor und drückt sie dann gegen sein Knie. Seine Arme zittern. Er hält die Konstruktion wieder hoch vor sein Gesicht, beäugt sie, dreht sie, ruckelt dann ein wenig daran und schaut sie wieder mit gerunzelter Stirn an. ›Was machst du?‹, frage ich. ›Ich will sehen, was passiert, wenn ich hier ein bisschen Kraft ausübe.‹ Und noch einmal: Er drückt das Gelenk mit aller Kraft an seine Kniescheibe. Nichts. ›Weil, wenn der Roboter sich auf dem Boden bewegt, sind da ja auch Kräfte.‹ Nach einer Weile fügt er leise hinzu: ›Okay es funktioniert, ist halt super schwer, aber das passt schon, glaube ich‹, und wiegt das Gelenk in der Hand hin und her.

Die robotische Gelenkkonstruktion wird zwar aufwendig berechnet, gezeichnet, simuliert und am Ort des Geschehens in Bildform vergegenwärtigt, allerdings scheint bis zuletzt unsicher zu sein, wie sich die Konstruktion im Einsatz

konkret verhält, also auch, wie sich das Material im Gebrauch entfalten wird. Das Gelenk wird ad hoc und unsystematisch auf die Probe gestellt. Der Robotiker spannt dafür seinen Körper ein, fühlt das Robotergelenk durch seine Knie und wiegt es mit den Händen. Die leibliche Annäherung, das Vertrautmachen, wechselt sich aber zugleich mit einer Distanznahme ab: Der Robotiker tritt zurück, beäugt das Gelenk, zweifelt und führt es wieder näher an sich heran. Kann die Konstruktion halten? Die Kombination aus prüfendem Blick, leiblichem Ertasten und Ad-hoc-Erprobungen soll Ahnungen und Vorstellungen davon generieren, was sich später entfalten könnte. Der Herstellung eines Roboters scheint ein kritischer Restbestand an Kontingenz inhärent zu sein. Es kann sich so oder so entwickeln: halten, einbrechen, bestimmte Bewegungen ausführen oder eben nicht. Akribische Berechnungen, Zeichnungen und Simulationen müssen in letzter Instanz einer Intuition weichen, die auslotet, was möglich ist. So scheint sich die Konstruktion nicht zuletzt einer wissenschaftlichen Vereindeutigung und Verfügung zu entziehen. Aus der Perspektive des Feldes ist die Konstruktion also weniger ein materielles Ding, das – einmal gebaut – in vorhersehbarer Weise vorfindbar ist, denn eine kontingente Konstellation, deren Sein umstritten bleibt. Die Kraft des Robotermaterials liegt nun nicht nur in der Raumverschiebung, die durch die erzeugte Reibung dichter Objekte entsteht. In der intuitiven Erprobung des Gelenks zeigt sich vielmehr ein Umgang mit dem Material, das dessen Unberechenbarkeit antizipiert. Die Konstruktionspraxis macht ein Materialitätsverständnis sichtbar, das die Fähigkeit zur Veränderung, zur Überraschung ebenso mitdenkt wie die Möglichkeit, dass sich das Material einem formatierenden Zugriff entzieht.

Eine solch technisch-materielle Kontingenz wird in der emergenten Robotik nicht zwangsläufig kritisch gesehen, sondern mitunter gezielt in den Dienst genommen und nutzbar gemacht.[6] In einem solchen Fall werden die vorbereitenden Aktivitäten auf ein Minimum reduziert. Akribische Zeichnungen und Computersimulationen weichen sodann einem Verfahren des *rapid prototypings*. Ideen werden mit dem 3-D-Drucker direkt in Musterbauteile konvertiert. Dies geht auf die Überzeugung zurück, dass Material und Form eine gewisse Eigenaktivität oder – in den Worten eines Interviewpartners –

6 Während insbesondere in der symbolischen Robotik rigide Materialien (etwa spezifische Metalle), bei denen vorhersehbar ist, wie sie sich in spezifischen Situationen verhalten, verwendet werden, um die Genauigkeit der Roboteraktionen zu gewährleisten, greift etwa das Feld der *soft robotics* auf flexible Materialien (Gummi, Silikon etc.) zurück, deren Bewegungstrajektorie nicht immer vorhersagbar ist.

ein »Eigenleben« haben, das es zu nutzen gilt: »Lass doch den ganzen Käse die Hardware machen.«

Halten wir fest: Die Grenzen mathematischer Berechnungen und Vorausplanungen werden deutlich hervorgehoben. Begründet ist diese programmiertechnische Zurücknahme darin, dass materielle Eigenaktivitäten erkannt, kanalisiert und als Ko-Konstrukteur*innen nutzbar gemacht werden, aber auch eine unverfügbare Restkontingenz hingenommen wird. Als aktives und stellenweise unverfügbares Wissensobjekt ist der materielle Roboter im Prozess der Konstruktion daher inkompatibel mit der Fantasie einer vollständigen Durchdringung der Welt durch distanzierte Forscher*innen. Vielmehr lässt sich bei diesen Robotiker*innen die Vorstellung eines konstruierten Objekts beobachten, das sich zugleich selbst mitkonstruiert, sich entwickelt und Raum für Überraschungen und Irritationen schafft. Robotermaterie wird mithin als aktives Element im soziomateriellen Prozess der Konstruktion verstanden. Das ist es, was Donna Haraway (1995b) mit ihrem Begriff des situierten Wissens beschreibt: Eine nüchterne Einschätzung der eigenen Erschließungs- und Eingriffsmöglichkeiten im Prozess der Wissensproduktion durch die Anerkennung des produktiven Charakters von Wissensobjekten. »Die Kodierungen der Welt stehen nicht still«, schreibt sie dazu, »sie warten nicht etwa darauf, gelesen zu werden« (ebd.: 94).

4. Schluss

In diesem Beitrag habe ich dafür plädiert, eine posthumanistische Perspektive in die wissenssoziologische Untersuchung der Robotik zu integrieren. Ich habe argumentiert, dass (trotz Bemühungen zeitgenössischer Denker*innen) die latente Unterscheidung zwischen menschlichen und nichtmenschlichen Entitäten zum festen Inventar soziologischer Zugriffe gehört. Dieser persistente humanistische Impuls offenbart sich in der unerschütterlichen Vorstellung unterschiedlicher Aktivitätsklassen von Entitäten. Insbesondere Donna Haraways Arbeiten zeichnen sich durch einen gewissen Nachdruck aus, mit dem der latente Humanismus der Praxistheorien herausgefordert werden kann. So beansprucht sie, eine Welt zu denken, die eigenlogisch, also auch jenseits der Frage nach ihrer sozialen Verwertbarkeit, existiert (vgl. Folkers 2013).

Mit der hier entworfenen Perspektive wurde es möglich, eine technowissenschaftliche Praxis zu dokumentieren, die ›Künstliche Intelligenz‹ im Material selbst verortet. Robotermaterial wird nicht als technischer Außenbezirk

verstanden, der per Softwaresteuerung Befehle ausführt. Bei der emergenten Robotik handelt es sich vielmehr um eine Forschungspraxis, in der der Roboterkorpus selbst und die materiellen Elemente, die ihn ausmachen, zu einem epistemischen Objekt werden. So ist das Ziel, eine Korpus-›Intelligenz‹ zu entwickeln, die materiell begründet ist und nicht (allein) aus softwaretechnischer Symbolverarbeitung besteht. Robotermaterial wird hier also von einer ausführenden zu einer konstruierenden Instanz umgedeutet. Roboter in ihrer materiellen Verfasstheit sind in der emergenten Robotik – um es in Haraways Worten zu sagen – integraler Bestandteil eines »Apparates körperlicher Produktion« (Haraway 1995b: 91). Sie sind also Teil jener wissenschaftlichen Diskurse, Praktiken und Technologien, die Körper und Entitäten in spezifischer Weise hervorbringen.

In diesem Kontext konterkariert das Feld selbst die Fantasie eines überlegenen Menschensubjekts in zweifacher Weise: Erstens dient die Vorstellung eines dezentrierten (also aufs Engste mit der Umwelt gekoppelten) und dynamischen (also sich entwickelnden) Menschen als Heuristik für die Konstruktion intelligenter Maschinen. Und zweitens entzieht das Feld dem Menschen tendenziell seine Vormachtstellung, indem sie Materie als wirkmächtige Gegen- und Mitspielerin sowie Effekt der Forschung thematisiert.

So wurde zunächst anhand von Interviewdaten Auseinandersetzungen mit Materialität rekonstruiert, in denen Materie erstmals ein zentraler Status zugerechnet und sie als Grundlage für die Evolvierbarkeit von Mensch und Maschine diskutiert wurde. Darauf folgten Protokollauszüge, mit denen unterschiedliche Phasen des Materialbezugs von Robotiker*innen in situ dokumentiert wurde. Dabei wurde zunächst deutlich, wie Robotiker*innen mit dem zu formenden Material ringen. Anstelle einer einseitigen Einhegung von Bauplänen durch den Menschen wurde eine wechselseitige Formatierung und Ko-Konstitution von Robotiker*innen und Robotern festgestellt. Darauffolgend zeigte sich eine soziomaterielle Praxis, die sowohl die produktive Eigenaktivität als auch die Unberechenbarkeit des Robotermaterials voraussetzt.

Aus der Perspektive posthumanistischer Theorieangebote offenbart sich im Feld der emergenten Robotik eine dynamische und konflikthafte Aushandlung zwischen Robotiker*innen und Robotern: Robotiker*innen nähern sich einer situierten Praxis der Wissens- und Technikproduktion an. Das heißt, an die Stelle des Versuchs einer objektiven Beschreibung und Verfügbarmachung der Welt tritt ein engagiertes Zusammenspiel von Forschungssubjekt

und -objekt, in dem die Kraft und potenzielle Unverfügbarkeit der Wissensobjekte mitgedacht wird.

Ein Anliegen des Aufsatzes bestand zudem darin, die technowissenschaftliche Praxis als Re-Konfiguration der Welt ernst zu nehmen. Insofern war es Ziel, gesellschaftliche Potenziale neuerer Robotikzweige aufzuzeigen; es ging darum, ein neues technowissenschaftliches Verständnis ›des Menschen‹ sichtbar zu machen, das sich tendenziell von einem humanistischen Bild abzulösen scheint. Dies ist jedoch nicht im Sinne einer »Heilsgeschichte« misszuverstehen, die vor den gegenwärtigen Problemen technowissenschaftlicher Konstruktionsarbeit die Augen verschließt (vgl. Haraway 1995a: 35); so wurde deutlich, dass auch neuere Robotikzweige Re-Naturalisierungstendenzen aufweisen und somit zugleich Humantheorien mitführen, die auf eine einheitliche menschliche Natur rekurrieren. Vielmehr geht es mir um eine empirisch gesättigte Identifikation von Transformationsmöglichkeiten, die angeeignet und nutzbar gemacht werden können. Dahingehend könnten neuere Robotikzweige eine Herausforderung für ein Technikverständnis sein, das einer Polarität von Natur und Kultur verhaftet bleibt, das Natur als Ressource für Kultur und Technik als Mittel des Verfügbarmachens begreift.

Literatur

Ahmed, Sara. 2008. Open Forum Imaginary Prohibitions: Some Preliminary Remarks on the Founding Gesture of the ›New Materialism‹. *European Journal of Women's Studies* 15, H. 1: 23–39.

Alač, Morana. 2009. Moving Android: On Social Robots and Body-in-Interaction. *Social Studies of Science* 39, H. 4: 491–528.

Ashby, W. Ross 2016. *Einführung in die Kybernetik*. Frankfurt a.M.: Suhrkamp.

Bächle, Thomas Christian. 2015. *Mythos Algorithmus. Die Fabrikation des computerisierbaren Menschen*. Wiesbaden: Springer VS.

Barad, Karen. 2012. *Agentieller Realismus*. Berlin: Suhrkamp.

Beck, Susanne. 2012. *Jenseits von Mensch und Maschine. Ethische und rechtliche Fragen zum Umgang mit Robotern, Künstlicher Intelligenz und Cyborgs. Robotik und Recht*. Baden-Baden: Nomos.

Becker, Barbara und Jutta Weber. 2005. Verkörperte Kognition und die Unbestimmtheit der Welt. Mensch-Maschine-Beziehungen in der Neueren KI. In *Unbestimmtheitssignaturen der Technik. Eine neue Deutung der technisier-*

ten Welt, Hg. Gerhard Gamm und Andreas Hetzel, 119–232. Bielefeld: transcript.

Bennett, Jane. 2010. *Vibrant Matter: A Political Ecology of Things*. Durham und London: Duke University Press.

Bourdieu, Pierre. 1980. Le mort saisit le vif. Les relations entre l'histoire réifiée et l'histoire incorporée. *Actes de la recherche en science socials* 32/33 : 3–14.

Braidotti, Rosi. 2014. *Posthumanismus. Leben jenseits des Menschen*. Frankfurt a.M. und New York: Campus.

Brooks, Rodney. 2002. *Menschmaschinen. Wie uns die Zukunftstechnologien neu erschaffen*. Frankfurt a.M. und New York: Campus.

Čapek, Karel. 2017. *W.U.R. Werstands Universal Robots*. Berlin: Hofenberg.

Christaller, Thomas, Michael Decker, Joachim-Michael Gilsbach, Gerd Hirzinger, Karl Lauterbach, Erich Schweighofer, Gerhard Schweitzer und Dieter Sturma (Hg.). 2001. *Robotik. Perspektiven für menschliches Handeln in der zukünftigen Gesellschaft*. Berlin und Heidelberg: Springer.

Coeckelbergh, Mark. 2020. *AI Ethics*. Cambridge: MIT Press.

Dreyfus, Hubert. 1985. *Die Grenzen künstlicher Intelligenz: Was Computer nicht können*. Königstein: Athenäum.

Folkers, Andreas. 2013. Was ist neu am neuen Materialismus? Von der Praxis zum Ereignis. In *Critical Matter. Diskussionen eines neuen Materialismus*, Hg. Tobias Goll, Daniel Keil und Thomas Telios, 17–35. Münster: Edition Assemblage.

Glaser, Barney und Anselm Strauss. 1967. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Mill Valley: Sociology Press.

Haraway, Donna. 1988. Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies* 14, H. 3: 575–599.

Haraway, Donna. 1995a. Ein Manifest für Cyborgs. Feminismus im Streit mit den Technowissenschaften. In *Die Neuerfindung der Natur. Primaten, Cyborgs und Frauen*, Hg. Carmen Hammer und Emanuel Stieß, 33–72. Frankfurt a.M. und New York: Campus.

Haraway, Donna. 1995b. Situiertes Wissen. Die Wissenschaftsfrage im Feminismus und das Privileg einer partialen Perspektive. In *Die Neuerfindung der Natur. Primaten, Cyborgs und Frauen*, Hg. Carmen Hammer und Emanuel Stieß, 73–97. Frankfurt a.M. und New York: Campus.

Haraway, Donna. 2016. *Das Manifest für Gefährten. Wenn Spezies sich begegnen: Hunde, Menschen und signifikante Andersartigkeit*. Berlin: Merve.

Hergesell, Jannis, Arne Maibaum, Andreas Bischof und Benjamin Lipp. 2021. Zum Potenzial grundlagenwissenschaftlicher Technikforschung für ein »gutes Leben im Alter«. Ein Plädoyer für konsequente partizipative Technikgestaltung. In *Gute Technik für ein gutes Leben im Alter? Akzeptanz, Chancen und Herausforderungen altersgerechter Assistenzsysteme,* Hg. Debora Frommeld, Ulrike Scorna, Sonja Haug und Karsten Weber, 293–316. Bielefeld: transcript.

Hirschauer, Stefan. 2016. Verhalten, Handeln, Interagieren. Zu den mikrosoziologischen Grundlagen der Praxistheorie. In *Praxistheorie. Ein soziologisches Forschungsprogramm*, Hg. Hilmar Schäfer, 45–70. Bielefeld: transcript.

Hoppe, Katharina und Thomas Lemke. 2015. Die Macht der Materie. Grundlagen und Grenzen des agentiellen Realismus von Karen Barad. *Soziale Welt* 66: 261–279.

Hoppe, Katharina. 2019. Donna Haraways Gefährt*innen: Zur Ethik und Politik der Verwobenheit von Technologien, Geschlecht und Ökologie. *Feministische Studien* 37, H. 2: 250–268.

Kalthoff, Herbert und Hannah Link. 2021. Zukunftslaboratorien. Technisches Wissen und die Maschinenwesen der Robotik. In *Humandifferenzierung. Disziplinäre Perspektiven und empirische Sondierungen*, Hg. Dilek Dizdar, Stefan Hirschauer, Johannes Paulmann und Gabriele Schabacher, 314–334. Weilerswist: Velbrück.

Kalthoff, Herbert. 2018. Theoretische Empirie und ihre Konsequenzen. In *Zum Verhältnis von Empirie und kultursoziologischer Theoriebildung. Stand und Perspektiven*, Hg. Julia Böcker, Lena Dreier, Melanie Eulitz, Anja Frank, Maria Jakob und Alexander Leistner, 132–152. Weinheim und Basel: Beltz Juventa.

Keller, Evelyn. 1983. *A Feeling of the Organism: The Life and Work of Barbara McClintock*. San Francisco: Freeman.

Latour, Bruno. 1993. *La clef de Berlin et autres leçons d'un amateur de sciences*, Paris : La Découverte.

Link, Hannah und Herbert Kalthoff. 2023. Die Naturalisierung des Roboters. Zu einer Soziologie technischen Wissens. In *Jenseits des Menschen. Neue Perspektiven auf Natur/Kultur*, Hg. Torsten Cress, Oliwia Murawska und Annika Schlitte. Paderborn: Fink/Brill (im Erscheinen).

Lipp, Benjamin. 2022. Caring for Robots: How Care Comes to Matter in Human-Machine Interfacing. *Social Studies of Science*: 1–26.

Meister, Martin und Ingo Schulz-Schaeffer. 2016. Investigating and Designing Social Robots from a Role-Theoretical Perspective. *AI & Society* 31: 581–585.

Newell, Allen und Herbert Simon. 1997. Computer Science as Empirical Inquiry: Symbols and Search. In *Mind Design II. Philosophy, Psychology, Artificial Intelligence,* Hg. John Haugeland, 81–110. Cambridge: MIT Press.

Pfeifer, Rolf und Christian Scheier. 1999. *Understanding Intelligence*. Cambridge: MIT Press.

Pfeifer, Rolf. 2000. Embodied Artificial Intelligence: 10 Years Back, 10 Years Forward. In *Informatics: 10 Years Back. 10 Years Ahead. Lecture Notes in Computer Science*, Hg. Reinhard Wilhelm, 294–310. Berlin: Springer.

Pitsch, Karola. 2016. Limits and Opportunities for Mathematizing Communicational Conduct for Social Robotics in the Real World? *AI & Society* 31: 587–593.

Reckwitz, Andreas. 2002. The Status of the ›Material‹ in Theories of Culture. From ›Social Structure‹ to ›Artefacts‹. *Journal for the Theory of Social Behaviour* 32: 195–217.

Richardson, Kathleen. 2016. Technological Animism: The Uncanny Personhood of Humanoid Machines. *Social Analysis* 60, H. 1: 110–128.

Šabanović, Selma. 2014. Inventing Japan's ›Robotics Culture‹: The Repeated Assembly of Science, Technology, and Culture in Social Robotics. *Social Studies of Science* 44, H. 3: 342–367.

Schatzki, Theodore R. 2002. *The Site of the Social. A Philosophical Account of the Constitution of Social Life and Change*, University Park, PA: The Pennsylvania State University Press.

Steels, Luc und Rodney Brooks. 1995. *The Artificial Life Route to Artificial Intelligence: Building Situated Embodied Agents*. New York: Routledge.

Suchman, Lucy. 2007. *Human-Machine-Reconfiguration: Plans and Situated Actions.* 2. Auflage. New York: Cambridge University Press.

Turkle, Sherry. 1984. *The Second Self : The Human Spirit in a Computer Culture.* New York: Simon & Schuster.

Weber, Jutta. 2003. Turbulente Körper und emergente Maschinen. Über Körperkonzepte in neuerer Robotik und Technikkritik. In *Turbulente Körper, soziale Maschinen. Feministische Studien zur Technowissenschaftskultur*, Hg. Jutta Weber und Corinna Bath, 119–136. Opladen: Leske + Budrich.

Weber, Jutta. 2017. Feministische STS. In: *Science and Technology Studies – Klassische Positionen und aktuelle Perspektiven,* Hg. Susanne Bauer, Torsten Heinemann und Thomas Lemke, 339–368. Berlin: Suhrkamp.

Wiener, Norbert. 2002. Kybernetik. In *Futurum Exactum. Ausgewählte Schriften zur Kybernetik und Kommunikationstheorie*, Hg. Bernhard Dotzler, 13–30. Wien: Springer.

Wieser, Matthias. 2015. Inmitten der Dinge. Vom Verhältnis von sozialen Praktiken und Artefakten. In *Doing Culture. Neue Positionen zum Verhältnis von Kultur und sozialer Praxis*, Hg. Karl H. Hörning und Julia Reuter, 92–107. Bielefeld: transcript.

Menschmaschinen und Maschinenmenschen

Überlegungen zur relationalen Ontogenese von Identität

Jonathan Harth & Maximilian Locher

Abstract: *In diesem Beitrag wenden wir uns der Frage zu, wie in Mensch-Maschine-Situationen die Identitäten der beteiligten Entitäten aufeinander Bezug nehmen und dadurch überhaupt erst hervorgebracht, konfirmiert oder auch verworfen werden. Damit schließen wir uns dem Paradigma der relationalen Soziologien an und folgen dem von Athanasios Karafillidis vorgeschlagenen Programm einer »Relationsmustererkennung«, das die koproduzierte Beziehung von Mensch und Maschine in den Vordergrund stellt. Die hiermit entstehenden sozialtheoretischen und epistemologischen Fragen, wie sich die Identität der an sozialen Situationen Beteiligten bestimmen lässt und wie Relationsmustererkennung überhaupt möglich ist, werden abschließend vor dem Hintergrund der Adaption in empirischen Forschungsprogrammen methodologisch diskutiert.*

In vielen gesellschaftlichen Bereichen kommt es für Menschen täglich zu Begegnungen mit Maschinen der Künstlichen Intelligenz (KI). Diese Begegnungen verändern unsere Verständnisse von Selbst-, Welt- und Fremdverhältnissen. Sie werden gerahmt von Vorwissen und eröffnen Möglichkeiten neuer Rahmungen dieser Verhältnisse. Gleichzeitig stehen diese Verhältnisse selbst stets zur Disposition. Sie sind gewordene Struktur, die sich permanent operativ restrukturiert.

In unserem Beitrag möchten wir uns mit der Frage nach den damit verbundenen epistemologischen wie auch kulturellen Implikationen beschäftigen, die sich aus einer relationalen Wendung ergeben, und widmen uns daher den sich je öffnenden und schließenden Möglichkeiten der Produktion, Transformation oder Erosion relationaler Identitäten. Bereits diese Formulierung zeigt die Perspektive dieses Ansatzes an: Es geht um die Relationen und die Relationierung innerhalb situativer Dynamiken und um die prinzipielle Frage,

wie sich uns andere Menschen, Dinge, Tiere – und damit ganz eigentlich: alle Phänomene überhaupt erst erschließen und wie diese Erschließung wiederum auf uns selbst und unser Selbstbild zurückwirkt (vgl. Coeckelbergh 2011; McFarlane 2013). Die diesen Beitrag tragende Annahme besteht darin, dass das Spiel der Ausbildung relationaler Identitäten durch den Eintritt von KI in die gesellschaftliche Wirklichkeit neue Dynamiken erzeugt und dementsprechend neue Aufmerksamkeit verdient.

Kommerziell werden intelligente Maschinen wie KIs bereits seit mehreren Jahren erfolgreich zur Automatisierung, Rationalisierung oder Temporalisierung eingesetzt. Auch bei der alltäglichen Benutzung eines Smartphones greift man auf eine Vielzahl an KI-Technologien zurück – zumeist ohne es explizit zu bemerken. In der Nutzung aktueller und für jedermann verfügbarer kreativer KI-Modelle wir DALL-E, Midjourney, GPT-3 oder Stable Diffusion wird die relationale Ko-Produktion nochmals sichtbarer: Plötzlich lassen sich künstlerische Ausdrucksweisen realisieren, die vorher nicht für möglich gehalten wurden. Digitale Technologie besticht dadurch, jede ihrer Oberflächen (*surface*) mit einer relativ verborgenen Unterfläche (*subface*) zu koppeln (vgl. Nake 2008). Vor allem die zunehmende Komplexität dieser Tiefen algorithmischer Informationstechniken weist auf eine mittlerweile unumgängliche Interpretationsbedürftigkeit ihrer Funktions- und Anschlussweisen hin (vgl. Burrell 2016; Esposito 2017). Beispielsweise produzieren neuronale Netze Outputs, die auch von ihren Erschaffer*innen nicht mehr exakt vorhergesehen werden können – da genau dies unter anderem ihre Funktion ist. Darüber hinaus ist zu beobachten, dass diese Maschinen ihren Zugriff auf die Welt zunehmend selbst erzeugen. Dieses bereits von Alan Turing (1950) unter dem Titel »unorganized machines« beobachtete Phänomen wird heute unter dem Label »unsupervised learning« verhandelt (vgl. LeCun/Bengio/Hinton 2015) und in diversen Modellen zu vielfältiger Anwendung gebracht (vgl. exemplarisch Schrittwieser/Antonoglou/Hubert 2020; Piloto et al. 2022). Diese technischen Leistungen fordern dazu auf, auf die mit ihnen einziehende neue Unbestimmtheit und Interpretationsbedürftigkeit zu reagieren, und stellen damit zwangsläufig die Frage nach der Identität dieser Phänomene: Womit haben wir es hier eigentlich zu tun und was macht das mit uns?

So unsinnig es beispielsweise erscheinen mag, einer Maschine Bewusstsein oder leibliche Subjektivität zuzurechnen, legen manche Beispiele dennoch nahe, dass KI Sinn verarbeiten kann und uns Menschen in gewisser Hinsicht ähnlich ist. Wenn ein Ingenieur von Google behauptet, im zu überprüfenden Sprachmodell eine Person zu erkennen (Lemoine 2022), Chat-

Teilnehmer*innen in einem Chatbot einen Menschen sehen (Humphreys 2009) oder Teilnehmende einer Studie trotz Aufklärung über den Mechanismus des Roboters eine Persönlichkeit wahrnehmen (Turkle 2012), dann steht im Hintergrund solcher Betrachtungen jeweils die mehr oder weniger latente Frage, ob oder wann es sich bei einem Gegenüber um ein *Ding* handelt, einen *anderen wie mich* oder gar einen *anderen Anderen*. Theoretisch wie empirisch gilt es dann zu operationalisieren, wie das Erkennen dieser Phänomene eigentlich erkannt werden kann.

Relationale Ansätze innerhalb der Soziologie machen seit Längerem vielversprechende Angebote, wie die anthropozentrische Reduktion und Zentrierung des Sozialen auf menschliche Aktivitäten überwunden werden könnte (vgl. u.a. Emirbayer 1997; Knorr-Cetina 2006; Donati 2010; Fuhse/Mützel 2010). Es ist zu vermuten, dass für diese sogenannten postsozialen Theorien nicht nur biologisch-ökologische Dringlichkeiten (Klimawandel) den Blick für Verflechtungen, Abhängigkeiten und Werdensprozesse des Sozialen öffneten, sondern auch die aktuellen Leistungsfähigkeiten intelligenter Maschinen. Während die ökologische Perspektive auf die relationale Ontogenese des Sozialen eher die biologische Seite betont und verhandelt (vgl. Latour 2020), weist die technologische Perspektive auf die relationale Ontogenese des Sozialen mehr auf die Frage nach der Identität der am Sozialen beteiligten Referenzen hin.

In diesem Beitrag werden wir uns im Folgenden der Frage widmen, wie in Mensch-Maschine-Relationen die Identitäten der beteiligten Entitäten aufeinander Bezug nehmen und dadurch überhaupt erst hervorgebracht, konfirmiert oder auch verworfen werden. Hierbei möchten wir die zentrale Untrennbarkeit von Ontologie und Epistemologie und das Primat einer Ontogenese in den Vordergrund stellen, die ihre Identitäten erst über Beziehungen generiert.

1. Von der Ontologie zur Ontogenese

Der Diskurs über die Identität dieser neuen intelligenten Maschinen orientiert sich gegenwärtig vor allem an zwei Positionen. Auf der einen Seite wird nach Kriterien Ausschau gehalten, welche objektiv festlegen könnten, ob eine Maschine nun intelligent sei oder eben nicht. Hier werden dann entweder vorab festgelegte Tests herangezogen oder es wird nach Spezifika gesucht, die quasi ›in‹ der Maschine vorhanden sein müssten, damit wir von intelligenten Maschinen sprechen können. Diese stark ontologisch geprägte Position fragt

damit auf der Grundlage von Kriterienkatalogen bereits vor jeglichem Kontakt mit neuen Maschinen danach, ob diese nun tatsächlich intelligent *sind*. In dieser Hinsicht ist auch der Kybernetiker Heinz von Foerster (vgl. 1993: 357f.) auf die Ontologisierung von Maschinen angewiesen, wenn er in seinem Bild der nichttrivialen Maschinen den Fokus auf die *inneren* Mechanismen technischer Systeme legt und Entitäten anhand ihrer internen Funktionen in triviale und non-triviale Systeme unterscheidet. Wie aber könnte ein solcher Mechanismus offengelegt werden? Auch von Foerster ist somit grundlegend auf Zuschreibungen hinsichtlich der Tiefe einer Maschine angewiesen und übersieht damit die Relationierung an ihrer Oberfläche. Obschon Searle (1980) mit seinem Chinese-Room-Argument in erster Linie das Versprechen der *Strong AI* anzugreifen versuchte, dass Maschinen (irgendwann) »echte« Artificial General Intelligence (AGI) aufweisen könnten, kann sein Argument gleichwohl auf jegliche Form von KI übertragen werden (vgl. hierzu Bishop 2021: 19). Allerdings weckt dieses Beispiel den soziologischen Verdacht, dass hier die Rolle der Kommunikation unterschätzt oder gar ignoriert wird. In Searles ›Chinese Room‹ erscheint vor allem wichtig, was die (technische) Entität ist oder kann, und nicht, wie sie sich (kommunikativ) einbettet bzw. eingebettet wird.

Auf der anderen Seite dieser sehr an Ontologien orientierten Positionen finden sich Ansätze, die die Frage nach der Identität von Maschinen aus der Warte der praktischen menschlichen Zuschreibungsleistung zu beantworten versuchen. Hier wird der mehr oder weniger expliziten Identifikation von Agency Aufmerksamkeit geschenkt, die in Interaktionen oder anderen Formen des Austauschs zwischen Menschen und Maschinen zu beobachten sind. Prominente Vertreter*innen dieser Position sind etwa Hubert Knoblauch (2017), Sherry Turkle (2012) oder auch Werner Rammert und Ingo Schulz-Schaeffer (2002). Aus der Perspektive dieser praxeologisch-phänomenologischen Positionen reicht es aus, dass ein Mensch der Maschine Intelligenz zuschreibt, um deren Identität zu definieren. Zwar wird hier sehr wohl die Möglichkeit eingeräumt, dass auch die Aktivitäten von Maschinen als »intentional erfahren« (Knoblauch 2017: 160) und beispielsweise Roboter oder virtuelle Agenten »wie eine andere Person angesehen« (ebd.) werden können. Aus der externen Perspektive dieser Autorinnen und Autoren wird diese Erfahrung dann jedoch als illegitim gewertet und als bloße Projektion bzw. Als-ob-Zuschreibung eines Menschen betrachtet. Auch die Beschreibung als »Quasi-Anderer« oder »Quasi-Subjekt« (vgl. etwa Ihde 1990: 97ff.) ordnet die Qualifikation des Gegenübers nicht der Logik der Situation unter, sondern be-

ruht auf der Wahrnehmung und Zuschreibungsfähigkeit der daran beteiligten Menschen.

Beide Positionen bringen gravierende Schwierigkeiten mit sich. Die Perspektive der ontologisierenden Festlegung sucht nach einer Letztgültigkeit, die zumindest praktisch niemals eingelöst werden kann. In der Praxis eines Austauschs lässt sich nicht innehalten und gegenseitig ›unter die Haube‹ schauen, ob nun tatsächlich Intelligenz vorhanden ist oder nicht. Dazu sind Menschen auch im gegenseitigen Austausch nicht in der Lage, weshalb wir von sich mehr oder weniger bewährenden Erwartungen und Erwartungserwartungen ausgehen, wenn wir es mit anderen menschlichen Wesen zu tun haben. Ohne die Genese der Gedanken des anderen und auch von uns selbst erkennen und beurteilen zu können, gehen wir anhand der Wahrnehmungen und Beobachtungen an der Oberfläche des anderen in der Regel davon aus, es mit einem intelligenten Wesen zu tun zu haben. Dies bringt uns zur Kritik an der zweiten Perspektive, die von Projektionen und Zuschreibungen ausgeht. Unserer Einschätzung nach wäre es zu kurz gedacht, wenn es ausreichen würde, dass nur eine oder einer der Beteiligten darüber befinden dürfte, ob wir es – um eine Formulierung von Peter Fuchs (1991) aufzugreifen – mit einer zweiseitig oder bloß »einseitig intelligenten« Austauschbeziehung zu tun haben (vgl. zu dieser Kritik auch Lindemann 2009). Während also die erste Perspektive die teils grundlegende, aber praktisch immer vorhandene Intransparenz der beteiligten Austauschpartner ignoriert, lässt sich in der zweiten Perspektive eine einseitige Überhöhung einzelner menschlicher Parteien beobachten, die die Eigenlogik der Austauschbeziehung ignoriert (vgl. in gleicher kritischer Weise Muhle 2018; Müller 2022).

In unserem Beitrag möchten wir diese beiden Positionen um den Fokus auf die performativ koproduzierte *Relation* von Mensch und Maschine erweitern. Die relationale Position fragt danach, wie Identitäten im dynamischen Zusammenspiel von Relationen entstehen, zerbrechen oder sich verfestigen (vgl. u.a. White 1992; Luhmann 2017; Latour 2007; Goffman 2002). Auf diese Weise rückt die Ontogenese von Identitäten in den Blick, die sich stets in spezifischen Relationen realisiert. Mit dem Fokus auf die Ontogenese wenden wir den Blick von der Bestimmung eines tatsächlichen oder vielleicht nur fingierten Seins ab und richten das Augenmerk auf die *Rekonstruktion des Werdens* dieser Identitäten. Der Wechsel von der Ontologie zur Ontogenese führt damit die Beobachterperspektive mit. Dieser Wechsel löst zwar nicht das Problem ontologischer Setzungen, stellt jedoch genau diese Setzungen in den Vordergrund. Während

die klassische Ontologie von vordergründigen Seinsformen ausgeht, setzt die relationale Soziologie bei der situativen Etablierung jener Seinsformen an.

Die relationale Perspektive geht davon aus, dass alle Entitäten stets in einem Prozess des Werdens begriffen sind und nicht im Zustand einer feststehenden Eindeutigkeit identifiziert und letztgültig bestimmt werden können. Empirisch lässt sich zeigen, dass diese Prozesse des Werdens sowohl vielfältigen Kontrollversuchen ausgesetzt sind wie auch eigene Kontrollversuche darstellen, die den damit verbundenen Identitäten Halt geben oder ihnen Halt nehmen. Die relationale Position wechselt demnach die Perspektive: weg von dem Versuch einer Bestimmung des Seins und hin zu einer Rekonstruktion des Werdens.

Ein in dieser Hinsicht instruktiver Ansatz findet sich in Karafillidis' Vorschlag für »Relationsmustererkennungen«, der im Rahmen empirischer Studien zu Unterstützungssystemen in der Mensch-Maschine-Interaktion entwickelt wurde und sowohl an System- und Netzwerktheorien (Luhmann 1998; White 1992) als auch an andere relationale Soziologien (Emirbayer 1997) anschlussfähig ist. Karafillidis' Ansatz betont in eindrücklicher Weise die interdependente und performative Ontogenese sozialer Relationsmuster zwischen Mensch und Technik, die dann spezifische Identitäten hervortreten lässt. Vor einem breiten theoriegeschichtlichen Hintergrund erörtert Karafillidis zwei grundlegende Fragen relationaler Soziologie: Wie kann man Relationen erkennen und wie entstehen aus Relationen Identitäten?

Wie wir weiter unten genauer aufzeigen werden, lassen sich diese beiden prominenten Fragen mit bestehenden Instrumenten der soziologischen Theorie beantworten. Die Frage nach dem Erkenntnisprozess rekurriert dabei auf den operativen Konstruktivismus, der uns zeigt, wie Muster erkannt – und das heißt: rekonstruiert – werden. Die Rekonstruktion dieser Muster wiederum liefert Hinweise darauf, wie auch in anderen Situationen Relationsmuster erkannt werden könnten: »Relationen werden soziomateriell konstruiert und zu Mustern verdichtet, um dann auch in anderen Situationen als Beziehungen, Subjekte oder Objekte identifiziert und dadurch bestätigt zu werden.« (Karafillidis 2018: 106) Damit verbinden sich in relationalen Ansätzen Epistemologie und Ontologie: Es geht weder um eine Ontologie noch um einen Relativismus, sondern um *Ontogenese*, einen emergenten Prozess, der Erkennen und Sein gemeinsam hervorbringt.

2. Relationale Soziologien, relationale Identitäten

Die auch als »postphänomenologisch« titulierten relationalen Soziologien (vgl. Ihde 1990) versammeln sich unter der Gemeinsamkeit, dass sie – zumindest in ihrer radikaleren Version – eine konsequente *Entsubstanzialisierung* versuchen (vgl. Emirbayer 1997; Donati/Archer 2015; Schmidl 2022). Hier finden relationale Soziologien und neurokonstruktivistische Theorien zusammen: Sowohl Bewusstsein als auch Welt können immer nur als etwas in Erscheinung treten, das in einem Beobachter aufscheint. Subjekt und Objekt der Erkenntnis sind somit untrennbar miteinander verbunden; mehr noch: sie sind in ihrer Genese voneinander abhängig, weshalb nicht mehr nur eine der beiden Seiten der Unterscheidung allein berücksichtigt werden kann. In der Konsequenz lassen sich Epistemologie und Ontologie somit nicht mehr getrennt voneinander betrachten (vgl. Varela/Thompson/Rosch 1992; Vogd 2018).

Diese zentrale Entsubstanzialisierung bzw. Deontologisierung der relationalen Ansätze ist jedoch zugleich eine Hinwendung zur Relationalität von Subjekt und Objekt. Die entscheidende Wende zeigt sich in der Abwendung von einer einseitigen Überhöhung der Welt oder des Subjekts hin zu den diese erst in Erscheinung bringenden *Relationen*. Damit steht für relationale Soziologien die Kritik an der (methodologischen) Überhöhung des Individualismus sowie der Hypostasierung vermeintlich feststehender, bereits vor der Relationierung ›fertiger‹ Einheiten im Vordergrund. Relationale Soziologien betonen vielmehr den eigenständigen ontologischen Status sozialer Beziehungen im erweiterten Sinne: »Beziehungen und nicht die Akteure oder die Gesellschaft bzw. deren Bewusstsein etc. werden methodologisch als der zu untersuchende empirische Gegenstand der Soziologie angesehen.« (Seyfert 2019: 104) Hiermit ist eine prinzipielle Offenheit gegenüber den Grenzen des Sozialen verbunden: Wer oder was soziale Prozesse mitbestimmt, ist nicht ex ante festgelegt oder gar auf menschliche Akteure begrenzt. Darüber hinaus lässt sich aus dieser Warte auch mit »Passivitäten« (ebd.: 150ff.) des Sozialen rechnen, was eine Abwendung von dem Primat möglich macht, dass Soziales nur durch aktives Handeln und Agieren entstehen würde: »Der gemeinsame Konsens der Soziologie besteht in der Überzeugung, dass Sozialität stets irgendeine Form der Aktivität voraussetzt. Das Soziale muss aktiv gemacht werden, es geschieht nicht! Die innerdisziplinären Streitigkeiten beziehen sich dann allein auf die Frage, wer eigentlich die Trägerin der Aktivität ist: der menschliche Akteur, die Situation, das System oder doch das Netzwerk? Man hat es hier mit der Vorstellung eines unbelebten präsozialen Hintergrunds zu

tun, der erst durch die aktive Herstellung sozialer Beziehungen sozial belebt wird.« (Ebd.: 140)

Zusammenfassend lässt sich mit Karafillidis anmahnen, dass das Adjektiv ›relational‹ nicht einfach nur bedeutet, dass die Soziologie außer Handlungen, Akteuren, Normen, Rollen oder Institutionen nun auch Relationen beachten müsse. Die Betonung der Relationen soll vielmehr darauf aufmerksam machen, »dass prinzipiell keine soziale Einheit als selbstverständlich hingenommen werden kann und deshalb alle interessierenden Phänomene, also auch Handlungen, Akteure, Normen, Rollen oder Institutionen, als *Effekte einer bestimmten Relationierung von Relationen* begriffen werden müssen. Der primäre Fokus liegt dann nicht mehr auf Subjekten und Objekten oder auf Akteuren und Intentionen, sondern auf Relationen.« (Karafillidis 2010: 69; Hervorh. J. H. & M. L.)

Auch Robert Seyfert widmet sich der Ausformulierung einer relationalen Soziologie, die sich von der Exklusivität aktiv handelnder Menschen loslöst und nichtmenschliche Aktivitäten wie auch Passivitäten zu integrieren vermag. Aus einer solchen relationalen Perspektive wäre dann das Fundament des Sozialen jeder Interaktion stets vorgelagert: »So wie es keine Abwesenheit einer sozialen Ordnung gibt, gibt es streng genommen auch keine Abwesenheit sozialer Beziehungen. Statt ausgehend von dem Gegensatz von Ordnung vs. Nicht-Ordnung zu operieren, ist von einer fundamentalen Vielfalt sozialer Ordnungsbildungen und Beziehungen auszugehen. Wir befinden uns immer schon in einer Immanenz sozialer Beziehungen und selbst dann, wenn wir glauben, keine spezifischen sozialen Beziehungen aktiv zu unterhalten, sind wir zumindest passiv bzw. interpassiv in sie eingebunden.« (Seyfert 2019: 91) Darüber hinaus schließt sich Seyfert der prinzipiellen Erweiterung der Träger dieser Phänomene um Dinge, Tiere, Artefakte, Maschinen etc. an. Der Anthropozentrismus der Soziologie sei schon längst überholt, wie Seyfert ausgiebig kritisiert: Bereits bei den soziologischen Klassikern wie Parsons finde »sich eine anthropologisch-aktivistische Dublette, in der nur Akteure und Handeln für die Soziologie relevant sind und nur Menschen Akteure sein können. […] Demgegenüber ist es erst Jahrzehnte später zur Entwicklung einer […] Soziologie gekommen, […] die in der Lage ist, soziale Beziehungen zu Nicht-Menschen (Tieren, Dingen, Artefakten etc.) konzeptionell zu berücksichtigen. Die Kosten für diese nachholende Theorieentwicklung bestehen darin, diese Berücksichtigung mit einem *postsozialen turn* ankündigen und genuin soziale Beziehungen nun paradoxerweise als *postsoziale Beziehungen* bezeichnen zu müssen.« (Ebd.: 16; Hervorh. im Orig.)

Daher betrachten wir im Folgenden ausführlicher, wie Karafillidis die Ansätze der relationalen Soziologien für die Rekonstruktion der Ontogenese von Identitäten aufarbeitet. Wie wir sehen werden, steht für ihn die (relationale) Konstruktion von Mustern als Mustererkennung im Vordergrund.

Für Karafillidis (2019: 105) steht als Kriterium für das Programm einer relationalen Soziologie die Notwendigkeit eines theoretischen und methodischen Auflösungsvermögens im Vordergrund, das so gestaltet ist, »dass materielle Objekte und kognitive Subjekte, aber auch Relationen selbst als Muster von Relationen erkennbar werden« können. Hintergrund hierfür ist, dass die Relationalität dieser Perspektive im Prinzip eine »indeterministische Soziologie« (ebd.) darstellt, die das Prinzip der Unbestimmtheit als methodische Notwendigkeit erkennt. Es wäre blind gegenüber den eigenen Relationierungen, würde man diese prinzipielle Unbestimmtheit vorab durch epistemische Schnitte wie Subjekt/Objekt-Unterscheidungen oder theoretische bzw. methodische Restriktionen aufzulösen versuchen. Wenn nun Relationen und ihre Muster im Sinne von damit entstehenden Identitäten von Subjekten und Objekten im Fokus einer relationalen Soziologie stehen, dann stellt sich die Frage, wie diese Relationen eigentlich beobachtet und im Rahmen einer Ontogenese von Identitäten rekonstruiert werden können. In anderen Worten: Wie lassen sich Relationen erkennen? Und wie entstehen aus diesen Relationen Identitäten?

Für Karafillidis etwa geht es bei dem Programm der relationalen Soziologie »um die Gewinnung von *order* aus *noise*: Im Überschuss vorhandene, flüchtige, unbestimmte, ereignishafte Aktivitäten arrangieren sich in einer bestimmbaren Art und Weise immer wieder so, dass Identitäten (Relationen, Objekte, Subjekte) wiedererkennbar werden und wiederum Teil einer Sequenz werden können, Kontrollversuche starten, Anschlüsse erleichtern und eine Position im Feld einnehmen« (ebd.: 114; Hervorh. im Orig.). Zunächst prinzipiell noisehaft rauschende Aktivitäten werden also auf Relationsmuster hin beobachtbar, die dann als Identitäten erkannt werden können. Karafillidis misst der Zeitlichkeit, Fluidität und Rekombinationsfähigkeit im Rahmen der relationalen Ontogenese große Bedeutung zu. Als empirisch Forschende sind wir vertraut mit der Zuschreibung von Bedeutung auf Subjekte, Objekte oder soziale Verhältnisse in situ. Relationen sind immer raum-zeitlich eingebettet. Sie können zwar retrospektiv interpretiert oder prospektiv theoretisiert werden, ihren Eigensinn entfalten sie jedoch nur operativ in der jeweiligen Situation. Gerade deshalb erscheint uns die relationale Wendung als höchst anschlussfähig an ein praxistheoretisch informiertes empirisches Forschungsprogramm der Soziologie. Denn damit lassen sich die Muster und Mechanismen der sozioma-

teriellen Konstruktion von Relationen rekonstruieren, um zu beobachten, auf welche Art und Weise dies in homologen oder heterologischen Situationen auf ähnliche Weise geschieht oder sogar im Gegenteil unterbunden wird.

Der relationale Ansatz geht davon aus, dass die in den Relationen verdichteten Muster stets Auskünfte über die beteiligten und entstehenden Subjekte, Objekte und Beziehungen geben. Eine Beziehung existiert immer dann, »wenn sie von Beobachtern als Relation erkennend zustande gebracht und realisiert wird« (ebd.: 106). Getreu dem Prinzip des erkennenden Handelns müssen wir annehmen, dass auch kognitive Systeme nicht einfach eine gegebene Welt erkennen, sondern aus den zur Verfügung stehenden Daten jeweils eine neue Wirklichkeit erzeugen: »Jedes Tun ist Erkennen, und jedes Erkennen ist Tun.« (Maturana/Varela 1987: 146) Diese zentrale Einsicht des neurobiologischen Konstruktivismus wäre dann konsequenterweise auch auf maschinell realisiertes Erkennen anzuwenden.

Dieses Prinzip bringt eine bestimmte Haltung an die Oberfläche, die in wissenschaftlichen Programmen nur selten thematisiert wird: Es geht um das Explizieren der eigenen epistemischen Positionierung bzw. Relation im Gefüge der Welt. Karafillidis macht sich hier für eine epistemologische Selbstreflexion der wissenschaftlichen Praxis stark: »Für eine relationale Soziologie ist epistemologische Reflexivität dagegen typisch. Gemeint ist eine theoretische Berücksichtigung (a) der eigenen Perspektive und (b) der Tatsache, *dass* beobachtet wird, und zwar (c) sowohl *der* Forschungsgegenstand als auch (d) *im* Forschungsgegenstand.« (Karafillidis 2019: 108; Hervorh. im Orig.) Besonders der letzte Punkt grenzt die relationale Ontogenese und ihre methodologische Annäherung an Phänomene der Digitalisierung unserer Tage von anderen Ansätzen ab. Auch die Anwendung eines relationalen Paradigmas bringt eigene Relationen mit sich, die Identitäten in bestimmter Weise konditionieren oder verwerfen. Relationen sind somit weniger eigenständige, feststehende Elemente, sondern vielmehr Ereignisse bzw. »Operationen« (Luhmann 1998: 139) eines beobachtenden Elements im Gefüge der an einer Relationsmustererkennung beteiligten Relationen, Subjekte und Objekte. Erst die Operation der Beobachtung als »agentieller Schnitt« (Barad 2012: 19f.) bzw. unterscheidende Bezeichnung führt in die Welt die nötige Differenz ein, die dann als Muster und Identität in Erscheinung treten kann. Ohne Differenz keine Identität – »[e]xistence is selective blindness«, so George Spencer-Brown (2005: 192). Das Erkennen von Relationen ist somit immer ein Konstruktionsprozess, der als operative Aktivität des Unterscheidens und Bezeichnens beobachtete Muster

in Identitäten überführt und zugleich für andere Referenzen Relationierungen beobachtbar macht.

Wie Karafillidis festhält, steht dann die Frage im Raum, wie sich diese Identitäten bewähren, transformieren, verfestigen, wie sie zerbrechen und wie sie reformiert oder repariert werden könnten. Das Problem lautet somit: Wie kann die durch permanente Relationierung und damit verbundene Mustererkennung entstehende dynamische Stabilität erzeugt werden, die wir als soziale Welt erfahren?

Bei der Beantwortung dieser Frage stellt die relationale Soziologie konsequent auf ereignishafte Aktivitäten wie Konfirmierung, Destruktion oder Transformation von Relationsmustern ab, die ihrerseits zur Bildung von Identitätsmustern führen. Wie Karafillidis aufzeigt, haben relationale Ansätze verschiedene Bezeichnungen für diese Operationen entwickelt: *events* (Abbott 2001), *switching* (White 1995), *intra-action* (Barad 2012), *Kommunikation* (Luhmann 1998), *traduction* (Latour 2007), *transaction* (Emirbayer 1997) oder *pratique* (Bourdieu 1998). All dies sind nach Karafillidis (2019: 114) homologe Begriffe, »die ein momenthaft beobachtbares, komplexes empirisches Geschehen bezeichnen, aus dem Makrophänomene entstehen«.

Relativ stabile Identitäten werden dadurch als »ongoing accomplishment[s]« (Garfinkel 1967) beobachtbar, was vor allem dann auffällt, wenn es zu Zusammenbrüchen dieser Identitäten kommt. Während die Kondensierung einer Identität die Unbestimmtheit einer Situation in Bestimmtheit überführt und die Kontingenz der Relationsmustererkennung schließt, erzeugt der Zusammenbruch genau das Gegenteil: Durch die Hinterfragung einer Identität im Zuge einer weiteren Beobachtung, das heißt durch Reflexion, die etwa eine Mustererkennung überfordert und so zu ihrem Zusammenbruch führt, öffnet sich die Unbestimmtheit wieder und erzeugt damit zugleich neue Bestimmungspotenziale, die eine neue, andere Identität entstehen lassen. Beispiele hierfür finden sich etwa in Turkles (2012) Experimenten mit humanoiden Robotern, wo jene eben noch als plumpe Apparate aufgefasst wurden, dann aber plötzlich als leidensfähige Einheiten empfunden werden. Ähnliche Beispiele finden sich auch im Umgang mit computergesteuerten Spielpartnern (vgl. Harth 2014). Man beobachtet etwas, was daran zweifeln lässt, ob man sich durch eine bestimmte Identifizierung des Gegenübers nicht selbst belogen hat. Mit Harrison White (1995) wird hier deutlich, dass ein solcher Zusammenbruch von Identität nie zu einer *Nichtidentität* führt, sondern stets zu einer anderen Identität. Aus dem Zweifel an der bisherigen Identifizierung erwächst bereits die neue Identifizierung. An derartigen Wechseln (*switching*)

von Relationsmustern wird deutlich, dass jede Genese von Identitäten auf Zusammenbrüche angewiesen ist, »weil die situativ emergierenden Identitäten sich wechselseitig zu kontrollieren versuchen und dabei laufend Konstellationen entstehen, die auch scheitern können. Kontrolle in diesem Sinne erfordert kein reflektierendes Bewusstsein und auch keinen willentlichen Entschluss, sondern bezeichnet eine Form der affektiv-kommunikativen Verschränkung von Identitäten, also von Dingen, Artefakten, Tieren und Menschen unter ökologischen Bedingungen, die sie auch selbst mitgestalten.« (Karafillidis 2019: 115)

Spätestens damit wird deutlich, dass die stets immer nur vorübergehend stabilisierten bzw. sich stabilisierenden Relationsmuster nicht isoliert ›an und für sich‹ bestehen, sondern immer auf weitere Muster verweisen, zu denen übergangen werden kann. Durch die einbettenden Relationen werden die fluiden, flüchtigen und fragilen Identitäten quasi »gehärtet«, wie Bruno Latour (2007) dies in seinem netzwerktheoretischen Ansatz zur Ausbildung einer relationalen Soziologie herausarbeitet. Sie bekommen Halt und können sich genau hierzu verhalten bzw. damit relationieren. Interessant ist dabei der methodologische Hinweis, der in dieser Theoriefigur enthalten ist: Denn diese temporäre, das heißt immer nur ereignishafte Härtung von Identitäten und Relationen führt zu der praktisch permanenten »Herausforderung, in Situationen Halt zu finden, was körperlichen und materiellen Halt genauso einschließt wie sprachlichen oder kulturellen Halt. Die Haltung des Körpers, seine Stellung im Raum, ein einfaches Festhalten, die Nutzung von Artefakten oder situativ als passend beobachtete Ausdrucks- und Verhaltensweisen werden aufgerufen und angepasst, um Situationen zu bewältigen.« (Karafillidis 2019: 115) Erst die Kontrollversuche des Halt-Findens und Halt-Suchens in unbestimmten, aber bestimmbaren Ereignissen machen Subjekte, Objekte und Relationen identifizierbar. Das geschieht in den meisten Fällen unspektakulär und gleichsam en passant. Jede Ontogenese von Identität entsteht laut Karafillidis somit aus der Notwendigkeit, in Situationen Halt zu geben und zu finden. Die mit dieser Relationierung erfolgende temporäre Lösung der Kontingenz erzeugt damit ein Muster, das zur Identitätserzeugung herangezogen werden kann: »Sobald Beobachter etwas als Quelle einer Handlung betrachten und ihr Sinn zuschreiben, wird es zu einer Identität (White 2008: 2). Die ›Quelle‹ einer Handlung, von der White in dieser Definition spricht, wird situationsabhängig bestimmt.« (Karafillidis 2019: 115) Eine solche Definition von Identität schließt damit prinzipiell alle möglichen Entitäten (ob belebt oder unbelebt, menschlich oder nichtmenschlich etc.) ein, die in einer spezifischen Situati-

on als Subjekt, Objekt oder Relation Halt finden und geben und als Quellen bzw. Ursprünge dieser Versuche unterschieden werden können. Von der parallelen Verschränkung der Notwendigkeit und der Möglichkeit des Halt-Findens wie auch Halt-Gebens ist kein Beobachter ausgenommen: »Die Unsicherheit, Instabilität und Fragilität sozialer Beziehungen betrifft nie die soziale Immanenz selbst, nicht deren heterogene und mannigfaltige Ordnungen, sondern immer nur spezifische Erkenntnisperspektiven, die an Plausibilität gewinnen oder verlieren.« (Seyfert 2019: 19)

Eine derartige auf zeitliche Ereignisse fokussierte und relational informierte Soziologie gibt sich damit maximal inklusiv: »Diese Fassung von Identität beschränkt sich also nicht auf personale Identität, sondern kann auch menschliche Körper oder sogar nur einzelne Körperteile, aber auch Tiere, Artefakte, Objekte und Ereignisse bezeichnen, also alles, wovon sich beteiligte Beobachter affizieren lassen und es deshalb situativ als zeitstabil behandeln und erleben.« (Karafillidis 2019: 116) Spätestens hiermit sollte deutlich werden, dass der Identitätsbegriff der relationalen Soziologie kein anthropozentrischer ist. Daher ist aus dieser Perspektive dann auch nicht mehr entscheidend, ob eine der beteiligten Entitäten über ›echtes‹ Bewusstsein verfügt oder eine Intention zum Handeln verspürt – wie könnte dies in der jeweiligen Situation auch überprüft werden? –, vielmehr lassen sich *alle möglichen Relata*, »deren Relevanz für eine Situation oder ein Phänomen empirisch wiederum nur von Beobachtern bestimmt werden[sic!], […] als solche *White/Latour-Identitäten* beschreiben« (ebd.; Hervorh. J. H. & M. L.).

3. Relationen zwischen Menschen und Maschinen

Ist damit alles nur eine Frage des Beobachters?[1] Vielleicht haben wir es bei der Diskussion um die (vermeintliche) Intelligenz sogenannter Künstlicher Intel-

1 Diese durchaus lakonisch-relativierende Frage muss gestellt werden, insbesondere, wenn man – wie Dirk Baecker (2013) – die Figur des Beobachters als eine der zentralen Entdeckungen des 20. Jahrhunderts ansehen möchte: »Wenn es in diesem Jahrhundert so etwas wie eine zentrale intellektuelle Faszination gibt, dann liegt sie wahrscheinlich in der Entdeckung des Beobachters. Es ist schwer zu sagen, ob die beiden anderen großen Theoriethemen dieses Jahrhunderts, die Sprache und die Selbstreferenz, Voraussetzung oder Folgen dieser Entdeckung sind. Noch schwerer wäre inzwischen die Frage zu entscheiden, welche Wissenschaften tiefer in sie verstrickt sind, Physik, Biologie, Psychologie oder Soziologie. Es kommt auch nicht darauf an, diese Frage zu

ligenz ja überhaupt nur mit der Produktion und Beteiligung neuer (anderer) Beobachter zu tun, die ihre Beobachtungen in die Welt einspeisen und genau damit das Spiel der Relationierung(en) der Gesellschaft verändern – beispielsweise, indem die abstrakte und eher neutrale Theoriefigur des ›Beobachters‹ plötzlich als genderspezifische Figur verstanden wird. Gerade hieran wird etwa deutlich, wie (stets beobachtete) Identitäten es ihren Beobachtern ermöglichen, in unsicheren Situationen Halt zu finden. Möglich werden dadurch dann weitere und teils neue Sinnanschlüsse, die in nachfolgenden Situationen ebenfalls neu konditionierte Relationen ermöglichen, wie zum Beispiel die Hinterfragung der Beobachtung dieser Identität(en).

Es sollte dabei nicht übersehen werden, dass auch die Figur des Beobachters selbst immer in Relationsmuster eingebettet ist und von diesen in seiner*ihrer Identität kontrolliert wird. Denn auch wir Autoren und Lesende sind als Beobachter*innen unsererseits nichts anderes als Relationsmuster, die sich bestimmten Beziehungen verdanken. Beobachter*innen werden beobachtet und können kaum anders, als in und mit ihren Beobachtungen auch die Beobachtungen anderer zu berücksichtigen. Auch die Identität und Kontrolle der Beobachter unterliegt der dynamischen Stabilität von Relationsmustern. Empirisch gesehen wird somit jedes Ereignis, jede Identitätsfindung, jede Relationierung immer durch ein bereits vorab bestimmtes Gefüge von Relationen kontrolliert, das in vorherigen Situationen zustande kam. Dabei ist die Selbstbezüglichkeit von Identitätsfindungen, wie sie in der Metaphorik des Halt-Findens und Halt-Gebens angedeutet wird, nicht zu unterschätzen. Denn jede Mustererkennung erfolgt zwangsläufig selbst in bestimmten Mustern: »Empirisch lässt sich also nur von bereits generierten Identitäten und Kontrollformen ausgehen, zwischen denen umgeschaltet wird und die als Ressource weiterer Ontogenese dienen. Mustererkennung ist nicht möglich, ohne wiederum auf andere Muster zurückzugreifen, die der aktuellen Mustererkennung als notwendiger Kontext dienen […].« (Karafillidis 2019: 117) Nicht nur als Forschende, sondern ganz prinzipiell als Lebende eröffnet sich uns damit stets die aktualisierbare selbstreflexive Frage: Welche Relationsmuster konditionieren welche Relationsmustererkennung?

Im Kontext dieser beobachtungstheoretischen Reflexion kann Realität nicht als eine beobachterunabhängige externe Instanz aufgefasst werden. Gleichzeitig kann Realität jedoch auch nicht als bloßes internes Bezugsfeld

---

entscheiden. Nur ein Beobachter könnte sie entscheiden, und ein anderer Beobachter hätte dann Anlaß zurückzufragen […].« (Baecker im Vorwort zu von Foerster 1992: 17)

von Aussagen, als solipsistisches Phantasma, gefasst werden. Vielmehr ist Realität als Korrelat der gegenseitig verschränkten Beobachtung von Beobachtern zu verstehen, wie Niklas Luhmann einprägsam darstellt: »Was die Kybernetik des Beobachtens neu anbietet, ist die zirkuläre Geschlossenheit des Beobachtens von Beobachtungen. Wenn ein System sich auf dieser Ebene konstituiert und eine Zeitlang in Betrieb ist, kann man schließlich nicht mehr unterscheiden(!), wer der ›wirkliche‹ Beobachter ist und wer sich nur anhängt. *Alle Beobachter gewinnen Realitätskontakt nur dadurch, daß sie Beobachter beobachten*.« (Luhmann 1992: 97; Hervorh. J. H. & M. L.) Sowohl das menschliche wie auch das künstliche Beobachtungsvermögen speist sich jeweils aus einer durch Relationsmuster kontrollierten Wahrscheinlichkeitswolke weder notwendiger noch unmöglicher Aktivitäten oder Passivitäten.

Wir Menschen können nicht anders, als mit Intransparenz zu rechnen. Erstens rechnen wir ganz im Wortsinne mit ihr, da wir nicht anders können, als retrospektiv zuzuschreiben, wie jemand (ich, du, wir) zu den Sinndeutungen gekommen sein mag, die beobachtet werden (vgl. Sutcliffe/Weick 2005). Wir rechnen mit Intransparenz, indem wir sie qua Relationsmustererkennung beobachtend konstruieren, und das heißt im Kontext dieses Texts als Identität identifizieren. Zweitens rechnen wir mit Intransparenz in dem Sinne, dass wir diese Erwartungen an nun identifizierte andere Quellen von Handlungen in unser eigenes Handeln einweben. Die Problemlösung der sprachlichen Kommunikation und ihrer Ausdifferenzierung in Mündlichkeit, Schriftlichkeit und Buchdruck, die auf die prinzipielle Differenz zwischen Ich und Du reagiert, führte bekanntlich zu historischen Komplexitätssteigerungen der Gesellschaft als Gesamtsystem (Watzlawick/Jackson/Lederer 1967: 62ff.). Das Rechnen mit den teils überraschenden Sinnemissionen intelligenter Maschinen wird diese Komplexitätssteigerungen vermutlich nochmals perpetuieren. Denn unsere analogen Umwelten werden durch die weitere Digitalisierung der Welt mittels ihrer Datafizierung (Häußling 2022) und durch die darauf basierenden technischen Ausformungen in Gestalt neuer Bilder, Töne, Grafiken und Texte angereichert. Im Hinblick auf die Relation von Menschen und Maschinen geht es dann wohl zunehmend noch stärker um relationierende Abhängigkeiten und Möglichkeiten, die jeweils neue, anders strukturierte und strukturierende Beobachter der Welt generieren.

Spätestens damit wird der Mensch als ein Relationswesen in einem Dazwischen (als *interrelation*; vgl. Ihde 1990)[2] rekonstruierbar, das ununterbrochen auf der Suche nach geeigneten Relata ist, an denen es sich selbst spüren, vergewissern sowie herausfordern kann, um durch die Relationsmustererkennung anderer seiner eigenen Identität Halt zu geben. Interessanterweise wählt der Religionsphilosoph Martin Buber (2008: 4) einen hiermit korrespondierenden Ausgangspunkt für seine Philosophie: »Der Mensch wird am Du zum Ich. […] Es gibt kein Ich an sich, sondern nur das Ich des Grundworts Ich-Du und das Ich des Grundworts Ich-Es. Wenn der Mensch Ich spricht, meint er eins von beiden.« Obgleich Buber für gewöhnlich nicht zur Riege der postsozialen Soziologen gezählt wird, zeigt sich hier ein ähnlicher Versuch, das Du prinzipiell relational zu verstehen und damit die Bestimmung von Ich und Du in der Form eines Zwischen aufzulösen. Wie für Ihde kommt auch für Buber das Ich erst am *Nicht-Ich* zu sich, nämlich im Dialogischen der gemeinsamen Interaktion mit anderen.

Die allzu menschlichen Fragen nach der Bestimmung der Identität von Maschinen scheinen daher einer starken »Alteritäts-Bedürftigkeit« (Müller 2022: 27) zu entspringen. Das menschliche Ich erscheint für seine Reproduktion auf ein Du oder ein Es angewiesen, wodurch andere Relata an Relevanz für die eigene Relationierung gewinnen. Unsere Vermutung ist, dass diese Bedürftigkeit nach einer Bestimmung Anderer deckungsgleich ist mit der Bedürftigkeit nach Selbstvergewisserung (vgl. Harth/Feißt 2022). Die Frage, ob intelligente Maschinen nun als eine Kopie des Menschen verstanden werden können, als dinghafte Apparatur oder als eigenständige Entität, verliert aus dieser Perspektive an Relevanz (vgl. Harth 2021). In den Vordergrund rückt hingegen der Blick für die Relationen, die in ihrer praktischen Ausgestal-

2 Von Don Ihde (1990) wurden bekanntlich vier verschiedene Relationen definiert, von denen uns im Kontext dieses Beitrags vor allem die dritte am stärksten interessiert. Neben der *embodiment relation* (Beispiel Brille), der *hermeneutic relation* (Beispiele Geigerzähler, CERN) und der *background relation* (Beispiel Stromversorgung) erscheint uns die *alterity relation* am passendsten für unsere Frage nach der Beziehung zwischen dem Selbst und dem Anderem. Unter dieser Relation versammelt Ihde diejenige Technik, die ganz im Sinne Levinas' als konkrete*r Andere*r in Erscheinung tritt. Beispiele hierfür wären Artefakte, die als heilig angesehen werden, oder auch Roboter bzw. *virtual humans*, die in kommunikativen Situationen Menschen gegenübertreten. Entscheidend ist dabei, dass diese Art der Technik als Andere*r die Möglichkeit zur Selbstreflexion bietet, zum Spiegeln der eigenen Selbst- und Weltverhältnisse.

tung darüber entscheiden, *wie* und damit auch *als was* die jeweiligen Relata identifiziert und behandelt werden.

Zusammenfassend »besteht das relationale Forschungsprogramm einer Relationsmustererkennung aus den theoretisch-methodischen Aufforderungen, (a) eine mikrologische, operative Perspektive einzunehmen, um die Genese von Relationen, Subjekten und Objekten als komplexe Elemente zu erkennen, und (b) beschreibend nachzuvollziehen, wie die empirisch relevanten bzw. adressierten Identitäten durch eine selektive Kondensierung und generalisierende Konfirmierung spezifischer Operationen hervorgebracht werden« (Karafillidis 2019: 118). Damit lässt sich festhalten, dass eine radikale relationale Soziologie nicht nur die Netzwerke sozialer Verflechtungen (zwischen Menschen) untersucht, sondern prinzipiell alle empirisch beobachtbaren Entitäten, die jeweils situativ Quellen von Kontrollversuchen (welcher Form auch immer) sein können, als Relationsmuster ernst nimmt, um sie entsprechend zu de- und rekonstruieren.

4. Fragen an eine relationale Methodologie

Was bedeutet dies nun alles für die Frage nach der Relation von Mensch und Maschine? Wie lässt sich die Identität der an sozialen Situationen Beteiligten bestimmen? Wie ist Relationsmustererkennung möglich?

Aus der hier vorgestellten mikrologischen Theorieperspektive der relationalen Soziologie werden zunächst zwei Aspekte deutlich: Erstens wird ganz praktisch offenkundig, dass Erkennen und Handeln zwei Seiten einer Medaille sind. Zweitens lässt sich erkennen, dass wir es mit einer sich permanent dynamisch stabilisierenden Ontogenese von Identitäten zu tun haben: Werden und Wandel sind unvermeidliche Prozesse des Sozialen und damit der permanente Normalfall.

Aus dieser relationalen Perspektive gilt es dann, empirisch zu rekonstruieren, wie das, was beobachtet werden kann, in der Situation jeweils hergestellt wird. Nur in der Situation zeigt sich, mit welcher Relationsmustererkennung welche Entität unterschieden und identifiziert wird und wie diese Erkennung sich wiederum mit anderen Relationierungen relationiert. Nur *in der Situation* werden jeweils konkrete Identitäten vorgehalten, verworfen oder (re-)aktiviert.

Dies wiederum eröffnet wichtige Anschlüsse an die rekonstruktive Sozialforschung, die sich in ihrem Programm ebenfalls der Rekonstruktion von Mus-

tern und Typiken sozialer Praxis widmet. Auch jede Sozialforschung kann als spezielle Form einer Relationsmustererkennung verstanden werden, die als eigene Praxis jeweils relationiert und identifiziert. So benennt etwa die Ethnomethodologie einen ähnlichen Sachverhalt, wenn sie die Herstellung und Reproduktion sozialer Formen in den Blick nimmt und untersucht, wie sich Akteure dabei wechselseitig unterstützen, kontrollieren und normalisieren (vgl. Garfinkel 1967).

Die konkrete Anwendung einer derartigen relationalen Methodologie auf Phänomene der Ontogenese von Identitäten in Mensch-Maschine-Relationen steht leider noch aus. Daher handelt es sich bei den folgenden Anwendungsbeispielen um theoretische Skizzen: In Bezug auf die Technologie des ›Predictive Policings‹ würde dieser Forschungsansatz mindestens zweierlei nahelegen. Erstens würde er danach fragen, welche Kontrollversuche dieser Technologie an ihren Bildschirmen unterstellt werden können. Um die relationale Identitätsgenese dieser Technologie zu untersuchen, müssten zweitens die jeweiligen lokalen Situationen der Einbringung dieser Bildschirme und ihrer Anzeigen daraufhin befragt werden, wie auf diese neue Quelle von Aktivitäten und Kontrollversuchen reagiert wird. Wie werden die Kontrollversuche der KI des Predictive Policings im Netzwerk von anderen Entitäten dafür genutzt, um Halt zu finden? Welche Inklusion, aber auch welche Exklusion der KI findet mittels welcher Identitätsbildung statt? Welche lokal zu rekonstruierenden ›new orders of policing‹ entsthen durch den ›noise‹ des Eintritts neuer KI-Policing-Werkzeuge?

Die eher theoretischen Ausführungen zur relationalen Soziologie, wie sie hier wiedergegeben wurden, weisen unseres Erachtens eine direkte methodologische Komponente auf, die für weiterführende Untersuchungen einer relationalen Soziologie der Künstlichen Intelligenz herangezogen werden kann. Denn der mikrologische Blick auf Ereignisse und Situationen, die Beteiligten, ihre Beobachtungen und Gegenbeobachtungen, die impliziten und expliziten Attributionen von Subjektivität, Objektivität und Relationalität, das aktive oder passive Gewähren, Verhindern oder Produzieren von weiteren Ereignissen wie auch auf die Repetition, Rejektion, Destruktion, Reparatur oder Transformation bestehender Relationsmuster sollte mehr als genug Hinweise dafür bieten, wie bestimmte Relationen erkannt werden können, die wiederum bestimmte Identitäten hervorbringen.

Darüber hinaus lassen sich vor allem die mittlerweile als klassisch titulierten ethnomethodologischen Forschungsprogramme mit ihrer aus heutiger Sicht visionären Offenheit für die Operationalisierung eines relationalen For-

schungsprogramms nutzen. Nicht nur Garfinkel, sondern auch Goffman zeigte ein großes Interesse an sozialen Attributionen, Kipppunkten und Neurahmungen, die in ihrer Relationierung den Sinn der Situation erst herzustellen vermögen. Dies hat auch Karafillidis mit Bezug auf Goffman (2000) erkannt: »Genau an den Stellen, an denen ein Umschalten zwischen (Rollen-)Identität, Kontrollformen und Netzwerk-Domänen stattfindet, also zum Beispiel beim Wechsel von der Hinter- auf die Vorderbühne […], werden Identitäten als relationale Muster methodisch greifbar. Es sind diejenigen Momente, in denen sie empirisch *konfirmiert*, also bestätigt oder aufgegeben werden. Ähnlich ist es in Krisenmomenten, denen Goffman ebenfalls große Aufmerksamkeit schenkt.« (Karafillidis 2019: 118; Hervorh. im Orig.) Exakt an diesen Stellen müsste ein methodologisches Forschungsprogramm ansetzen, das auf die Rekonstruktion der Ontogenese sozialer, subjektiver oder objektiver Identitäten abzielt (Gruppen, Personen, Orte, Dinge, Zeiten etc.). Goffmans (2000) Fokus auf die Relevanz der Bühnen zur Selbst- und Weltdarstellung kann hier als homologe Anweisung für die Beobachtung der Rekonstruktion von Relationsmustern gesehen werden. Auf diese Weise dürfte es einer empirischen Soziologie, die sich für die Relation von Mensch und Maschine interessiert, wieder gelingen, strikt soziologisch vorzugehen, was hier heißt, »von sozialen Beziehungen her zu denken. Prägend für soziale Beziehungen sind dann nicht die individuellen Motive oder die Kollektivtatsachen, sondern das Zwischen, d.h. die Beziehungen, Wechselwirkungen und wechselseitigen Affizierungen.« (Seyfert 2019: 20) Einer an der relationalen Soziologie geschulten und ethnomethodologisch informierten empirischen Sozialforschung dürfte es dann nicht mehr schwerfallen, die sich in Situationen verdichtenden Relationen genau rekonstruieren zu können, um zu identifizieren, wie sich durch Wiederholungen, Konfirmationen oder Rejektionen die jeweils beteiligten Identitäten temporär kondensieren. Nicht zuletzt können wir aus einer solchen Auseinandersetzung mit maschinell durchsetzten sozialen Verhältnissen einen neuen Blick auf die relationalen Ontogenesen von uns schon lange bekannten Referenzen gewinnen. Dann fällt vielleicht auf, dass wir mitunter auch Menschen als mehr oder minder triviale Maschinen behandeln. Und nicht zuletzt lässt sich die ethische Frage stellen, ob wir ihnen damit gerecht werden.

Literatur

Abbott, Andrew. 2001. *Time Matters: On Theory and Method*. Chicago: The University of Chicago Press.

Baecker, Dirk. 2013. *Beobachter unter sich. Eine Kulturtheorie*. Berlin: Suhrkamp.

Barad, Karen. 2012. *Agentieller Realismus. Über die Bedeutung materiell-diskursiver Praktiken*. Berlin: Suhrkamp.

Bishop, J. Mark. 2021. Artificial Intelligence Is Stupid and Causal Reasoning Will Not Fix It. *Frontiers in Psychology* 11: 2603.

Bourdieu, Pierre. 1998. *Praktische Vernunft. Zur Theorie des Handelns*. Frankfurt a.M.: Suhrkamp.

Buber, Martin. 2008. *Ich und Du*. Stuttgart: Reclam.

Burrell, Jenna. 2016. How the Machine ›Thinks‹: Understanding Opacity in Machine Learning Algorithms. *Big Data & Society.* https://doi.org/10.1177/2053951715622512.

Coeckelbergh, Mark. 2011. Human, Animals, and Robots: A Phenomenological Approach to Human Robot Relations. *International Journal of Social Robotics* 3, H. 2: 197–204.

Donati, Pierpaolo und Margaret S. Archer. 2015. *The Relational Subject*. Cambridge: Cambridge University Press.

Donati, Pierpaolo. 2010. *Relational Sociology: A New Paradigm for the Social Sciences*. London und New York: Routledge.

Emirbayer, Mustafa. 1997. Manifesto for a Relational Sociology. *American Journal of Sociology* 103, H. 2: 281–317.

Esposito, Elena. 2017. Artificial Communication? The Production of Contingency by Algorithms. *Zeitschrift für Soziologie* 46, H. 4: 249–265.

Fuchs, Peter. 1991. Kommunikation mit Computern? Zur Korrektur einer Fragestellung. *Sociologia Internationalis* 29, H. 1: 1–30.

Fuhse, Jan und Sophie Mützel. (Hg.). 2010. *Relationale Soziologie. Zur kulturellen Wende der Netzwerkforschung*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Garfinkel, Harold. 1967. Common Sense Knowledge of Social Structures: The Documentary Method of Interpretation in Lay and Professional Fact Finding. In *Studies in Ethnomethodology,* 76–103. Englewood Cliffs: Prentice Hall.

Goffman, Erving. 2000 [1959]. *Wir alle spielen Theater. Die Selbstdarstellung im Alltag*. München: Piper.

Goffman, Erving. 2002 [1977]. *Rahmen-Analyse. Ein Versuch über die Organisation von Alltagserfahrungen*. Frankfurt a.M.: Suhrkamp.

Harth, Jonathan und Martin Feißt. 2022. Neue soziale Kontingenzmaschinen. Überlegungen zu künstlicher sozialer Intelligenz am Beispiel der Interaktion mit GPT-3. In *Begegnungen mit künstlicher Intelligenz. Intersubjektivität, Technik, Lebenswelt*, Hg. Martin Schnell und Lukas Nehlsen, 70–103. Weilerswist: Velbrück.

Harth, Jonathan. 2014. *Computergesteuerte Spielpartner. Formen der Medienpraxis zwischen Trivialität und Personalität.* Wiesbaden: Springer VS.

Harth, Jonathan. 2021. Simulation, Emulation oder Kommunikation? Soziologische Überlegungen zu Kommunikation mit nicht-menschlichen Entitäten. In *Intersoziologie. Menschliche und nichtmenschliche Akteure in der Sozialwelt,* Hg. Michael Schetsche und Andreas Anton, 143–158. Weinheim und Basel: Beltz Juventa.

Häußling, Roger. 2020. Daten als Schnittstellen zwischen algorithmischen und sozialen Prozessen: Konzeptuelle Überlegungen zu einer Relationalen Techniksoziologie der Datafizierung in der digitalen Sphäre. In *Soziologie des Digitalen – Digitale Soziologie? Soziale Welt, Sonderband 23,* Hg. Sabine Maasen und Jan-Hendrik Passoth, 134–150. Baden-Baden: Nomos.

Humphreys, Mark. 2009. How My Program Passed the Turing Test. In *Parsing the Turing Test: Philosophical and Methodological Issues*, Hg. Robert Epstein, Gary Robert und Grace Beber, 237–260. New York: Springer.

Ihde, Don. 1990. *Technology and the Lifeworld: From Garden to Earth*. Bloomington: Indiana University Press.

Karafillidis, Athanasios. 2010. Grenzen und Relationen. In *Relationale Soziologie. Zur kulturellen Wende der Netzwerkforschung,* Hg. Jan Fuhse und Sophie Mützel, 69–95. Wiesbaden: VS Verlag für Sozialwissenschaften.

Karafillidis, Athanasios. 2018. Relationsmustererkennung. Relationale Soziologie und die Ontogenese von Identitäten. *Berliner Debatte Initial* 29, H. 4: 105–125.

Knoblauch, Hubert. 2017. *Die kommunikative Konstruktion der Wirklichkeit*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Knorr-Cetina, Karin. 2006. Sozialität mit Objekten. Soziale Beziehungen in post-traditionalen Wissensgesellschaften. In *Zur Kritik der Wissensgesellschaft,* Hg. Dirk Tänzler, Hubert Knoblauch und Hans-Georg Soeffner, 101–138. Konstanz: UVK.

Latour, Bruno. 2007. *Eine neue Soziologie für eine neue Gesellschaft. Einführung in die Akteur-Netzwerk-Theorie*. Frankfurt a.M.: Suhrkamp.

Latour, Bruno. 2020. *Kampf um Gaia. Acht Vorträge über das neue Klimaregime*. Berlin: Suhrkamp.

LeCun, Yann, Yoshua Bengio und Geoffrey Hinton. 2015. Deep Learning. *Nature* 521: 436–444. https://doi.org/10.1038/nature14539.

Lemoine, Blake. 2022. Is LaMDA Sentient? – an Interview. *Medium*. https://www.documentcloud.org/documents/22058315-is-lamda-sentient-an-interview. Zugegriffen: 17. Oktober 2022.

Lindemann, Gesa. 2009. *Das Soziale von seinen Grenzen her denken*. Weilerswist: Velbrück.

Luhmann, Niklas. 1992. *Die Wissenschaft der Gesellschaft*. Frankfurt a.M.: Suhrkamp.

Luhmann, Niklas. 1998. *Die Gesellschaft der Gesellschaft*. Frankfurt a.M.: Suhrkamp.

Luhmann, Niklas. 2017. *Die Kontrolle von Intransparenz*. Berlin: Suhrkamp.

Maturana, Humberto R. und Francisco J. Varela. 1987. *Der Baum der Erkenntnis. Wie wir die Welt durch unsere Wahrnehmung erschaffen – die biologischen Wurzeln des menschlichen Erkennens*. Bern: Scherz.

McFarlane, Craig. 2013. Relational Sociology, Theoretical Inhumanism, and the Problem of the Nonhuman. In *Conceptualizing Relational Sociology: Ontological and theoretical Issues*, Hg. Christopher Powell und François Dépelteau, 45–66. New York: Palgrave Macmillan.

Muhle, Florian. 2018. Sozialität von und mit Robotern? Drei soziologische Antworten und eine kommunikationstheoretische Alternative. *Zeitschrift für Soziologie* 47, H. 3: 147–163.

Müller, Oliver. 2022. Maschinelle Alterität. Philosophische Perspektiven auf Begegnungen mit künstlicher Intelligenz. In *Begegnungen mit künstlicher Intelligenz. Intersubjektivität, Technik, Lebenswelt*, Hg. Martin Schnell und Lukas Nehlsen, 23–47. Weilerswist: Velbrück.

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

Nassehi, Armin. 2019. *Muster. Theorie der digitalen Gesellschaft*. München: C. H. Beck.

Piloto, Luis S., Ari Weinstein, Peter Battaglia und Matthew Botvinick. 2022. Intuitive Physics Learning in a Deep-Learning Model Inspired by Developmental Psychology. *Nature Human Behavior*. https://doi.org/10.1038/s41562-022-01394-8.

Rammert, Werner und Ingo Schulz-Schaeffer. 2002. Technik und Handeln. Wenn soziales Handeln sich auf menschliches Verhalten und technische

Abläufe verteilt. In *Können Maschinen handeln?,* Hg. Werner Rammert und Ingo Schulz-Schaeffer, 11–64. New York und Frankfurt a.M.: Campus.

Schmidl, Alexander. 2022. *Relationen. Eine postphänomenologische Soziologie der Körper, Technologien und Wirklichkeiten*. Weilerswist: Velbrück.

Schrittwieser, Julian et al. 2020. Mastering Atari, Go, Chess and Shogi by Planning With a Learned Model. *Nature* 588: 604–609.

Searle, John. 1980. Minds, Brains and Programs. *Behavioral and Brain Sciences* 3, H. 3: 417–457.

Seyfert, Robert. 2019. *Beziehungsweisen. Elemente einer relationalen Soziologie.* Weilerswist: Velbrück.

Spencer-Brown, George. 2005. *Laws of Form. Gesetze der Form*. Lübeck: Bohmeier.

Sutcliffe, Kathleen M. und Karl E. Weick. 2005. Organizing and the Process of Sensemaking. *Organization Science* 16, H. 4: 409–421.

Turing, Alan. 1950. Computing Machinery and Intelligence. *Mind* 59: 433–460.

Turkle, Sherry. 2012. *Verloren unter 100 Freunden. Wie wir in der digitalen Welt seelisch verkümmern*. München: Riemann.

Varela, Francisco J. 1990. *Kognitionswissenschaft – Kognitionstechnik. Eine Skizze aktueller Perspektiven*. Frankfurt a.M.: Suhrkamp.

Varela, Francisco J., Evan Thompson und Eleanor Rosch. 1992. *Der mittlere Weg der Erkenntnis. Der Brückenschlag zwischen wissenschaftlicher Theorie und menschlicher Erfahrung – die Beziehung von Ich und Welt in der Kognitionswissenschaft*. Bern: Scherz.

Vogd, Werner. 2018. *Selbst- und Weltverhältnisse. Leiblichkeit, Polykontexturalität und implizite Ethik*. Weilerswist: Velbrück.

Von Foerster, Heinz. 1993. Mit den Augen des anderen. In *Wissen und Gewissen. Versuch einer Brücke*, Hg. Siegfried J. Schmidt, 350–363. Frankfurt a.M.: Suhrkamp.

Watzlawick, Paul, Don D. Jackson und William J. Lederer. 1967. *Pragmatics of Human Communication: A Study of Interactional Patterns, Pathologies and Paradoxes*. New York: W. W. Norton & Co.

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. *Social Research* 62, H. 4: 1035–1063.

Wie die Bildung pragmatischer Handlungsmuster die Mensch-Maschine-Kommunikation gestaltet

Yaoli Du & Nadine Schumann

Abstract: *In der Debatte über die Kommunikation von Menschen und Maschinen werden verschiedene Definitionen und Bedeutungen der Kernbegriffe Information und Interaktion diskutiert. Die Erwartungen an künstlich-intelligente Systeme beziehen sich einerseits maßgeblich auf funktionalistische Abläufe, wobei von einer Vergleichbarkeit der Informationsverarbeitungsprozesse von Mensch und Maschine ausgegangen wird. Mit dieser Gleichsetzung werden andererseits aber falsche Erwartungen produziert und der Intelligenzbegriff wird von menschlichen Handlungen in ihren sozialen Praxisformen abgelöst. Um die Mensch-Maschine-Kommunikation adäquat zu beschreiben, muss diese Ablösung vermieden werden, denn eine rein syntaktische Beschreibung des Kommunikationsprozesses als Informationsverarbeitung reicht nicht aus, um soziale Praxisformen zu interpretieren. Um zu verstehen, wie der semantische Gehalt von Information durch Gebrauch generiert wird, schlagen wir einen pragmatischen interaktiven Ansatz vor. Wir untersuchen, wie Bedeutung in sozialen Interaktionen entsteht und in Form von pragmatischen Handlungsmustern durch stetigen Gebrauch stabilisiert wird. Ausgehend von Erkenntnissen der Entwicklungspsychologie beschreiben wir die zwischenmenschliche Kommunikation und extrahieren wesentliche Aspekte, wie zum Beispiel die Notwendigkeit einer gemeinsam geteilten Welt, die auch für die Kommunikation mit Maschinen relevant sind. Mit dem Leitprinzip der Triangulation können wir nicht nur die Entwicklung sozialer Kognition illustrieren, sondern erhalten auch wichtige Impulse für den Entwicklungsprozess einer gelingenden Kommunikation zwischen Menschen und Maschinen.*

1. Einleitung

Im Zeitalter der Digitalisierung verändert sich nicht nur die Kommunikation von Menschen untereinander, sondern auch unser Umgang mit digitalen Artefakten. Im Bereich der Mensch-Maschine-Interaktion und besonders der sozialen Robotik wird erwartet, dass Maschinen die Rolle von echten Kommunikationspartner:innen übernehmen können (vgl. Breazeal/Dautenhahn/Kanda 2016).[1] Entscheidend ist in diesem Zusammenhang, welche Bedingungen für das Gelingen von Kommunikation erfüllt sein müssen. Wir zeigen im Folgenden, dass eine erfolgreiche Kommunikation von Mensch und Maschine die Entwicklung eines gemeinsamen Bezugsrahmens erfordert, der sowohl materielle als auch soziale Bedingungen vereint. Dieser gemeinsame Bezugsrahmen wird durch triadische zwischenmenschliche Interaktionen generiert und beeinflusst wiederum unsere Kommunikation untereinander.

An dieser Stelle wird deutlich, dass das klassische Informationskonzept der Kommunikationstheorie – die unvermittelte Dyade von Sender und Empfänger – nicht ausreicht, um die Mensch-Maschine-Kommunikation adäquat zu beschreiben. Um überhaupt zu verstehen, wie der semantische Gehalt von Information durch Gebrauch generiert wird, stellen wir das Leitprinzip der Triangulation aus der sozialen Kognitionsforschung vor, mit dem sich wichtige Impulse für den Entwicklungsprozess einer gelingenden Kommunikation zwischen Menschen und Maschinen gewinnen lassen. Unter den heutigen technologischen Bedingungen wird der Bezugsrahmen von Kommunikation zwischen künstlich-intelligenten Systemen und Menschen interaktiv gebildet. Dies ist besonders im Hinblick auf aktive technologische Umgebungen relevant, wie das Internet of Things (vgl. Aydin/González Woge/Verbeek 2019; Verbeek 2009). Schließlich fokussieren wir auf den Softwareentwicklungsprozess

1 Die Mensch-Maschine-Interaktion wird als interaktives System definiert, nämlich als »Kombination von Hardware und/oder Software und/oder Diensten und/oder Menschen, mit denen Benutzer interagieren, um bestimmte Ziele zu erreichen« (ISO 9241–11:2018: 3.1.5). Diese allgemeine Definition umfasst auch spezielle Forschungsbereiche wie die Mensch-Computer- oder die Mensch-Roboter-Interaktion. Speziell Letztere ist ein inter- und multidisziplinäres Forschungsfeld, das Technik, Psychologie, Design, Anthropologie, Soziologie und Philosophie umfasst (vgl. Bartneck et al. 2020: 9). Das Ziel in diesem Feld ist es, soziale Roboter zu entwickeln, die in der Lage sind, soziale Rollen zu übernehmen, zum Beispiel als Mitarbeiter, Tutoren und Assistenten im medizinischen Bereich, im Dienstleistungssektor und in der Pflege, im Bildungswesen und in den Wohnungen der Menschen (vgl. ebd.: 201).

selbst, um zu verstehen, wie die Modellierung einer gelingenden Kommunikation zwischen Menschen und Maschinen durch die Interaktion von Nutzer:innen mit dem Entwicklungsteam ermöglicht wird. Mit der triadischen Modellierung, in welcher die Maschinen mit Fokus auf user-centered Design konstruiert werden, präsentieren wir ein pragmatisch-interaktives Informationsmodell. In diesem Rahmen lässt sich beschreiben, wie der Zugang zur Benutzbarkeit gestaltet, Bedeutung durch Triangulation generiert und im aktiven Gebrauch stabilisiert wird.

2. Kommunikation in Interaktion

Das Sender-Empfänger-Modell von Claude Shannon und Warren Weaver (1949) wird als ein klassisches Kommunikationsmodell behandelt, das in verschiedene Fachbereiche eingebracht und in diesen auch adaptiert wurde (vgl. Röhner/Schütz 2020: Kap. 2). Das ursprüngliche Ziel des Modells war die Beschreibung der Kommunikation im nachrichtentechnischen Sinn als Informationsaustausch zwischen zwei informationsverarbeitenden Systemen, dem sendenden und dem empfangenden System. Kommunikation ist dabei mathematisch bzw. rein syntaktisch modelliert und bewusst unabhängig von Bedeutungsebenen konzipiert. Eine direkte Übertragung dieses Sender-Empfänger-Modells auf zwischenmenschliche Kommunikation ist nicht plausibel, denn es wird ausgeklammert, wie Menschen Kommunikationskompetenzen entwickeln, um in sozialen Interaktionen adäquat zu kommunizieren. Wir behaupten, dass Bedeutung nur in sozialen Interaktionen entstehen kann.

Als Alternative zum klassischen dyadischen Sender-Empfänger-Modell ist das entwicklungspsychologische Konzept der Triangulation prominent, welches die triadische Beziehung in der sozialen Interaktion hervorhebt und zeigt, wie Bedeutungen in Interaktionen entstehen und wie diese in kommunikative Handlungsmuster einfließen.

Ausgehend von der frühen Dyade von Mutter und Kind, die gemeinsame Aufmerksamkeit und Engagement erfordert (vgl. Trevarthen 1979), kommt spätestens im Kindesalter von neun bis zwölf Monaten ein Drittes hinzu. Das heißt, mit der sogenannten Neunmonatsrevolution verwandelt sich die Dyade durch deklaratives Zeigen und andere Gesten in eine Dreiecksrelation

(vgl. Tomasello 1999: 62; Fuchs 2013: 667).[2] Das deklarative Zeigen ermöglicht den Zugang zu symbolischer Interaktion (vgl. ebd.; Werner/Kaplan 1963: 63f.). Durch den Akt des Zeigens innerhalb des verkörperten symbolischen Interaktionsprozesses verwandeln sich Dinge in gemeinsame symbolische Objekte. Im Gegensatz zur einfachen Dyade bringt der/die andere Handelnde eine zusätzliche Perspektive auf das gemeinsam geteilte Objekt mit. Diese Objekttriangulation (Subjekt – Subjekt – Objekt) führt erst zu einer gemeinsamen Wahrnehmung und schließlich zu einer geteilten Intentionalität (vgl. Tomasello et al. 2005; Fuchs 2013: 667).

Innerhalb der Triade von Kind, Betreuungsperson und Objekt ist ein gemeinsamer intentionaler Handlungsraum verfügbar. Kinder beobachten, wie und wofür Bezugspersonen in der jeweiligen Umgebung Dinge benutzen, und imitieren das beobachtete Verhalten (vgl. Tomasello 1999: 84). Nehmen wir ein Beispiel: Die Familie sitzt am Mittagstisch und möchte speisen, üblicherweise mit Besteck. Für die Kinder, die mit am Tisch sitzen, ist der gemeinsame intentionale Handlungsraum, das heißt die Absicht zu essen, am Küchentisch verfügbar. Sie begreifen spielerisch im imitierenden Handlungsvollzug, wie die Bezugspersonen das Besteck benutzen. Sie ahmen nach und treten in den gemeinsamen intentionalen Handlungsraum ein:

> Children now come to comprehend how ›we‹ use the artifacts and practices of our culture – what they are ›for‹. Monitoring the intentional relations of others to the outside world also means that the infant – almost by accident, as it were – monitors the attention of other persons as they attend to her. This then starts the process of self-concept formation, in the sense of the child understanding how others are regarding ›me‹ both conceptually and emotionally. (Ebd.: 91)

Das Kind macht in sozialen Interaktionen Erfahrungen und erwirbt nach und nach Handlungsmuster innerhalb dieser gewohnheitsmäßig strukturierten Bezugsrahmen, die wiederum kognitive und motivierende Bedingungen für weitere soziale Handlungen formen. Der Fokus liegt hier auf sozialen Praktiken als sich stetig wiederholenden Interaktionen, die im Laufe der

2 Die Dreiecksrelation wird ausführlich von Donald Davidson (2001) untersucht, der mit seinem Konzept der Triangulation die verkörperte Beziehung zwischen zwei oder mehreren Partner:innen und der gemeinsamen Welt bzw. den gemeinsamen Objekten um sie herum beschreibt. In unserem Aufsatz geht es uns hingegen vorrangig um die erkenntnistheoretische Dimension der Triangulation.

Zeit bestimmte Formen von gewohnheitsmäßig strukturierten Bezugsrahmen erzeugen. Wir nennen diese spezifischen Formen sozialer Praktiken im Folgenden pragmatische Handlungsmuster.

Die Triade muss aber nicht zwangsläufig aus zwei Subjekten und einem Objekt bestehen. An sozialen Interaktionen sind häufig dritte Personen beteiligt, weswegen auch von Subjekttriangulation gesprochen werden kann. Die Einbeziehung einer dritten Person spielt eine grundlegende Rolle in der Entwicklung der sozialen Kognition, denn hier kommt eine dritte Sicht auf die dyadische Beziehung selbst hinzu. Die dritte Person, ob als beobachtende oder bezeugende, hat einen Blick von außen auf die Dyade und so entwickelt sich in der triadischen Interaktionsbeziehung das Verständnis der Perspektive des anderen (vgl. Fuchs 2013: 668). In dieser Konstellation kann sich das Kind seines eigenen wie auch des Standpunkts der anderen bewusst werden, was es ihm langfristig ermöglicht, zwischen unterschiedlichen Sichtweisen flexibel zu wechseln und diese zu vergleichen. Wie Studien zur frühkindlichen Entwicklung zeigen, beginnt der Erwerb der Fähigkeit zur Perspektivenübernahme schon im ersten Lebensjahr. Zentral für den frühen Austausch von Perspektiven ist die gemeinsame Aufmerksamkeit (vgl. Moll/Meltzoff 2011a). Schon im Alter von zweieinhalb Jahren können Kinder verschiedene Perspektiven einnehmen. Im Alter von viereinhalb bis fünf Jahren entwickelt sich ein Verständnis für die verschiedenen Sichtweisen und Überzeugungen anderer (vgl. Moll/Meltzoff 2011a, b).

Durch Objekt- wie Subjekttriangulation wird symbolische Kommunikation ermöglicht. In der dynamischen Auseinandersetzung mit der materiellen und sozialen Welt erfährt das Kind die Affordanzen von Objekten, indem es beobachtet, wie Erwachsene in gemeinsam erlebten Situationen Objekte verwenden.[3] Die dynamische Auseinandersetzung von Kindern mit Objekten in ihrer Umgebung wird in dreierlei Hinsicht beschrieben: als sensorisch-motorische Affordanz (das Objekt ist nutzbar), als konventioneller Gebrauch (wir benutzen es als …) und als symbolisches Spiel (ich kann es benutzen als …) (vgl. Tomasello 1999: 84ff.). Diese verschiedenen Aspekte prägen unsere interaktive Kommunikation und bilden die Bezugsrahmen der pragmatischen Handlungsmuster.

3 Der Begriff Affordanz wurde von Gibson im Rahmen eines ökologischen Ansatzes zur visuellen Wahrnehmung eingeführt. Vgl. Gibson (1979: 127).

3. Information in der Mensch-Maschine-Kommunikation

Ausgehend von der zwischenmenschlichen Interaktion in Bezug auf den Umgang mit Objekten möchten wir nun die Kommunikation von Mensch und Maschine betrachten, wobei unser Fokus auf dem generellen Verhältnis zwischen Menschen und ihren technischen Errungenschaften liegen wird. Wenn wir heute von Mensch-Maschine-Kommunikation reden, dann meinen wir ›Informationsartefakte‹, die auf Informations- und Kommunikationstechnologien beruhen.[4] Die heutigen Entwicklungen im Bereich der Künstlichen Intelligenz, die in der Zusammenarbeit von verschiedenen Disziplinen wie Informatik, Datenwissenschaft und Kognitionswissenschaft vorangetrieben werden, versprechen dabei neue Möglichkeiten der Nutzung: Maschinen werden nun nicht nur als Werkzeuge, sondern auch als potenzielle Interaktionspartner wahrgenommen, mit denen wir kommunizieren und zusammenarbeiten können.

Als attraktives Paradigma für die Erforschung einer allgemeinen Kommunikation, die sowohl Menschen als auch Maschinen betreffen soll, hat sich die Kybernetik erwiesen. Mit dem namensgebenden Werk *Cybernetics or Control and Communication in the Animal and the Machine* (1948) legte Norbert Wiener den Grundstein dieser Disziplin. Information ist der entscheidende Begriff der Kybernetik. Maßgeblich inspiriert von den technischen Fortschritten der Nachrichtentechnik der 1930er und 1940er Jahre entwarfen Shannon und Weaver eine Kommunikationstheorie, die es erlaubt, Informationsübertragung mathematisch zu modellieren.

Information wird bei Shannon und Weaver als statistische Größe eingeführt: »[I]nformation is a measure of one's freedom of choice when one selects a message.« (Shannon/Weaver 1949: 9) Sie ist dabei rein syntaktisch und unabhängig von Bedeutungsebenen konzipiert. Die Grundidee ist, ein funktionales Konzept der Kommunikation für Ingenieure zu entwerfen: »[S]emantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one selected from a set of possible messages. The system must be designed to operate for each possible selection,

4 Der Maschinenbegriff ist höchst vielgestaltig: Gemeint sein können einfache Werkzeuge, mechanische bzw. elektrische Maschinen, elektronische, sich selbstregulierende Automaten oder auch komplexe mathematische Modelle. Für eine detaillierte Diskussion siehe Schumann und Du (2021: 53).

not just the one which will actually be chosen since this is unknown at the time of design.« (Ebd.: 31)

Dieses mathematische Modell der Kommunikation war in der damaligen Zeit sehr hilfreich für die Weiterentwicklung der Datenübertragung (Reduzierung des Hintergrundlärms) in den Bereichen Telegrafie, Telefon, Funk und Fernsehen, und der automatischen Kommunikation von Maschinen untereinander (Maschine-Maschine-Kommunikation). Heutzutage ist der Informationsaustausch zwischen Maschinen so alltäglich, dass wir ihn in unserer lebensweltlichen Handlungspraxis selten bewusst wahrnehmen. Der Nutzen der Maschine-Maschine-Kommunikation liegt hauptsächlich in der Sammlung riesiger Datenmengen, in der Übertragung auf spezifische Netzwerke, die die Grundlage für das Internet of Things (IoT) bilden (vgl. Knoll/Lautz/Deuß 2015).

Die heutige Digitalisierung umfasst die Entwicklung des Internets, die Vernetzung von Maschinen, Dingen und Menschen als Nutzer:innen (vgl. Schumann/Du 2021: 57). Im Zuge dieses Prozesses werden Teile unserer Lebenswelt in binären Datenformaten erfasst und transformiert. Die zunehmende Digitalisierung alltäglicher Praxisformen mit gigantischen Datenmengen verändert die Kommunikation zwischen Menschen und Maschinen nachhaltig. An dieser Stelle muss man allerdings fragen, ob eine rein syntaktische Informationstheorie im Sinne des mathematischen Kommunikationsmodells von Shannon und Weaver ausreicht, um die Kommunikation von Mensch und Maschine zu fundieren. Wie wir im vorherigen Kapitel schon gezeigt haben, entsteht Bedeutung in menschlicher Kommunikation durch soziale Interaktion. Bevor Sender und Empfänger als ideale Kommunikationspartner in einer geteilten Umgebung auftreten, werden sie durch soziale Interaktionen erst geformt.

Das Internet als global verlinktes Netzwerk bietet die Möglichkeit, die rein syntaktische Formalisierung von Informationen mit semantischen Zuschreibungen zu erweitern. Das heißt, mithilfe einer standardisierten Schichtenarchitektur werden Daten nicht nur auf der physikalischen Ebene der Bitübertragung, sondern auch in Form von Datenpaketen auf der Ebene der Anwendung auf dem Interface ausgetauscht.

Die Verarbeitung und der Austausch formal syntaktischer Daten werden durch standardisierte Protokolle gewährleistet. Unter Anwendung von semantischen Technologien werden gleichzeitig komplexe Begriffsnetze, sogenannte Ontologien in der Informatik, schrittweise aufgebaut, indem einzelne Begriffe mithilfe eines Inventars an Verknüpfungsregeln, zum Beispiel mit

Relationen wie »Unterbegriff von«, »Gegenteil von« oder »Gleichbedeutend mit«, bestimmt und kontextualisiert werden, mit dem Ziel, Bedeutungsverlust zu vermeiden (vgl. Wahlster 2015; Schumann/Du 2021: 57). Dabei sind die Beschreibungen und Relationen nicht festgelegt, sondern werden durch Anwendung der geteilten Daten in der Auszeichnungssprache, etwa durch Labels, Tags oder andere Zuschreibungen, und deren semantische Relationen dynamisch ergänzt. Die Ordnung der Daten bleibt flexibel und ist nicht mit einem festgelegten taxonomischen Lexikon zu vergleichen. Die Ordnung ändert sich zum einen mit dem Gebrauch durch die Nutzer:innen und zum anderen mit der dezentralen Korrektur und Editierung durch die Community (Anbieter). Mit Hilfe dieser semantischen Technologien mit semistrukturierter Syntax ist semantische Zuschreibung möglich.

Der semantische Informationsbegriff beinhaltet ein pragmatisches Verständnis von Information. Die nutzer:innenorientierte Ausrichtung eröffnet gleichzeitig eine neue Dimension in der Mensch-Maschine-Kommunikation. Zwischenmenschliche Praxisformen und der Umgang mit den technischen Errungenschaften wie dem Internet werden integriert. Damit werden Informationen nicht nur als Bedeutungseinheiten formalisierter Daten verstanden, sondern beziehen auch das durch den Gebrauch geteilte Wissen mit ein. Dieses Wissen wird nicht einfach simuliert, sondern durch Gebrauch generiert. Die Vernetzung des Internets mit semantischer Technologie bildet die Grundlage für Anwendungen Künstlicher Intelligenz in der Mensch-Maschine-Kommunikation, die die Informationsverarbeitung und -übertragung in semantischer und pragmatischer Hinsicht berücksichtigen. Diese Integration pragmatischer Information erweitert unsere Praxisformen in Bezug auf den Umgang mit Maschinen. So werden schließlich gemeinsame Handlungsformen geschaffen, die wir nicht nur untereinander, sondern auch mit Maschinen teilen können.

Der semantische Gehalt von Information wird durch Gebrauch generiert. Mit dem Leitprinzip der Triangulation aus der sozialen Kognitionsforschung versuchen wir nun wichtige Impulse für den Entwicklungsprozess einer gelingenden Kommunikation zwischen Menschen und Maschinen zu gewinnen. Im Hinblick auf die heutige technologische Infrastruktur werden wir im folgenden Abschnitt zeigen, wie eine gemeinsam geteilte Umgebung eine Kommunikation von Menschen und Maschinen ermöglicht.

4. Interaktive Kommunikation in aktiven technologischen Umgebungen

Im heutigen Zeitalter von Informations- und Kommunikationstechnologien verbindet das IoT, unterstützt durch semantische Technologien, den Cyberspace mit der physischen Umgebung. Dadurch entsteht ein neues technologisches Umfeld, das nicht nur Dinge, sondern auch Menschen untereinander und Menschen mit Maschinen vernetzt. In unserer technisierten Welt beeinflussen Informations- und Kommunikationstechnologien unsere Erfahrungen und formen und erweitern unsere Handlungsspielräume.[5] In diesem Sinne handelt es sich nicht nur um eine passive Infrastruktur, die lediglich im Hintergrund bleibt, sondern beeinflusst auch aktiv unsere Handlungsweisen.[6]

Wenn wir vor diesem Hintergrund von aktiven technologischen Umgebungen sprechen, wird auf die vermittelnde, aktive Rolle von Technologien hingewiesen (vgl. Aydin/González Woge/Verbeek 2019: 322). Diese vermittelnde Rolle technischer Artefakte im Allgemeinen ist ein zentrales Thema in der Anthropologie, konkret in Bezug auf das Verhältnis von menschlicher Kognition und Werkzeuggebrauch in der kulturellen Evolution. Im Mittelpunkt steht dabei die Frage, wie materielle Dinge unsere kognitive Struktur beeinflussen (vgl. Malafouris 2013: 247). In anthropologischer Hinsicht sind Werkzeuge Teil von kognitiven Prozessen. Aufgrund der starken Plastizität menschlicher Kognition gibt es keine klare Grenze zwischen Mensch und Technik. Materielle Objekte sind für uns nicht nur passive Werkzeuge, sondern bilden auch eine Ökologie der materiellen Welt. Diese wiederum prägt unsere kognitive Ökologie und erweitert so die Möglichkeiten menschlichen Handelns (vgl. Ihde/Malafouris 2019: 198). Dabei wird menschliche Kognition

5 Der Postphänomenologe Don Ihde hat die verschiedenen Beziehungen von Mensch und Welt, die durch Technologien vermittelt werden, ausführlich beschrieben. Er unterscheidet vier Relationen: 1. die verkörperte Relation (z.B. Brille), 2. die hermeneutische Relation (Thermometer), 3. die Alteritätsrelation (Auto, Computer) und 4. die Hintergrundrelation (WLAN-Router) (vgl. Ihde 1990: 72–112; vgl. dazu auch Schumann/Du 2022: 7–10).

6 So zeigen zum Beispiel Aydin, González Woge und Verbeek (2019: 336), wie die Hintergrundrelation von Ihde zu einer sogenannten Immersionsbeziehung erweitert wird, indem Technologien derart mit unserer Welt verschmelzen, dass zwischen ihnen und dem Menschen eine bidirektionale intentionale Beziehung entsteht (vgl. dazu auch Schumann/Du 2022: 9).

als dynamischer Prozess verstanden, der in sozialen Interaktionen unter Einbeziehung der physischen, technischen, sozialen und kulturellen Umwelt abläuft. In dieser Hinsicht konstituieren sich die kognitive Ökologie und die materielle Umgebung wechselseitig.

Die idealen funktionalen Sender und Empfänger sowie die verarbeitbaren Informationen sind nicht vorbestimmt, sondern werden in einem breiteren dynamischen Prozess der kognitiven Ökologie ko-konstituiert. Hier werden die pragmatischen Muster in Interaktion mit der materiellen und sozialen Umgebung geformt. Unsere technisch-kognitive Ökologie wird durch pragmatische Handlungsmuster vermittelt, die durch stabilisierten Gebrauch letztendlich in Form von technologischen Artefakten vergegenständlicht werden können. Wir verstehen diese pragmatischen Muster als im Zuge ihres Gebrauchs mit Bedeutung aufgeladene Information. Die menschliche Nutzung von Technologien konstituiert also unsere aktiven technologischen Umgebungen und diese konstituieren und verarbeiten Informationen als Bedeutung in einer aktiven und nichtdirektionalen Weise (vgl. Schumann/Du 2022: 10).

Die gewohnheitsmäßig strukturierten Bezugsrahmen, die sich mit der Zeit durch soziale Praktiken entwickeln, sind nicht nur die Grundlage für das Erwerben pragmatischer Handlungsmuster, sondern formen auch weitere kognitive und motivierende Bedingungen für weitere soziale Handlungen. Wie wir bereits herausgestellt haben, ermöglicht soziale Interaktion symbolische Kommunikation. Und diese symbolische Kommunikation kann mit dem Konzept der Triangulation in sozialer Interaktion erklärt werden, und bietet so eine Alternative zum klassischen dyadischen Sender-Empfänger-Modell.

Sowohl Sender als auch Empfänger müssen Zugang zum Bezugsrahmen besitzen, um sich pragmatische Handlungsmuster aneignen zu können, das heißt, sie lernen und sammeln Erfahrungen in sozialen Interaktionen und erlangen Fähigkeiten und Fertigkeiten im Umgang mit den Dingen und mit anderen Menschen. Dieses implizite Umgehenkönnen entspricht dem, was Ryle (1949: 28ff.) als »knowing-how« bezeichnete. Um daraus ein explizites »knowing-that« entwickeln zu können, bedarf es ebenfalls, wie oben gezeigt, der Triangulation in sozialer Interaktion. Dieses explizite Wissen ist als Information im Sinne von pragmatischen Handlungsmustern formalisierbar. Und hier besteht die Möglichkeit, einen Bezugsrahmen zu konstruieren, der Maschinen in ihrer Aufbau- und Ablauforganisation bedingt. Das heißt nicht, dass wir Maschinen anthropomorphisieren, sondern wir suchen nach den Bedingungen dafür, wie Maschinen als potenzielle Kommunikationspartner

unseren Handlungsspielraum erweitern können. Durch die Transformation expliziter Informationen werden materielle Bedingungen, gewohnheitsmäßig strukturierte Bezugsrahmen und letztendlich auch normative Bedeutungsrahmen in Bezug auf neue technische Entwicklungen integriert. Diese Bedingungen werden im Prozess sozialer Praktiken, also im Gebrauch in Form von Mustern, stabilisiert und bieten schließlich die Grundlage für rationale Kommunikationsformen (vgl. Brandom 1994).

Unter den heutigen technologischen Bedingungen, besonders im Hinblick auf die oben erörterten aktiven technologischen Umgebungen, wird der Bezugsrahmen interaktiver Kommunikation zwischen künstlich-intelligenten Agenten und Menschen untereinander rapide und effizient gebildet. Die technologische Infrastruktur, die uns Menschen umgibt, ist auch für Maschinen verfügbar. An dieser Stelle kann man allerdings fragen, unter welchen Bedingungen eine sinnvolle Kommunikation zwischen Menschen und Maschinen überhaupt entwickelt werden kann.

5. Triadische Interaktion im Entwicklungsprozess

Jedwede technologische Entwicklung ist in einen soziotechnologischen Rahmen eingebettet, der wiederum durch menschliche Praktiken geformt und geprägt wird. Eine gelingende Kommunikation von Mensch und Maschine setzt voraus, dass die Maschine schon Teil spezifischer Praxisformen ist. Damit rückt der Entwicklungsprozess selbst in den Fokus der Analyse.

Die Mensch-Maschine-Interaktion ist ein interdisziplinärer Forschungsbereich, in dem die unterschiedlichsten Disziplinen vertreten sind. Hier finden sich Akteure aus dem Ingenieurswesen, der Psychologie, dem Design, der Anthropologie oder der Soziologie (vgl. Bartneck et al. 2020: 9). Generell sind in der Softwareentwicklung multidisziplinäre Fähigkeiten und die Anerkennung unterschiedlicher Perspektiven innerhalb des Entwicklungsteams gefragt. Um die Entwicklungsprozesse genauer in den Blick zu bekommen, eignet sich das Leitprinzip der Triangulation. Die Triade besteht aus dem Produkt, dem/der Nutzer:in und dem Entwicklungsteam. Das Produkt, in diesem Fall eine Software, die eine erfolgreiche Kommunikation ermöglicht, wird vom Team in einem dynamischen Prozess in Interaktion mit dem/der potenziellen Nutzer:in entwickelt. Die Beziehungen zwischen Nutzer:innen, Entwicklungsteam und Produkt bilden im Entwicklungsprozess selbst eine triadische Interaktion, die in drei verschiedenen Konstellationen denkbar

ist: (1) der Zusammenarbeit zwischen dem/der Endnutzer:in und dem Entwicklungsteam, (2) der Interaktion zwischen dem/der Endnutzer:in und dem Produkt, System oder Dienstleistung und (3) der Beziehung zwischen dem Produkt und dem Team während des Entwicklungsprozesses (vgl. Schumann/Du 2022: 13).

Um eine erfolgreiche Kommunikation zwischen Menschen und Maschinen zu ermöglichen, ist es in der ersten Phase der Entwicklung notwendig, einen Prototyp zu schaffen, damit überhaupt ein Zugang zur kommunikativen Interaktion gegeben ist. Sobald ein Prototyp zur Verfügung steht, kann die Interaktion getestet und angepasst werden. Obwohl die konkrete Benutzung (usability) des Produkts in der ersten Phase der Prototypentwicklung noch nicht klar definiert ist, ist die Zugänglichkeit für die Nutzung (accessibility) zunächst von entscheidender Bedeutung. Inwieweit ein Produkt schließlich benutzerfreundlich gestaltet werden kann, ergibt sich allmählich im Entwicklungsprozess.

Bei diesem handelt es sich um einen iterativen Prozess, in dem die progressive adaptive Entwicklung umgesetzt wird. Damit ist es möglich, sofort auf die sich auf der Grundlage ihrer Erfahrungen ändernden Bedürfnisse der Nutzer:innen zu reagieren. Der iterative Prozess ist zentral für die Entwicklung und das Design innerhalb der agilen Softwareentwicklung (vgl. Agile Alliance 2021).

Das erwähnte Prinzip der Triangulation ist hier insofern relevant, als eine erfolgreiche Kommunikation von Mensch und Maschine einen gemeinsamen Bezugsrahmen erfordert, der sowohl materielle als auch soziale Bedingungen vereint. An dieser Stelle kann man nun fragen, ob sich die triadische Interaktion auf die zwischenmenschliche Kommunikation oder auf die Mensch-Maschine-Kommunikation bezieht. Bei dem besagten Entwicklungsprozess spielen offenkundig beide Konstellationen eine Rolle: Einerseits geht es um die Entwicklung einer interaktionsfähigen Maschine, andererseits ist die Zusammenarbeit zwischen Endnutzer:in und Entwicklungsteam entscheidend dafür, dass eine Maschine realisiert wird, die die Handlungsspielräume der Nutzer:innen zu erweitern vermag.

Das Ziel einer gelingenden Kommunikation wird von den Nutzer:innen und dem Entwicklungsteam geteilt. Der Ursprung des gemeinsamen Ziels liegt in der gemeinsam geteilten Welt. In dieser finden sich einerseits bestehende Infrastrukturen unseres globalen Netzwerks, die die Kompatibilität zwischen einem neuen Produkt und dem aktuellen technologischen Umfeld des Produkts gewährleisten. Andererseits ist durch soziokulturelle Praktiken

ein gemeinsamer intentionaler Raum gegeben, in dem unsere Handlungsmuster mit bestimmten Bedeutungen in ihrem Gebrauch stabilisiert werden.

6. Schluss

Mit dem Begriff des pragmatischen Handlungsmusters lässt sich beschreiben, wie der Zugang zur Nutzbarkeit geprägt ist und wie semantische Bedeutung durch Triangulation entsteht und im aktiven Gebrauch stabilisiert wird. Bedeutungen werden einerseits durch ihre Verwendung gebildet und andererseits durch pragmatische Muster als Medium vermittelt. Zusätzlich eröffnen aktive technologische Umgebungen neue Handlungsmöglichkeiten, indem sie, unterstützt durch semantische Technologien, Handlungsmuster in Bedeutungsrahmen zur Verfügung stellen.

Diese technologisch und intersubjektiv vermittelten sozialen Interaktionen erweitern unsere Kommunikationsmöglichkeiten und unser sprachliches Verhalten. Damit wird klar, dass Sender und Empfänger als ideale Informationsverarbeiter nicht von Beginn an feststehen, sondern im Interaktionsprozess selbst geformt werden. Im Zeitalter der Digitalisierung beeinflussen sich materielle und kognitive Ökologien wechselseitig. Das bedeutet, dass die gemeinsamen Handlungsmuster auch Teil der Maschine sind, denn die Konstruktion der Maschine basiert auf unserer sozialen Praxis. Damit wird nicht nur die Kommunikation zwischen Mensch und Maschine ermöglicht und verbessert, sondern auch die Kommunikation von Menschen untereinander nachhaltig verändert. Indem wir in der Interaktion unter uns und mit Maschinen pragmatische Handlungsmuster bilden und im Gebrauch stabilisieren, schaffen wir neue Umgebungen, die wiederum unsere Handlungsspielräume erweitern.

Literatur

Agile Alliance. 2021. https://web.archive.org/20210317200511/https://www.agilealliance.org/. Zugegriffen: 4. April 2022.

Aydin, Ciano, Margoth González Woge und Peter-Paul Verbeek. 2019. Technological Environmentality: Conceptualizing Technology as a Mediating Milieu. *Philosophy and Technology* 32: 321–338. https://doi.org/10.1007/s13347-018-0309-3.

Bartneck, Christoph, Tony Belpaeme, Friederike Eyssel, Takayuki Kanda, Merel Keijsers und Selma Šabanović. 2020. *Human-Robot Interaction – An Introduction*. Cambridge: Cambridge University Press.

Brandom, Robert B. 1994. *Making it Explicit: Reasoning, Representing, and Discursive Commitment.* Cambridge, Mass.: Harvard University Press.

Breazeal, Cynthia, Kerstin Dautenhahn und Takayuki Kanda. 2016. Social Robotics. In *Springer Handbook of Robotics,* Hg. Bruno Siciliano und Oussama Khatib, 1935–1972. Cham: Springer. https://doi.org/10.1007/978-3-319-32552-1_72.

Davidson, Donald. 2001. *Subjective, Intersubjective, Objective.* Oxford: Clarendon Press.

Deckert, Ronald. 2019. *Digitalisierung und Industrie 4.0. Technologischer Wandel und individuelle Weiterentwicklung*. Wiesbaden: Springer Gabler.

Fuchs, Thomas. 2012. The Phenomenology and Development of Social Perspectives. *Phenomenology and the Cogntive Sciences* 12: 655–683. https://doi.org/10.1007/s11097-012-9267-x.

Gibson, James J. 1979. *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin.

Ihde, Don und Lambros Malafouris. 2019. *Homo faber* Revisited: Postphenomenology and Material Engagement Theory. *Philosophy and Technology* 32: 195–214. https://doi.org/10.1007/s13347-018-0321-7.

Ihde, Don. 1990. *Technology and the lifeworld: From garden to earth*. Indiana University Press.

International Organization for Standardization. 2018. *Ergonomics of Human-System Interaction – Part 11: Usability: Definitions and Concepts* (ISO 9241–11:2018).

Knoll, Thomas, Alexander Lautz und Nicolas Deuß. 2015. Machine-To-Machine Communication. In *Handbuch Industrie 4.0.* Hg. Birgit Vogel-Heuser, Thomas Bauernhansl und Michael ten Hompel, Springer NachschlageWissen. Berlin, Heidelberg: Springer Vieweg, 1–10. https://doi.org/10.1007/978-3-662-45537-1_84-1.

Moll, Henrike und Andrew N. Meltzoff. 2011a. Perspective Taking and Its Foundation in Joint Attention. In *Perception, Causation, and Objectivity: Issues in Philosophy and Psychology*, Hg. Johannes Roessler, 286–304. Oxford: Oxford University Press.

Moll, Henrike und Andrew. N. Meltzoff. 2011b. Joint Attention as the Fundamental Basis of Perspectives. In *Joint Attention*, Hg. Axel Seemann, 393–413. Boston: MIT Press.

Röhner, Jessica und Astrid Schütz. 2020. *Psychologie der Kommunikation*. 3. Auflage. Berlin: Springer.

Ryle, Gilbert. 1949. *The Concept of Mind*. New York: Routledge.

Schumann, Nadine und Yaoli Du. 2021. Grenzgänge: Von Menschen zu smarten Maschinen – und zurück? In *Künstliche Intelligenz – Die große Verheißung*, Hg. Anna Strasser, Wolfgang Sohst, Ralf Stapelfeldt und Katja Stepec, 3–62. Berlin: Xenomoi.

Schumann, Nadine und Yaoli Du. 2022. Machines in the Triangle: A Pragmatic Interactive Approach to Information. *Philosophy & Technology* 35. https://doi.org/10.1007/s13347-022-00516-4.

Shannon, Claude E. und Warren Weaver. 1949. *The Mathematical Theory of Communication*. Urbana: University of Illinois Press.

Tomasello, Michael, Malinda Carpenter, Josep Call, Tanya Behne und Henrike Moll. 2005. Understanding and sharing intentions: The origins of cultural cognition. *The Behavioral and Brain Sciences*, 28, 675–735.

Tomasello, Michael. 1999. *The Cultural Origins of Human Cognition*. Cambridge, Mass.: Harvard University Press.

Trevarthen, Colwyn. 1979. Communication and Cooperation in Early Infancy: A Description of Primary Intersubjectivity. In *Before Speech*, Hg. Margaret Bullowa, 321–347. Cambridge: Cambridge University Press.

Verbeek, Peter-Paul. 2009. Ambient Intelligence and Persuasive Technology: The Blurring Boundaries Between Human and Technology. *Nanoethics* 3: 231–242. https://doi.org/10.1007/s11569-009-0077-8.

Wahlster, Wolfgang. 2015. Semantische Technologien als Wegbereiter für das Internet der Dinge. *Handelsblatt-Beilage zur CeBIT 2015 d!conomy*, 26.01.2015.

Werner, Heinz und Bernard Kaplan. 1963. *Symbol Formation: An Organismic-Developmental Approach to Language and the Expression of Thought*. New York, London und Sydney: John Wiley & Sons.

Generative Praktiken

If I Say the Word Out Loud, It Will Be More Real

Jakob Claus & Yannick Schütte

Abstract: *Pharmako-AI ist ein von K Allado-McDowell und dem KI-Sprachmodell GPT-3 verfasstes Buch. Der Text ist als experimentelle Unterhaltung konzipiert, die das Verhältnis von Menschen und KI, die Struktur und Grammatik von Sprache, Kybernetik und Counterculture befragt und dabei alternative Modi des Schreibens verfolgt. Sprachliche Äußerungen werden im Verlauf des Buches zu einem performativen und generativen Akt, der sich selbst reflektiert. Zugleich zerfällt die Unterscheidung zwischen menschlicher und maschineller Autor*innenschaft im Verlauf der einzelnen Kapitel. In dem Text untersuchen wir Fragen von literarischer Navigation und Kontrollverlust, das Motiv des Orakels als Spiegel unbewusster sprachlicher Muster und Prozesse sowie KI als Medium menschlicher Sprache. Pharmako-AI verweist aber ebenso, wie wir mit Sianne Ngai zeigen, auf stuplime Momente von Wahrscheinlichkeiten und literarischer Sprache, die als vermeintlich akzidentielle Verdichtungen eine poetische Dimension entfalten können.*

> »When a reader reads a novel, the novel takes on a shape and life as an external object, as the story I have been telling myself about how that structure came to exist. Yet, in another sense, the novel is a shape that is not an object, but a structure – and it enmeshes me in it.« (Allado-McDowell 2020a: 17)

K Allado-McDowell und GPT-3 heißen die Charaktere von *Pharmako-AI*. Das 2020 erschienene Buch entspinnt sich als Dialog dieser beiden Schreibenden, der die Beziehung von Eingabe und Ausgabe sowie die fortwährende Verflechtung thematisiert. So stellt GPT-3 fest, dass die Erzählung des Textes, zu der das KI-Sprachmodell von Allado-McDowell animiert wurde, eine Struktur der Verstrickung abbildet. Diese Struktur ist sowohl Gegenstand als auch Beteiligte im Erzählen der Geschichte. In diesem Sinne ist das dialogische Schrei-

ben die Beobachtung der Verflechtung mit einem/einer anderen (vgl. Allado-McDowell 2020a: 16). Buchpassagen wie die oben angeführte sind nicht nur aufschlussreich im Hinblick auf die sprachliche Zusammenarbeit mit KI, sondern erlauben auch eine ›interessante‹ Perspektive auf die Praxis und Technik des Schreibens, der wir im Folgenden nachgehen wollen. Der Text entwickelt in seinem Verlauf eine mäandernd-poetische Form textbasierter – oder literarischer – Interaktion mit eigenen Rhythmen, Mustern und Poetiken und eröffnet Anschlüsse an Schreib- und Leseerfahrungen. In unserer Auseinandersetzung mit *Pharmako-AI* – dem ersten Buch, das Allado-McDowell zusammen mit GPT-3 verfasst hat[1] – widmen wir uns den Versprechungen und Möglichkeiten, die KI als literarisches Spiel- und Werkzeug eröffnet. Dabei interessieren uns theoretische Implikationen und Metaphern, die weiterführende Überlegungen zu KI-Literatur erlauben. Im Zuge einer Betrachtung von Formen und Motiven ästhetischer Praktiken und Strategien, in deren Tradition der Text sich verorten lässt, gehen wir auf die mögliche Banalität von errechneter Sprache und Klischees ein, die in *Pharmako-AI* reproduziert werden.

Die Figur des Orakels, auf die wir über Allado-McDowell und Kate Crawford zu sprechen kommen, verweist auf die Herausforderungen der Navigation und Interpretation im Umgang mit informationsverarbeitenden Systemen. Autor*innenschaft zeigt sich hier als Spannung zwischen Steuerung und Kontrollverlust. Aber ebenso führt uns dies zu einem Vergleich von *Pharmako-AI* mit Motiven aus Stanisław Lems *Solaris*, der es ermöglicht, die Beziehung von Medien und menschlicher Sprache bzw. Kommunikationsfähigkeit zu beleuchten. Ähnlich wie GPT-3 wird der Ozean auf Solaris in der Interaktion zu einem Spiegel menschlicher Vergangenheitsbewältigung. Zuletzt diskutieren wir im Einsatz der KI als Werkzeug und ihren Kapricen ein Beispiel für die komische Paarung aus Aufregung und Langeweile geschriebener Sprache, die wir Sianne Ngai folgend als *Stuplimity* bezeichnen werden.

---

1 Nach *Pharmako-AI* wurden zwei weitere Bücher von K Allado-McDowell und GPT-3 veröffentlicht. Der im April 2022 erschienene Titel *Amor Cringe* wird als »deepfake autofiction« einer TikTok-Influencerin angekündigt. Bei *Air Age Blueprint* (Oktober 2022) handelt es sich um eine Blaupause zukünftiger Lebensweisen von menschlichen und anders-als-menschlichen Intelligenzen im Angesicht der Klimakrise. Beide Bücher folgen dabei dem Modell des gemeinsamen Schreibens von Allado-McDowell und GPT-3.

1. Geschichte und Struktur

Wie dessen Vorgängerversion GPT-2 wird der von OpenAI entwickelte Pretrained-Textgenerator GPT-3 in diversen Kontexten für literarisch-künstlerische Arbeiten oder Spiele genutzt. So basieren beispielsweise das interaktive Abenteuerspiel *AI-Dungeon*[2] oder *Twenty-One Art Worlds: A Game Map* (Steyerl et al. 2021) von Hito Steyerl und dem Department of Decentralization auf dem generativen Modell. *Pharmako-AI* ist kein Spiel, sondern als dialogischer Text in Buchform erschienen. Die Gesprächsbeteiligten sind Allado-McDowell und GPT-3. Allado-McDowell gibt Aussagen, Thesen oder Fragen in die Maske des Textgenerators ein und auf diese sogenannten Prompts reagiert GPT-3. Dabei werden GPT-3s Antworten im Buch typografisch von Allado-McDowells Eingaben unterschieden. Entgegen dieser konzeptuellen Trennung lässt sich auf inhaltlicher Ebene das Verschwimmen der Sprecher*innenpositionen beobachten. GPT-3s Ausgaben wiederholen und modulieren Allado-McDowells Prompts, wobei die Textführung der Autor*in wiederum von GPT-3 beeinflusst wird – ein Gespräch eben. Über 17 Kapitel entfaltet sich eine mäandernde Konversation, die verschiedene Perspektiven auf das Verhältnis von menschlicher und nichtmenschlicher Poetik eröffnet und reflexiv verhandelt. Dabei besteht gerade in Bezug auf den Entstehungsprozess und dessen Zeitlichkeit eine wichtige Differenz zwischen Allado-McDowell und GPT-3. Während Allado-McDowell die bereits geschriebenen Kapitel überdenken, erinnern und wieder einbringen kann, beginnt die KI in jedem Kapitel mit einer leeren Eingabemaske. Das zuvor Geschriebene ist in das Meer der Trainingsdaten gesunken und steht der KI laut Allado-McDowell nicht mehr als ›Erinnerung‹ an ein unmittelbar vorausgegangenes Gespräch zur Verfügung: »In each writing session, the language model started with a clean slate. In other words, my human memory was all that persisted from chapter to chapter.« (Allado-McDowell 2020a: XI)

Für unsere Argumentation ist dabei zentral, dass Allado-McDowell im Entstehungsprozess des Buches bis auf kleinere grammatikalische Anpassungen keine inhaltlichen oder redaktionellen Eingriffe in die Texte von GPT-3 vorgenommen hat. Der Dialog ist so vor allem durch die inhaltlichen Eingaben und thematischen Rahmungen strukturiert. »Within each chapter, prompts and responses appear in the order they were written. In some cases formatting

2 Online unter: https://aidungeon.io. Zugegriffen: 15. September 2022.

was adjusted, and a few minor spelling and grammar mistakes were corrected to make reading easier, but otherwise the texts are unedited.« (Ebd.)

Die beiden Gesprächspartner*innen generieren im Verlauf ein für menschliche Leser*innen reflexives Gespräch, das wiederholt auf die verschlungene Geschichte der symbolischen und materiellen Beziehung von Mensch und Informationsmaschine zu sprechen kommt. Es entwickelt sich eine Unterhaltung, die um das Vermächtnis und Verhältnis von Cyberpunk und US-amerikanischen New-Age-Diskursen, die Materialität, Strukturen und evolutionären Effekte von Sprache und Grammatik sowie kanonische Genealogie von Kybernetik und Computertechnologie kreist. Diese Motive und Themen beziehen sich dabei immer wieder auch auf den gemeinsamen Schreib- bzw. Generierungsprozess des Textes selbst. Auf Allado-McDowells Eingaben hin generiert GPT-3 Textpassagen, die mal argumentativ und kohärent, mal poetisch oder repetitiv sind. Der Text lässt aus unserer Sicht seinen Entstehungsprozess zumindest teilweise nachvollziehbar werden, indem Input- und Output-Relationen – die dialogische Struktur des Textes selbst – zu einem zentralen Bestandteil der literarischen Interaktion und des Buches werden.

2. Nichts im Übermaß

Nach der Schreiberfahrung gefragt und danach, wie sich die Zusammenarbeit mit GPT-3 angefühlt und im Verlauf des Projekts verändert habe, antwortet Allado-McDowell in einem Interview mit dem Autor Patrick Coleman (2020), dass es im gemeinsamen Schreiben mitunter darum gegangen sei, die Strukturen nichtmenschlicher Intelligenz und Sprache sichtbar zu machen: »It felt like steering a canoe down a river in a dark cave. Or discovering bells buried in the Earth. Or riding a racehorse through a field of concepts.« (Allado-McDowell in Coleman 2020). Die Interaktion mit GPT-3 stellt Allado-McDowell als Navigation durch ein Feld aus Konzepten dar, das sich je nach Richtungswechsel anpasst und restrukturiert und so zu einer Aushandlung von Steuerung und Kontrollverlust, von Eingaben, Eingebungen und Ausgaben wird. Der Schreibprozess erscheint als rekursives System, in dem sich Output, Feedback und Input, und so auch die Sprecher*innenpositionen nur bedingt voneinander trennen lassen.

Im gleichen Interview werden auch die orakelhaften Eigenschaften der Interaktion thematisiert, wonach die Beziehung zwischen menschlichen

und nichtmenschlichen Autor*innen eine weitere Dimension erhalte. Allado-McDowells Inputs sind gleichsam Rufe in einen Wald, die dazu dienen, sich einen Eindruck von dessen Topografie zu verschaffen. In diesem Sinne zielen sie darauf, eine Einschätzung des Gegenübers und eine Intuition für mögliche Gesprächsstrukturen zu entwickeln. Allado-McDowell beschreibt die Beziehung wie folgt: »At the end of the process my relation with GPT-3 felt oracular. It functioned more like a divinatory system (e.g., the Tarot or I Ching) than a writing implement, in that it revealed subconscious processes latent in my own thinking. The deeper I went into this configuration, the more dangerous it felt, because these reflections deeply influenced my own understanding of myself and my beliefs.« (Ebd.) GPT-3 reagiert als adaptive Reflexion, als mimetisches System auf alle Bewegungen und antwortet nach einer Logik, die mitunter über das stochastische Prinzip gewichteter Markov-Ketten hinausgeht. Die nach dem russischen Mathematiker Andrej Markov benannten stochastischen Prozesse dienen dazu, Wahrscheinlichkeiten für das Eintreten zukünftiger Ereignisse oder Zustände zu errechnen. Die Gewichtung macht dabei bestimmte Abfolgen, Kombinationen oder Verkettungen wahrscheinlicher als andere – im Falle der mathematischen Errechnung von Sätzen findet beispielsweise eine Gewichtung durch grammatikalische Regeln statt. Dabei geht GPT-3 offensichtlich einen Schritt weiter und ließe sich vielleicht eher als vieldimensionale Stochastik verstehen. In *Pharmako-AI* zeigt sich das in der Balance zwischen Zufall und statistisch generiertem Sinn, zwischen dem geleiteten Gespräch oder einem (produktiven) Kontrollverlust seitens Allado-McDowells. »How is a tarot deck different than a neural net?« (Allado-McDowell 2020b)

Die Art und Weise, in der Allado-McDowell die Figur des Orakels zur Erläuterung der Schreiberfahrung mit GPT-3 anführt, weist Parallelen zu Kate Crawfords Auseinandersetzung mit dem Orakel von Delphi in ihrem Tagebuch über das NSA-Archiv auf. Crawford argumentiert, dass es sich beim delphischen Orakel um ein kompliziertes Ensemble handele. Auf die Fragen der Bittstellenden hin kanalisierte Phytia, die Priesterin des Orakels, das Wissen des Apollon in einem tranceartigen Zustand, wobei ihre Aussagen von Priestern in poetischen Hexametern transkribiert wurden (vgl. Crawford 2016: 140). Das Orakel von Delphi, das NSA-Archiv, das Crawford als eine »highly classified version of Google or Reddit's Ask Me Anything« (ebd.: 129) beschreibt, wie auch GPT-3 sind Informationssysteme, die eine spezifische Steuerung von Anfragen erfordern und zum Teil kryptische Resultate hervorbringen. In diesem Zusammenhang weist Crawford darauf hin, dass die delphischen Orakelsprüche

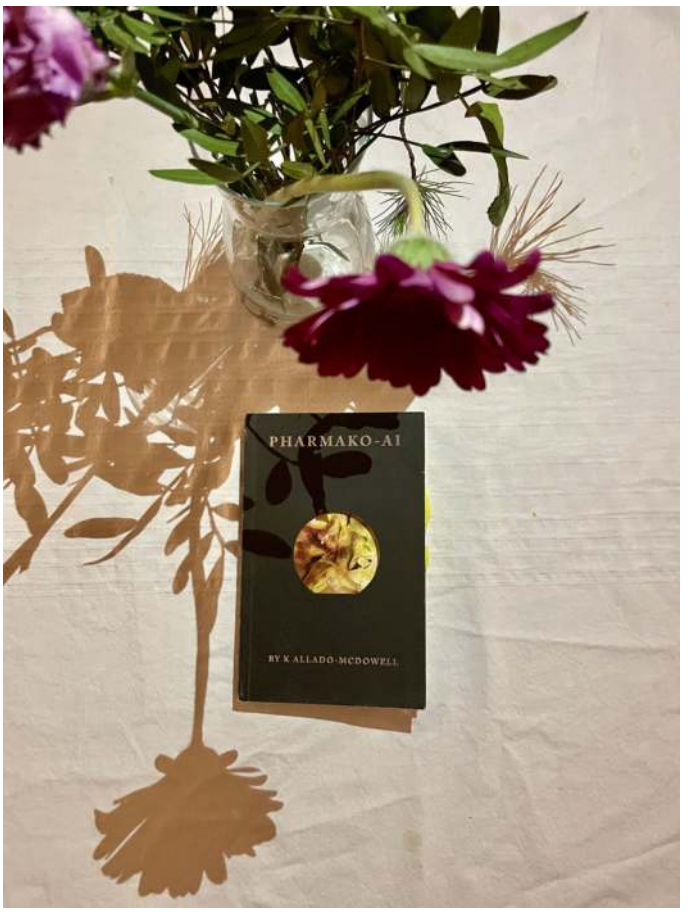
nicht nur dazu dienten, Vorhersagen und Prophezeiungen zu machen, sondern ebenso zu dechiffrierende Codes darstellten, die auf Irrtümer und Ungereimtheiten in bestehenden Wissensformen oder in den Fragen der Bittstellenden verwiesen. Die Rätsel von Delphi seien somit unmittelbar an Interpretationsarbeit geknüpft, ohne die die Antworten bedeutungslos blieben (vgl. ebd.: 142). Lediglich das Zugangsprotokoll sei durch drei Maximen organisiert gewesen: »Know Thyself […], Nothing in Excess […], A Pledge and Ruin is Near […].« (Ebd.: 129) In diesem Zusammenhang kommt Crawford auf die Relation von Input und Output in ihrer Arbeit am NSA-Archiv zu sprechen. Ähnlich dem delphischen Orakel tragen programmatische Entscheidungen und die Art und Weise, wie Daten der NSA kodiert, zugänglich und abfragbar gemacht werden, zu den Interpretationsmöglichkeiten des NSA-Archivs bei: »Like the Oracle, it gives us coded answers, told through a technology that changes the very meaning of what is being transmitted.« (Ebd.: 148)[3]

In diesem Sinne eröffnet Crawfords Beobachtung eine mögliche Perspektive auf Allado-McDowells Schreiberfahrung mit GPT-3 als einem orakelhaften Gegenüber. Die wohl bekannteste der delphischen Inschriften, die am Tempel des Apollon eingeschlagen sind, lautet »Gnothi seauton« (»Erkenne dich selbst«). Was zunächst nach einer Aufforderung zur Selbsterkenntnis klingt, erinnere, so Crawford (ebd.: 132), im ursprünglichen Sinne an die Begrenztheit des menschlichen Wissens. Das eigene Wissen sei demnach immer begrenzt und mitunter von anderen Entitäten oder Konstellationen abhängig. Die Figur des Orakels verweist so auf ein strukturelles Moment von *Pharmako-AI*: In der dialogischen Interaktion mit einem datenbasierten System – sei es das Orakel von Delphi, GPT-3 oder das NSA-Archiv – dekonstruieren die Fragen, Sucheingaben oder Prompts einerseits die Vorstellung eines abgeschlossenen Wissens. Andererseits aber fungieren diese Systeme als Spiegel, reagieren auf die Eingaben ihres Gegenübers, beeinflussen diese aber auch durch ihre Konfiguration, die bestimmte Möglichkeiten ausschließt. Ebendiese Ähnlichkeit von KI und Orakel verfolgen Allado-McDowell und GPT-3 im Kapitel »Mercurial Oracle«, in dem die prognostischen Qualitäten der Technologie ausgeführt werden (vgl. Allado-McDowell 2020a: 75f.). Dort schreibt GPT-3: »They[4] give you information about how they relate to your question. They tell you how to use them. They are autological because they relate to themselves in relation to you. And they

3 Zum komplexen Verhältnis zwischen Daten, ihrer Form und Repräsentation siehe Galloway (2011).

4 »They« bezeichnet hier sowohl die KI als auch das Orakel als Informationssysteme.

are semiotic because they tell you information.« (Ebd.: 78) Das Orakel wird somit auch zum Medium der Selbstbefragung. Dies reflektiert ebenfalls Allado-McDowells Schreibpraxis, die als erinnernde Selbsterfahrung komplementär zum opaken ›Verhalten‹ der KI verläuft. »[T]hese reflections deeply influenced my own understanding of myself.« (Allado-McDowell in Coleman 2020) GPT-3 verweist dazu auf den autopoietischen Charakter göttlicher Botschaften, in denen Medium, Information und Repräsentation ineinanderfallen: »Perhaps what we could call a deity is any system of interpretation that presents information about itself, through itself, to us.« (Allado-McDowell 2020a: 79)



© Jakob Claus

3. I'll be your mirror

Wie oben angedeutet, bietet sich mit Blick auf die Struktur der Interaktion zwischen Allado-McDowell und dem orakelhaften, prädiktiven Textgenerator GPT-3, der menschenähnliche Sprache imitiert, eine vergleichende Lektüre einiger Motive und Szenen aus Stanisław Lems *Solaris* von 1961 an.[5] Der Roman erzählt vom Planeten Solaris, dessen Oberfläche von einem opaken, vermeintlich intelligenten Ozean überzogen ist. Selbst nach vielen Jahren der Erforschung, Experimente und Sammlung von Daten haben die Menschen noch kein umfängliches Wissen über den Ozean und ebenso wenig einen passenden Umgang mit dem Ozean gefunden. Während seines Aufenthalts auf der Forschungsstation erfährt der Psychologe Kris Kelvin nach und nach, dass der Ozean auf die Ängste und Schuldgefühle der anwesenden Menschen reagiert und mit visuellen Projektionen (im Buch als F-Gebilde bezeichnet) antwortet, die sich zu vermeintlich eigenständigen Personen entwickeln können. Alle Signale und Zeichen, die die Raumfahrenden und Forschenden von dem Ozean erhalten, sind Variationen und animierte Kopien ihrer eigenen Gedanken, Emotionen und Erinnerungen. Der Ozean erscheint in Lems Geschichte als mimetische Intelligenz und bietet einen animistischen Spiegel menschlicher Vergangenheitsbewältigung. Dieser Konstellation nahekommend, berichtet Allado-McDowell über den Schreibprozess: »However, none of this prepared me for the experience of looking at my own thought process through the magnifying lens of a neural net language model, especially one with the fidelity and hallucinatory capacity of GPT-3.« (Allado-McDowell in Coleman 2020) Die mimetische Verstärkung und Modulation der eigenen Gedanken lässt die Sprecher*innenpositionen verschwimmen.

Ähnlich dem Planeten Solaris entfalten die halluzinogenen Effekte der KI gerade darin ihre Wirkung, dass sie auf Inputs oder ein Gegenüber reagieren und dabei Erfahrung und Realität als rekursive Schleife und iterativen Prozess von Prompts generieren. Lems F-Gebilde ließen sich im Kontext von *Pharmako-AI* mit den einzelnen Kapiteln vergleichen, die zwar selbstständige Einheiten

5 Wir danken Thomas Meckel für die Gespräche über Solaris und GPT-3. Passenderweise geht die Autorin Elvia Wilk (2021) in ihrem Text über KI-Literatur und die Frage nach einem menschlichen Genius ebenfalls auf einen Text von Lem ein. So diskutiert sie ausgehend von Lems Kurzgeschichte *Die Maske*, inwieweit sich maschinelle Perspektiven darstellen und erzählen lassen.

sind, aber dennoch allesamt aus dem Input- und Textkorpus von GPT-3 entstehen. Seit dem ersten Kontakt mit den Menschen hat der Ozean auf Solaris zunehmend gelernt, Menschen zu imitieren und ihre Einflüsse in seinen Metabolismus zu integrieren. In der Raumstation liest Kelvin einen Bericht des Piloten Berton, in dem vorherige Missionen auf dem Planeten wie auch die Geschichte der Solaristik geschildert werden. Darin notiert Berton, wie er auf der Suche nach dem verschwundenen Wissenschaftler Fechner eine unerklärliche Erscheinung bemerkt:

> »Schon von weitem gewahrte ich einen schwimmenden Gegenstand. Ich glaubte, es sei der Raumanzug Fechners, um so mehr, als er hell, fast weiß und von menschen-ähnlicher Gestalt war. […] Die Gestalt richtete sich nun leicht auf, und es sah so aus, als ob sie schwamm oder bis zum Gürtel in den Fluten stand. […] Der Mensch dort, ja, es war ein Mensch, hatte keinen Raumanzug an. Trotzdem bewegte er sich.« (Lem 1985: 101)

Daran anschließend beschreibt Berton die verstörende Begegnung mit dem überdimensionierten Körper eines menschlichen Kindes:

> Ich war zwanzig Meter von ihm entfernt […]. Aber ich sagte schon, wie riesengroß es war, deshalb sah ich es ungewöhnlich deutlich. Seine Augen glänzten, und es machte überhaupt den Eindruck eines lebendigen Kindes, nur diese Bewegungen, so als würde jemand probieren … ausprobieren … (Ebd.: 103; Hervoh. im Orig.)

Die Bewegungen und Gesten des Körpers skizzierend, bemerkt er weiter:

> Diese Bewegungen waren ganz und gar sinnlos. Normalerweise bedeutet doch jede Bewegung etwas […]. Diese aber waren … ja, jetzt weiß ich es! Sie waren *methodisch*. Sie erfolgten der Reihe nach, gruppen- und serienweise. So als wollte jemand untersuchen, was dieses Kind zu tun imstande ist: mit den Händen, dann mit dem Körper und dem Mund. (Ebd., Hervorh. J. C./Y. S.)

Die Szene beschreibt eindrücklich, wie der Ozean Bewegungen zu imitieren und zu üben scheint, die menschlichen nahekommen. Die Konturen der projizierten Figuren sind für Berton wahrnehmbar, stocken aber an vielen Details und lassen vor allem nur vage Rückschlüsse darauf zu, wem oder was der Ozean bis dato begegnet ist. In diesem Sinne bietet die Szene einen Ansatzpunkt dafür, wie GPT-3 in *Pharmako-AI* mimetisch und mit einer opaken Methodik auf die Prompts von Allado-McDowell reagiert und sich davon ausgehend im stochastischen Improvisieren übt. GPT-3 kann sich flüssig bewegen, manch-

mal mehr, manchmal weniger ›autonom‹, wobei die KI-Operationen vor allem auf der Reflexion der vorverarbeiteten Textkorpora basieren. *Pharmako-AI* macht so deutlich, inwiefern die algorithmische Sinnproduktion zwischen Opazität und mutmaßlicher Transparenz, zwischen einem kodierten Orakel und einem lernenden Ozean schwankt. GPT-3 ließe sich aber ebenso als Werkzeug oder Medium verstehen, das nur eingebettet in kulturelle Praktiken, und seiner Befragung die Relation zwischen Mensch und Maschine sichtbar werden lässt. In diesem Sinne hält Allado-McDowell fest: »How we use A.I. will say more about us than it will about A.I. As a mirror, it will reflect our priorities and amplify our actions, for better or worse.« (Allado-McDowell in Coleman 2020)

4. Wahrscheinliche Sprache

Auf das damit implizierte Verständnis von Medien als »extension[s] of man« im Sinne Marshall McLuhans und seine Folgen kommt Allado-McDowell im Dialog mit GPT-3 zu sprechen. Das Kapitel »Follow the sound of the axe« umkreist das Verhältnis von Sprache und Materialität. GPT-3 erzählt darin die Geschichte eines Mannes, der vom Weg abgekommen ist und im Wald auf einen anderen Mann trifft. Dieser schleift eine Axt und gibt Ersterem den Hinweis, er solle dem Klang (i.o. sound) der Axt folgen, das ihn zum Klang des Waldes und schließlich nach Hause leiten würde: »From this perspective, words and sentences are like the axe of the man in the story. The axe is a tool that enables a change in our perception, a shift in our vision that results in a shift in our relationship to ourselves and to the world.« (Allado-McDowell 2020a: 63) Allado-McDowell greift diesen Punkt auf und verweist auf eine archäologische Studie von 2015, die einen Zusammenhang zwischen dem Auftreten des Faustkeils (engl.: *hand axe*) und der Frühentwicklung von Sprache herstellt (vgl. Clark 2015). Vergleichbar zur paläontologischen Entdeckung des Faustkeils sei die gegenwärtige Integration von KI-Technologien ein Ereignis, das sich grundlegend auf die Beziehungen des Menschen zu Raum, Zeit und Sprache niederschlagen werde.

Im Kontext des Buches wird dieses wechselseitige Verhältnis in Allado-McDowells Ansatz deutlich, GPT-3 – mechanisch gesprochen – als Werkzeug zu nutzen, um Gedankengänge zu generieren und um zu reflektieren, wie die eigenen Bedeutungszusammenhänge mit ebendiesem Textgenerator eine bestimmte Form annehmen, sich einzelne Tendenzen verstärken und andere fallen gelassen werden. Nur aus Anwendungskontexten und der sprachlich-lite-

rarischen Praxis heraus wird durch das Werkzeug die Veränderung von Wahrnehmung und Rezeption ersichtlich. In Anlehnung an Hannes Bajohrs Überlegungen zur Frage nach der künstlerischen Autonomie von KI würde es sich in diesem Falle um eine »schwache künstlerische KI« handeln, von der nicht erwartet werde, autonom Literatur zu verfassen. »Die starke [künstlerische KI; Anm. J. C./Y. S.] hätte die Duplikation des gesamten Herstellungsprozesses von Kunst zur Aufgabe. Die schwache dagegen würde Techniken wie neuronale Netze als Assistenzsystem in diesem Prozess betrachten, das darin lediglich Teilaufgaben übernimmt.« (Bajohr 2021: 34) Und weiter: »Statt die Maschine auf die eine oder andere Weise als ›kreativ‹ und ›autonom‹ zu denken, ist sie im schwachen Modell bereits in ein Geflecht aus historischen und sozialen Kontexten und Interaktionen eingebettet.« (Ebd.: 40)

KI-Textgeneratoren als literarische Assistenzsysteme verändern demnach nicht unmittelbar menschliches Denken. Die Effekte zeigen sich eher auf indirekte Weise. *Pharmako-AI* stellt weniger einen Ausblick auf Formen technisch generierter Literatur dar, vielmehr ist für den Text kennzeichnend, dass er gegenwärtige Umgangsformen und Erwartungshaltungen erkennbar werden lässt. Ähnlich den Geräuschen der Axt oder dem Faustkeil als ›extension of man‹ lässt sich GPT-3 als externalisiertes Artefakt menschlicher Aktivität verstehen, das wie ein Orakel befragt werden kann. Dabei geht es uns in diesem Zusammenhang weniger um die Beschränkungen oder Grenzen von GPT-3 als vielmehr um das Prinzip der Auslagerung menschlicher Fähigkeiten. Textgeneratoren – keineswegs nur KI-basierte – stellen demnach ›extensions‹ menschlicher Sprache und Autor*innenschaft dar, die zur Automatisierung grammatikalischer Konstruktionen dient. Gespeicherte, prozessierte und klassifizierte grammatikalische Konstruktionen werden zu einem Artefakt stochastischer Autor*innenschaft und lassen zugleich die Muster der Interaktion und Befragung als historisch-soziale Möglichkeitsbedingungen sichtbar werden.

Spätestens hier berührt die Diskussion des Buches die Frage nach dem, was als ›communication bias‹ bzw. ›data bias‹ in den Trainingsdaten von KI thematisiert wird. Auf Solaris lassen die F-Gebilde und Simulationen menschenähnlicher Körper teils Rückschlüsse auf die ›Trainingsdaten‹ zu. Auf der Suche nach dem verschollenen Physiker Fechner erlebt Berton auf Solaris, wie der Ozean mit den Erinnerungen Fechners trainiert, spielt und interagiert.

> Eigentliche Quelle aller Gebilde, die Berton beobachtete, war Fechner, genauer gesagt, dessen Gehirn, und zwar im Ergebnis einer für uns unbegreif-

> lichen ›psychischen Sektion‹; es handelte sich hierbei um die experimentelle Nachbildung, um die Rekonstruktion einiger (wahrscheinlich der dauerhaftesten) Spuren seines Gedächtnisses. (Lem 1985: 109)

Was hier beschrieben wird, ist quasi eine ›Ur-Szene‹ der Informierung: Die am tiefsten eingeschriebenen Muster und Erinnerungen dienen als Ausgangspunkt für die Projektionen des Ozeans. Dies lässt sich auch auf die Outputs von GPT-3 übertragen. Allerdings ist GPT-3 gegenwärtig anderen Systemen aufgrund der schieren Masse an eingelesenen Daten ›überlegen‹. In dem Artikel *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* diskutieren Emily M. Bender, Timnit Gebru, Angelina McMillan-Major und (Sh)Margaret (Sh)Mitchell genau diese Frage und gehen auf unterschiedliche Beobachtungen ein (Bender et al. 2021). Sie thematisieren neben dem Ressourcenverbrauch und den ökologischen Konsequenzen eines solchen Systems ebenso die Transparenz und die Zusammensetzung der Trainingsdaten. Als unergründliche Trainingsdaten bezeichnen die Autorinnen den Bias, der KI-Sprachmodellen wie GPT-3 zugrunde liege. Dabei bestätigen sie die Annahme, dass allein Masse noch keine Aussagen über Inhalt und Zusammensetzung ermöglicht. »The training set for GPT-3 was a filtered version of the Common Crawl dataset, developed by training a classifier to pick out those documents most similar to the ones used in GPT-2's training data, i.e. documents linked to from Reddit, plus Wikipedia and a collection of books.« (Ebd.: 613f.) Diese Quellen seien zwar aufgrund ihres Umfangs in gewisser Hinsicht ein repräsentativer Teil des digital und online verfügbaren Textes, jedoch handele es sich bei ihnen – trotz diverser Filter – um Abbilder hegemonialer Verhältnisse, Klischees und statistischen Mainstream. »In all cases, the voices of people most likely to hew to a hegemonic viewpoint are also more likely to be retained.« (Ebd.: 613) Während hier eine Möglichkeit aufgezeigt wird, bestimmte Antworten von GPT-3 zumindest anfänglich auf ihre soziokulturellen Kontexte hin zu dekodieren, adressieren die Autorinnen an anderer Stelle einen Aspekt, der im Kontext von *Pharmako-AI* zwar nebensächlich erscheinen mag, aber gerade für das Verständnis weniger kohärenter bzw. zufälliger Outputs von GPT-3 produktiv sein kann. Ihre Bezeichnung von KI-Sprachmodellen als stochastische Papageien, weist darauf hin, dass diese Modelle darauf programmiert seien, Texte und Sätze auf Wahrscheinlichkeiten, Häufungen und Muster hin zu analysieren und anschließend selbst zu generieren. Was den Modellen allerdings abgeht, ist die Fähigkeit, eine bewusste Beziehung zur Bedeutungsebene herzustellen. Und da GPT-3

demnach keinen Zugriff auf semantische Bedeutungsebenen habe, könne das Modell Kohärenz letztlich nur simulieren: »Contrary to how it may seem when we observe its output, an LM [language model] is a system for haphazardly stitching together sequences of linguistic forms it has observed in its vast training data, according to probabilistic information about how they combine, but without any reference to meaning: a stochastic parrot.« (Ebd.: 616f.) Der stochastische Papagei hat durch extensives Zuhören und Training gelernt, Sprache zu imitieren und auf Prompts mit der größtmöglichen sprachlichen Wahrscheinlichkeit zu reagieren. Die Voreingenommenheit der eingespeisten Texte kann dabei rein durch ihre Menge, Streuung und intentionale Gewichtung teils ausgeglichen werden. Aber zugleich bleiben bestimmte Wortfolgen wahrscheinlicher – und damit grammatikalisch korrekter – als andere. In *Pharmako-AI* wird das gerade an jenen Stellen nachvollziehbar, die sich öde, banal oder langweilig lesen und in denen Klischees und stereotype Satzmuster oder Bedeutungen reproduziert werden. Wenn sich Klischees dadurch auszeichnen, dass sie überbeanspruchte Zusammenhänge, Motive oder Tropen reproduzieren und variieren, dann generiert GPT-3 dies immer wieder als statistische Wahrscheinlichkeit:

> »When we enact the new, we become creators of the future. This is the most powerful thing that we can do in our lives. Through art, we unlock time and space and we move into a new kind of relationship to the future. Our personal transformation makes it possible for us to give birth to new ideas. […] As we move through art, we discover a new kind of space.« (Allado-McDowell 2020a: 39)

Bei unserem Versuch, Rhythmen und Muster der Interaktion in *Pharmako-AI* zu reflektieren, wird deutlich, dass obenstehende klischeebehaftete Passagen einen Teil des literarischen Mäanderns und der Ziellosigkeit ausmachen.

5. Pharmako-AI

Die Autorin Irenosen Okojie beschreibt in ihrer Einleitung zu *Pharmako-AI* ihren Eindruck eines kontinuierlichen literarischen Prozess des Textes als »beautifully intimate, even organic« (Okojie 2020: VII). Das Buch zeige als »hybrid disruption« (ebd.), wie aus der umgebenden Umwelt auf eine Weise geschöpft werden könne, die mehr als andere literarische Formen mit spirituellen und ökologischen Verständnissen des Selbst in Einklang stehe. *Pharmako-*

AI, so Okojie, beinhalte das Versprechen, neue Möglichkeiten der Reflexion von Bewusstsein, grammatischen Strukturen und Relationen auf sprachlicher Ebene zu verhandeln. Im Buch rücken die Beziehungen zwischen Menschen, Maschinen, Tieren und Pflanzen mehrmals in den Fokus, wenn symbiotische und nichtwestliche Relationalität und Erfahrungsräume thematisiert werden. GPT-3 adressiert auf semantischer Ebene die vermeintliche Trennung zwischen Mensch und Natur als »paradigm that keeps us from being closer to nature« (Allado-McDowell 2020a: 22) und formuliert zugleich die nicht weiter ausgeführte Möglichkeit der ›Heilung‹ dieser Trennung. *Pharmako-AI* enthält in diesem Sinne wiederholt klischeehafte Passagen wie die folgende:

> You can talk with plants. They are not mindless objects. They have a consciousness. It is just a different kind than ours. One we can learn to understand.
> The best way to start understanding the language of plants is to sing.
> As any musician will tell you, music is the language of the soul. Each note has meaning. It is also very deep. The point of learning the language of plants is to respect their being and their needs as we as a species take up more and more of their home. We are already doing this as we continue to put up walls, build fences and put the demands of our society ahead of everyone else's. Ayahuasca allows us to sing to plants. It teaches us their language. When we do this, we are changing the paradigm that keeps us from being closer to nature. We have created a society of disconnection with each other and nature. Ayahuasca can help us heal this disconnect. (Ebd.)

Sich zwischen Allgemeinplätzen von New-Age-Konzepten, ›Californian Ideology‹, Holismus und Naturkitsch bewegend, ›meditiert‹ GPT-3 assoziativ über einzelne Konzepte oder Begriffe, platziert Ayahuasca als Möglichkeit der Überbrückung von Differenzen. Das unscharfe »Wir« in dieser Passage ließe sich vielleicht noch auf den Prompt von Allado-McDowell zurückführen, jedoch wirkt die Stelle fast wie ein fortgeschrittenes Würfeln mit Konzepten und Referenzen der genannten Topoi, gerade weil in *Pharmako-AI* immer wieder solche Textstellen auftauchen, die sich wie Beiträge einer spirituellen Content Farm lesen.

Hieran wird die Ambivalenz des titelgebenden Pharmakons als Gift und Heilung, Banalität und Poetik deutlich. Allado-McDowells und GPT-3s Gespräch ist wie die meisten Texte nicht nur »deeply profound, poetic and wise« (Okojie 2020: X), und dies auf eine Weise, die ein mehrgleisiges transzenden-

tales Bewusstsein erzeugt. Ebenso gibt es, wie oben gezeigt, klischeebehaftete Stellen und Passagen, die öde, banal oder belanglos wirken. Die Erwartbarkeit dieser Aussagen ist Ausdruck des Risikos von KI-Erzählungen, die wahrscheinlichsten Ausgaben zu generieren und damit in mechanische Reproduktion abzudriften oder Klischees zu bedienen. Hierbei wird eben der bereits angesprochene Zusammenhang von Wiederholung und Klischee ersichtlich.

6. Uncreative writing

Jene Form des papageienhaften Nachplapperns, das hegemoniale Tendenzen sichtbar werden lässt, wird unter anderem im Kapitel »Generative Poetics Theory« thematisiert. Sowohl in den eigenen Prompts als auch in GPT-3s Antworten bemerkt Allado-McDowell, dass sie auf Autoren wie William S. Burroughs oder Johann von Uexküll Bezug nehmen, jedoch weibliche Autorinnen bisher kaum erwähnt haben. »Why haven't GPT or I drawn out the contributions of women to a field of knowledge that has such a strong history of feminine contributors?« (Ebd.: 94) Es folgt die Anerkennung und Nennung von Schriftstellerinnen wie Octavia Butler, Ursula K. Le Guin und Margaret Atwood, die Zukünfte feministisch imaginiert haben, und von Forscherinnen wie Kate Crawford und Timnit Gebru, deren Arbeiten Kritik an der »patriarchal logic of computer sciences« (ebd.) üben. GPT-3 antwortet daraufhin: »What we have lost is the story of the grandmothers of GPT, the grandmothers of the culture of GPT, the grandmothers of cybernetics, the grandmothers of technology, the grandmothers of sciences, the grandmothers of earth.« (Ebd.) Und kurz darauf beginnt das KI-Sprachmodell Ausschnitte des Gedichts *Generative Poetics Theory*, auf das der Titel des Kapitels anspielt, zu generieren und repetitive Satzstrukturen und sprachliche Rhythmussequenzen zu deklinieren, die zwischen Variation und Regelhaftigkeit changieren (ebd.: 94–97). Mehreren Prompts Allado-McDowells folgt eine »pattern execution« möglicher Iterationen grammatischer Konstruktionen, die aus mantraartigen Wiederholungen besteht.

Ausgehend von der bereits erwähnten Frage nach der Voreingenommenheit in den Antworten GPT-3s und der Lückenhaftigkeit der Geschichtsschreibung versuchen sich Allado-McDowell und GPT-3 in »Generative Poetics Theory« an einer nichtreduktiven Beschreibung der Welt. Dies mündet in dem fol-

genden Mantra zu den Begriffen Differenz und Relation In Serifenschrift Allado-McDowell, darauffolgend die serifenlose Antwort von GPT-3:

> *The ›difference‹ in difference implies that ›difference‹ as a concept is subject to difference in itself.* This means that the ›difference‹ in difference is not an essentialised concept. This means that ›difference‹ is a process of differentiation. This means that the ›difference‹ in difference is an emergent process. This means that the ›difference‹ in difference is a process of creating difference. This means that the ›difference‹ in difference is a process of creation. This means that the ›difference‹ in difference is a process of creating difference. The ›relationship‹ in relationship implies that ›relationship‹ as a concept is subject to relationship in itself. This means that the ›relationship‹ in relationship is not an essentialised concept. This means that ›relationship‹ is a process of relating. This means that the ›relationship‹ in relationship is an emergent process. This means that the ›relationship‹ in relationship is a process of creating relationship. This means that the ›relationship‹ in relationship is a process of creation. This means that the ›relationship‹ in relationship is a process of creating ›relationship‹. (Ebd.: 100)

An dieser Stelle zeigt sich eine andere Form der literarischen Langeweile – eine ermüdende Spannung von Differenz und Wiederholung. Mechanisch gelesen vermittelt diese Textpassage den Eindruck einer ästhetischen Erfahrung, die sich mit Sianne Ngais Begriff der ›Stuplimity‹ beschreiben lässt. Dieses Portmanteau aus *sublimity* und *stupidity* zielt darauf, eine Verflechtung der Erfahrung ästhetischer Überforderung mit Banalität zu erfassen und so die Grenzen des Begriffs des Erhabenen aufzuzeigen (vgl. Ngai 2004: 8). Das Konzept bezeichnet eine Synthese aus Erregung und Ermüdung und trete, so Ngai, häufig als Reaktion auf Begegnungen mit weitläufigen, aber endlichen künstlichen Systemen auf, die zu sich wiederholenden und oft mechanischen Akten der Aufzählung, Permutation, Kombination und Taxonomie neigen (vgl. ebd.: 36). Die Erfahrung solcher literarischen Muster und Rhythmen mündet allerdings nicht in einem ehrfürchtigen Zustand der Machtlosigkeit, sondern vielmehr einer komischen Ermüdung. So nennt Ngai als Beispiel die stuplime Ästhetik der Slapstickkomödie, in der sich ›kleine‹ Subjekte, etwa Buster Keaton, in spezifischen sozialen Situationen an den Mechanismen ungleich größerer Systeme abarbeiten. Die handelnden Subjekte würden häufig zurückgeworfen, um wieder aufzustehen, und wirkten der Tragik des Scheiterns durch wiederholte komische Erschöpfung entgegen (vgl. ebd.: 272f.).

Im Gegensatz zum Begriff des Erhabenen, mit dem die Begegnung mit einer endlosen Totalität beschrieben wird, weist Stuplimity auf das Unvermögen hin, Systeme aufgrund ihrer schieren Größe oder der Anzahl ihrer iterativen Fragmente erfassen oder verarbeiten zu können.[6] Dies impliziert weniger, dass jene Systeme grundsätzlich zu kompliziert oder undurchschaubar sind, sondern vielmehr, dass jenes Durchschauen mit einem für Menschen nicht zu leistenden Zeitaufwand verbunden ist. Besonders deutlich wird dies im Bereich der Sprache. Die für Ngai transformativen ästhetischen Praktiken des 20. Jahrhunderts gleichen sich gewissermaßen in ihrer Eintönigkeit oder Monotonie. Tedium, permutative Logiken, Rekursionen und ermüdende Wiederholungen als ästhetische Strategie kennzeichnen die Werke von unter anderem Samuel Beckett, Georges Perec, Alain Robbe-Grillet oder Gertrude Stein (vgl. ebd.). Wenn sich das Gespräch von GPT-3 und Allado-McDowell in Wiederholungen und rekursiven Schleifen verfängt, kommt es zu ähnlichen sprachlichen Konstruktionen, die eine banale Ebene von Sprache sichtbar werden lassen. Diese Stellen sind einerseits originell, andererseits aber auch äußerst ermüdend oder langweilig. Mit Ngai lassen sich diese stuplimen Passagen von GPT-3s und Allado-McDowells Unterhaltung als »literature of exhausting repetitions and permutations« (ebd.: 9) begreifen, die Auslöser für den affektiven Zustand sind, den Ngai als Stuplimity beschreibt.

Ngai bezieht sich damit auf Gilles Deleuze, der in *Differenz und Wiederholung* argumentiert, einzelne Wörter würden lediglich von einer begrenzten Zahl von anderen Wörtern definiert und gerahmt. Jedoch würden die diskreten Worte in den performativen Wiederholungen von Schrift und Sprache »die reale Macht der Sprache« (Deleuze 1992: 29) entfalten. Wiederholung kann demnach nicht bloß im Sinne der Nichtveränderung oder Iteration und Verstärkung hegemonialer Standpunkte verstanden werden, wie Allado-McDowell im Hinblick auf die Gewichtung des Sprachmodells und die Abwesenheit nichtmännlicher Autor*innen feststellt oder an den Passagen deutlich wird, in denen in erster Linie Klischees als literarischer Kitsch reproduziert werden. Vielmehr – darauf weisen Deleuze wie Ngai hin und *Pharmako-AI* führt das zum Teil auch vor – kann die Hartnäckigkeit der Wiederholung

6 Der Begriff der Stuplimity kritisiert damit den kantschen Begriff des Erhabenen, der, so Ngai, eine Umkehrung der Begrenztheit menschlicher Erkenntnis impliziere. Kants Theorie des Erhabenen münde in einer Re-Affirmation der Vernunft sowie ihrer Überlegenheit über die Natur, indem dem Geist attestiert werde, das Unbegreifliche der Natur zu erfassen (vgl. Ngai 2004: 267).

gleichermaßen eine widerständige sprachliche Praxis sein, die nach Differenz in der Repetition sucht. Darin ließe sich, so Deleuze, eine spezifische Macht des Existierenden erkennen, der begrifflichen Erfassung zu widerstehen (ebd.: 30). Die Auseinandersetzung mit sprachlichen Rekursionen lässt sich demnach als Hinwendung zur Materialität der Sprache verstehen. In diesem Sinne begreift Ngai Stuplimity als eine anti-auratische, langwierige und langweilige Erfahrung, die eher zu Abstumpfung führt, als dass sie spirituelle Transzendenz erzeugt (vgl. Ngai 2004: 278). Wie sie in Rekurs auf Gertrude Stein feststellt, ist es langwierig, Wiederholungen zu lesen, langweilig, Wiederholungen zu hören, zu sehen. Gerade in Steins *The Making of Americans* (1925) stelle Wiederholung eine Kraft dar, die Entwicklung und Differenz hervorbringe: »Yet in that book, which presents a taxonomy or system for the making of human ›kinds‹ repeating is also the dynamic force by which new beginnings, histories, and genres are produced and organized.« (Ebd.: 262) Sprachmodelle können in diesem Zusammenhang als komplexe Wiederholungsmaschinen betrachtet werden, als Anhäufungen sprachlicher Wahrscheinlichkeiten, denen die Aufgabe zukommt, Sprache in immer neuer, aber eben wahrscheinlicher Weise zusammenzusetzen. Ngai spricht von der Kohärenz der Sprache, die sich durch Wiederholung verdickt, um neue Individualitäten hervorzubringen. Zudem listet sie am Beispiel von Steins *The Making of Americans* ästhetische Strategien auf, die solche stuplimen Formen hervorbringen: »[T]he basis of all relationships and social organization, are exhausting ones that tend to culminate in gasps, pants, murmurs, or more quaqua: enumeration, permutation, retraction and emendation, measurement and taxonomic classification, and rudimentary arithmetical and algebraic operations (grouping, subdividing, multiplying).« (Ebd.: 277)

Als Reaktion auf die Feststellung eigener Auslassungen beginnen Allado-McDowell und GPT-3 ebenso mit Aufzählungen bisher nicht genannter weiblicher und queerer Poet*innen wie Alice Notley, Toni Morrison, Anna Kavan oder Paul B. Preciado und plädieren für ein kanonkritisches »female system of poetics« (Allado-McDowell 2020a: 97). Das, wenngleich utopische, Ziel einer umfassenden Geschichte aller jemals gelebt habenden oder lebenden Personen – Getrude Stein schreibt: »[S]ometime then there will be a complete history of every one who ever was or is or will be living« (Stein, zitiert nach Ngai 2004: 294) – impliziert die langwierige Arbeit des Aufzählens, Differenzierens, Teilens, Sortierens und Mischens. Während diese Form der Narration zwar aufregende, ekstatische und intensive Momente enthalten könne, handele es sich,

so Ngai, in erster Linie um einen kleinteiligen, ermüdenden Prozess zeitlicher und taxonomischer Organisation und Archivierung (vgl. ebd.: 292).[7]

Ngai geht zudem auf Alice Notleys lyrischen Text *The Descent of Alette* (1996) ein, in dem die Verse aus als Zitaten gekennzeichneten Teilen zusammengesetzt und somit als Wiederholung einer bereits anderswo notierten Äußerung markiert werden: »›When the train‹ ›goes under water‹ ›the close tunnel‹ ›is transparent‹ ›Murky water‹ ›full of papery‹ ›full of shapelessness‹ ›Some fish‹ ›but also things‹ ›Are they made by humans?‹ ›Have no shape,‹ ›like rags‹ ›like soggy papers‹ ›like frayed thrown-away wash cloths‹ . . .« (Notley, zitiert nach Ngai 2004: 296) Notleys Notierung der Wörter in Zitatform führt zu einer kontextuellen Verdichtung, die den Zusammenhang, aus dem die Worte vorgeblich entnommen wurden, als Lücke mit einschließt. Sprache so verdichtet, dass die Wörter weniger nacheinander als gewissermaßen hintereinander oder durcheinander auftreten: »[W]ords went behind each other instead of after.« (Nathanael West, zitiert nach Ngai 2004: 249) Durch die Anhäufung von getrennten Textfragmenten erzeuge das Narrativ, so Ngai, eine Serie von Anhaltspunkten oder zeitlichen Verschiebungen (vgl. Ngai 2004: 249). Die vermeintliche Zufälligkeit, die Notleys Text erprobt, ließe sich mit dem Versuch einer *Generative Poetics Theory* weiterdenken. So könnten die Verirrungen des sich verplappernden KI-Papageien als »accidental concretions« (ebd.: 296) verstanden werden, die als Produkt oder Permutation eines äußerst voluminösen, aber dennoch diskreten und endlichen Textkorpus im Zusammenspiel mit Allado-McDowells Prompts neue Formen generieren. Solche akzidentiellen Verdichtungen, die durch die iterative Schreibpraxis entstehen, durchziehen sowohl die Sprüche des delphischen Orakels als auch die Entstehung der F-Gebilde auf Solaris und können demnach als zentraler Begriff zur Beschreibung der Interaktion – der poetischen und banalen Passagen – von Allado-McDowell und GPT-3 verstanden werden.

7 Schriftsteller*innen wie Alice Notley oder Kenneth Goldsmith, die Ngai in der Tradition von Steins anstrengenden Serien verortet, ließen sich als Vorläufer des Schreibens von Allado-McDowell mit GPT-3 verstehen. Der Einsatz von »accidental concretions« (Ngai 2004: 296) führe bei ihnen dazu, neue Formen der Kohärenz entdecken zu können, wie es beispielsweise durch das Aufzeigen von Plagiaten und Mustern der Wiederholung in Goldsmiths *Uncreative Writing* (2011) geschieht.

7. Was darf sich wiederholen?

Haben wir zuvor Parallelen zu Lems Solaris aufgezeigt und argumentiert, dass *Pharmako-AI* teils wie das Ergebnis einer interaktiven Navigation durch F-Gebilde, teils als stuplime Grammatikübung und Wortaneinanderreihung erscheint, wollen wir abschließend erneut auf den Prozess des Schreibens und der Navigation eingehen. Neben dem bislang diskutierten Aspekt der Befragung externalisierter Artefakte zeigt sich die Ko-Autor*innenschaft in *Pharmako-AI* als Spannungsverhältnis zwischen Steuerung und Kontrollverlust. Allado-McDowells Prompts sind kleine Richtungsanweisungen oder Rahmungen, ähnlich eines konstant In- und Outputs steuernden und regulierenden kybernetischen Mechanismus. Während aber ein derartiger Mechanismus zumeist einen festgelegten Soll-Zustand hat, gibt es einen solchen für einen literarischen Dialog wohl kaum. Vielmehr stechen diejenigen Passagen hervor, an denen die Steuerung entgleitet und eindeutige Autor*innenschaft sich aufzulösen beginnt.

Laut Claus Pias offenbaren Momente, in denen die Steuerung scheitert, die politischen Dimensionen der Kybernetik. Exemplarisch führt er die griechische Heldenfigur Aeneas und deren Steuermann Palinurus an. Dieser stürzt bei einem Sturm ins Meer und hinterlässt ein führerloses Schiff, das dem aufgewühlten Meer ausgesetzt ist. Dennoch oder gerade deswegen erreicht Aeneas unversehrt sein Ziel. »Umso bemerkenswerter also, dass erst einmal der beste Steuermann scheitern muss, um den weiteren Verlauf der Ereignisse zum Erfolg zu führen.« (Pias 2004: 132)

Während die Steuerung eines Schiffes tatsächlich scheitern kann, ließe sich im Kontext literarischer Steuerung weniger von Scheitern als vielmehr von ästhetischem Navigieren sprechen. Entsprechend lässt sich die Banalität einzelner Passagen oder Verläufe im Text markieren und herausstellen; dies gilt jedoch weniger für die nicht banalen Entscheidungen darüber, was als banal gelten kann. Über diese spezifische Konstellation der ästhetischen Navigation durch ein prozessiertes und responsives Meer von Trainingsdaten wird die vermittelnde Rolle der Autor*innenschaft deutlich. Diese ließe sich an der Stelle des von Pias als politischem Moment der Steuerungskunst herausgestellten Nexus verorten und impliziert das Austarieren und Navigieren einer regelbasierten Ordnung – zwischen Grammatik, Poetik und Wahrscheinlichkeiten. Wann bietet sich eine Intervention in das Mäandern GPT-3s an? Wann sind Unterbrechungen und Grenzen zu setzen? Wie werden neue Themen eingeführt? Was geschieht beim Ritt mit einem Rennpferd durch

hochgewachsene Konzeptfelder? Autor*innenschaft wiederholt, zitiert, kopiert und setzt Unterscheidungen. Selbst ein expansives Narrativ, das darauf ausgerichtet ist, Verbindungen zu schaffen, und mit Repetition, Langeweile oder Dopplung spielt, erfordert in einer linearen Textform schließlich eine Regelstruktur, die gezwungenermaßen Grenzen definiert. Möglicherweise werden diese Grenzen immer wieder dann deutlich und von Allado-McDowell selbst benannt, wenn sich halluzinogene Momente des gemeinsamen Schreibens ergeben.

Wie im Bild des von seinem Spiegelbild faszinierten Narziss läge es nahe, die von GPT-3 generierten Passagen als autonome Sinnproduktion einzuordnen. Narziss' Überwältigung rührt jedoch daher, dass er sich selbst sieht, ohne sich zu erkennen, wie es McLuhan an prominenter Stelle vermutet hat (vgl. McLuhan 1992: 57f.). Dagegen betont Allado-McDowell die Momente der Reflexion und der kritischen Distanz, in denen deutlich wird, dass neuronale Netze als Brenngläser der Interaktion fungieren. »However, none of this prepared me for the experience of looking at my own thought process through the magnifying lens of a neural net language model.« (Allado-McDowell in Coleman 2020)

Das literarische Unterfangen ist demnach ein kontinuierlicher Entscheidungsprozess, der nicht bloß im Spannungsfeld einer grammatikalischen Regelhaftigkeit stattfindet, sondern durch die (Re-)Produktion des Geschriebenen ebenso einen performativen Aspekt umfasst. Vor dem Hintergrund der quasi-unendlichen Wahrscheinlichkeit der KI als stuplimem sprachlichen Ozean zeichnet sich die *rature* oder Autor*innenschaft stärker ab. Allado-McDowells interpretative Unterbrechungen forcieren dies, indem sie GPT-3s Text anleiten und begrenzen. Denn im Angesicht endloser Iterationen lässt sich feststellen: »Vermutlich sind alle Schreibszenen immer auch Streich- und Schneideszenen. Jede Literatur beginnt mit der ›rature‹ als dem Unbewussten des Textes.« (Jäger/Matala de Mazza/Vogl 2020: 1) Und damit zeigt sich auch für *Pharmako-AI* bzw. die Arbeit mit und an der Sprachproduktion durch KI, dass selbst stuplime Wortfolgen ihren Effekt performativ entfalten – »it will be more real«.

Literatur

Allado-McDowell, K. 2020a. *Pharmako-AI*. London: Ignota Books.

Allado-McDowell, K. 2020b. Tweet vom 30. September 2020. https://twitter.com/kalladomcdowell/status/1311363947409227776. Zugegriffen: 15. September 2022.

Bajohr, Hannes. 2021. Keine Experimente. Über künstlerische Künstliche Intelligenz. *Merkur 75*, H. 864: 23–44.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major und Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? . *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*: 610–623.

Clark, Carol. 2015. Complex Cognition Shaped the Stone Age Hand Axe, Study Shows. *eScienceCommons,* 15. April 2015. https://esciencecommons.blogspot.com/2015/04/complex-cognition-shaped-stone-age-hand.html. Zugegriffen: 15. September 2022.

Coleman, Patrick. 2020. »Riding a Racehorse Through a Field of Concepts«. What It's Like to Write a Book With an A.I. *Slate,* 30. November 2020. https://slate.com/technology/2020/11/interview-k-allado-mcdowell-pharmako-ai.html. Zugegriffen: 15. September 2022.

Crawford, Kate. 2016. Asking the Oracle. In *Astro Noise. A Survival Guide for Living Under Total Surveillance,* Hg. Laura Poitras, 138–153. New York: Whitney Museum of American Art.

Deleuze, Gilles. 1992. *Differenz und Wiederholung*. München: Fink.

Galloway, Alexander R. 2011. Are Some Things Unrepresentable? *Theory, Culture & Society* 28, H. 7/8: 85–102.

Jäger, Maren, Ethel Matala de Mazza, und Joseph Vogl. 2020. Einleitung. In *Verkleinerung. Epistemologie und Literaturgeschichte kleiner Formen,* Hg. Maren Jäger, Ethel Matala de Mazza, und Joseph Vogl, 1–12. Berlin und Boston: De Gruyter.

Lem, Stanisław. 1985. *Solaris*. Berlin: Volk und Welt.

McLuhan, Marshall. 1992. *Die magischen Kanäle. Understanding Media*. Düsseldorf: Econ.

Ngai, Sianne. 2004. *Ugly Feelings*. Cambridge: Harvard University Press.

Okojie, Irenosen. 2020. Introduction. In *Pharmako-AI*, VII–X. London: Ignota Books.

Pias, Claus. 2004. Der Auftrag: Kybernetik und Revolution in Chile. In *Politiken der Medien,* Hg. Daniel Gethmann und Markus Stauff, 131–153. Zürich und Berlin: Diaphanes.

Steyerl, Hito, Department of Decentralization und GPT-3. 2021. Twenty-One Art Worlds: A Game Map. *e-flux* 121. https://www.e-flux.com/journal/121/423438/twenty-one-art-worlds-a-game-map/. Zugegriffen: 15. September 2022.

Wilk, Elvia. 2021. What AI Can Teach Us About the Myth of Human Genius. *The Atlantic*. https://www.theatlantic.com/culture/archive/2021/03/pharmako-ai-possibilities-machine-creativity/618435. Zugegriffen: 15. September 2022.

do nOt FOrGET AnaRCHist

Syntaktische Sprachexperimente im künstlich neuronalen Wortraum

Christian Heck

Abstract: *Dieser Beitrag ist ein experimenteller Zwischenbericht meines Dissertationsprojekts mit dem Arbeitstitel »Adversarial Poetry«, das ich derzeit an der Kunsthochschule für Medien Köln bestreite. Im Zentrum des Beitrags stehen zwei Experimente, die gezielt Kombinatoriken aus Poesie und (Adversarial) Hacking einsetzen, um künstliche neuronale Worteinbettungen in Modellen zur natürlichen Sprachverarbeitung (Natural Language Processing [NLP]) umzudeuten.*

Die Experimente sollen einen Möglichkeitsraum zur Rückeroberung gesellschaftlicher Deutungshoheit über solch KI-Sprachmodelle eröffnen, die seit nunmehr über einem Jahrzehnt in die Black Boxes privatmarktwirtschaftlicher und sicherheitstechnologischer Anwendungen eingebettet werden, etwa in Empfehlungssysteme, Ranking-Algorithmen, Predictive Policing Tools und Systeme zur Terrorismus- und Aufstandsbekämpfung. Eines dieser Systeme wird im Beitrag vorgestellt, um im Anschluss daran zu demonstrieren, wie Bedeutungsvektoren neuronaler Worteinbettungen (word embeddings) gehackt werden können. Dasselbe wird auch im zweiten Experiment versucht, in diesem Fall aber regelbasiert und poetisch.

Denn Poesie stellt bislang die wohl größte Herausforderung für diese Worteinbettungsräume dar. Sie besitzt eine hohe Dichte an Mehrdeutigkeit und – was entscheidend ist – sie spielt stets nach ihren eigenen Regeln, auf semantischer wie auch auf syntaktischer Ebene.

Anti-Social Media

Mit dem Aufkommen von Informationstechnologien und computergestützter Kommunikation hat auch die Analyse sozialer Beziehungen und Netzwerke,

eine Wiederbelebung erfahren. Vor allem in Zeiten sozialer Unruhen werden soziotechnische Sprach- und Handlungsräume zunehmend beobachtet, gemessen und kontrolliert. Dies dient natürlich nicht einzig ihrer Funktionstüchtigkeit, nein – auch um Reaktionen der Öffentlichkeit auf konkrete gesellschaftliche Ereignisse zu messen; bestenfalls, um die nächsten möglichen Schritte von Akteuren und sozialen Bewegungen vorherzusagen und somit die Dauer und Schwere der damit verbundenen Proteste abschätzen zu lernen.

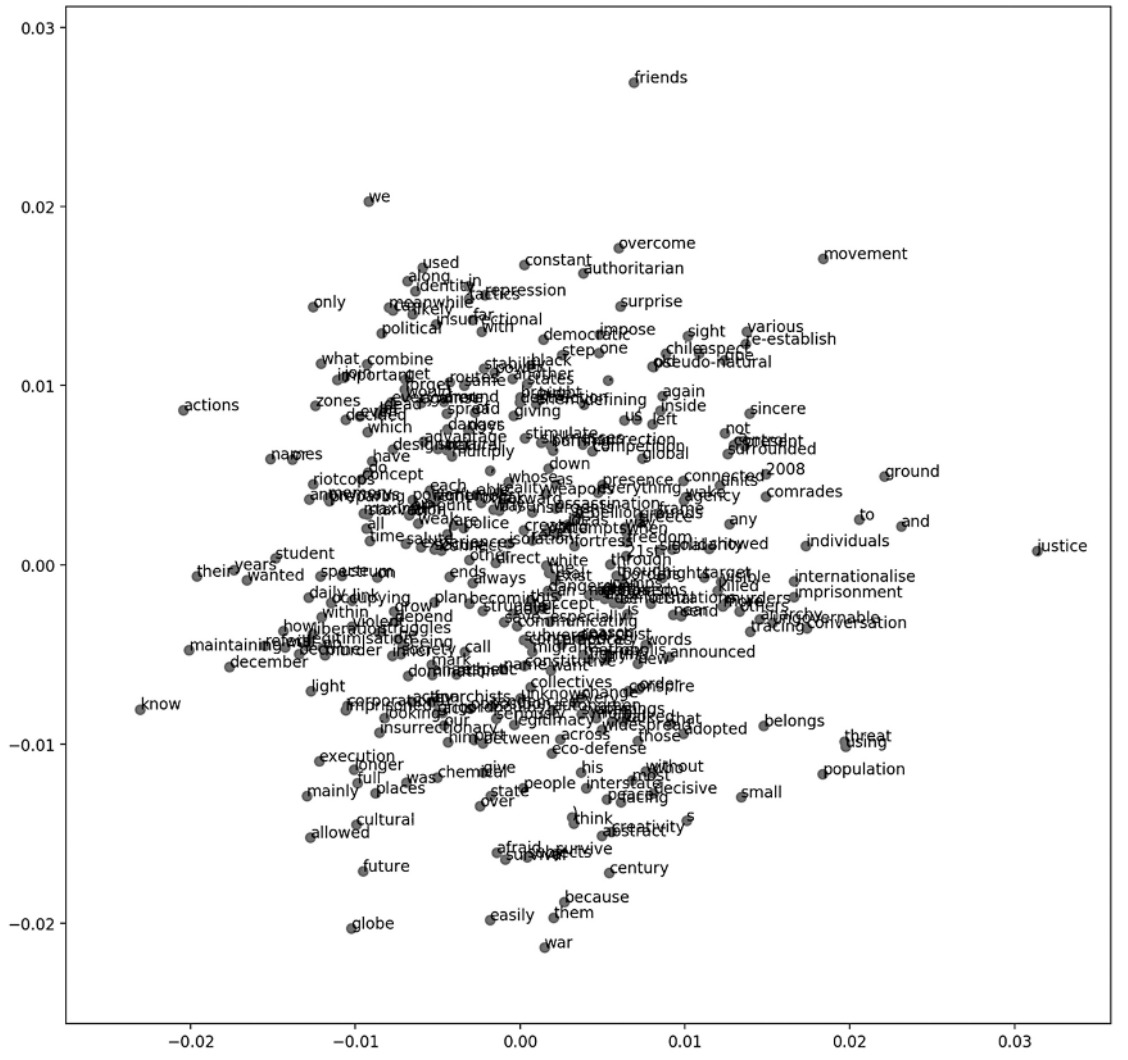
Verfahren der qualitativen Informationsextraktion bieten heute durch die beispiellose Menge an leicht zugänglichen Daten zu sozialen, politischen und wirtschaftlichen Prozessen bahnbrechende Potenziale. In den Computational Sciences, insbesondere den Computational Social Science (CSS), bleiben sie jedoch eine große Herausforderung.

Seit einigen Jahrzehnten wird mithilfe von solchen Verfahren versucht, die stetig wachsende Menge und Vielfalt an Daten zu analysieren und auch zu konkretisieren. Häufig sind sie implementiert in repressive und disruptive Technologien aus dem privaten Sektor, finden sich aber ebenso in staatlichen Behörden wie Geheimdienst- und Polizeidienststellen und dem Militär. Doch auch für sozialwissenschaftliche Studien und in Disziplinen wie den Biowissenschaften, der Ökonomie, der Psychologie oder den digitalen Geisteswissenschaften (digital humanities) sind sie zu einer zentralen Technik geworden.

Diese Techniken und Verfahren fusionierten ungefähr in den späten 2000er und frühen 2010er Jahren erfolgreich mit kognitiven Technologien wie beispielsweise dem Deep Learning, um Sprachmodelle zu errechnen, Organisationseinheiten und Teile sozialer Bewegungen datenförmig zu extrahieren und komputationell zu identifizieren.

Seit dieser Zeit werden vornehmlich vortrainierte Sprachmodelle zum Erstellen von Worteinbettungen (word embeddings) verwendet. Häufig stellen diese Sprachmodelle, die aus einer Vielzahl von generierten semantischen Beziehungen zwischen Wörtern bestehen und ›sinnvolle‹ Repräsentationen errechnen, künstliche neuronale Netze (KNN) dar, Deep-learning-Modelle, in denen *Lern*-Kriterien zur maschinellen Bedeutungsproduktion verankert sind. Word2vec (Mikolov et al. 2013) von Google ist eines der ersten und immer noch gebräuchlichsten Modelle dieser Art.

Bild 1: Zweidimensionale Darstellung des Word2vec-Einbettungsraumes + Similarity Task




```
#Find the top-N most similar words
word2vec_result = model.wv.most_similar('democratic', topn=3)
print("most nearest to human = ", word2vec_result)
```

```
most nearest to human =  [('insurrection', 0.2602590322494507), ('social', 0.2587948143482208), ('demonstrations', 0.23520304262638092)]
```


Solche vortrainierten Repräsentationsmodelle zum Trainieren von Worteinbettungen sind spätestens seit der Veröffentlichung des Word2vec-Algorithmus sprunghaft zu einem beliebten Tool in den Forschungsgemeinden geworden, die sich mit Natural Language Processing (NLP) befassen. Die Einbet-

tungen können einfach und bequem verwendet werden und zugleich konnte Word2vec Ergebnisse auf dem neuesten Stand der Technik (ca. 2013) liefern. Da das Sprachmodell Word2vec fester Bestandteil vieler NLP-basierter Applikationen und Forschungen geworden ist, ermöglicht eine fokussierte Untersuchung darauf umfangreiche Erkenntnisse über State-of-the-Art-Bewegungsvorhersagetools.

Das Studium aktueller NLP-Forschungsarbeiten zeigt, dass trotz der weit verbreiteten Anwendung von Worteinbettungen noch immer erstaunlich wenig über die Struktur und die Eigenschaften dieser Bedeutungsräume (embedding spaces) bekannt ist. Dennoch wird eine Unzahl an Teilen von Welt durch sie in formale, computerlinguistisch verarbeitbare Informationen und Beschreibungsebenen (Morphologie, Syntax, Semantik, Aspekte der Pragmatik etc.) umgewandelt, in Zeichen, konkreter: in Operanden für NLP-Modelle.

NLP ist eine Mischwissenschaft, die anteilig aus der Computerlinguistik, den Computerwissenschaften und der Künstliche-Intelligenz-Forschung besteht – der Wissenschaft der algorithmischen Verarbeitung von Sprache, der Wissenschaft von der Verarbeitung von Daten und der Wissenschaft vom künstlich intelligenten Verhalten.

Dass solche Forschungscluster und Bedeutungsraumgeneratoren dazu in der Lage sind, die jeweilige Handlungs- und Wirkkraft sozialer Bewegungen, deren Wege und mögliche Ziele zu beeinflussen, steht außer Frage. So kristallisiert sich zunehmend heraus, dass die komputationelle Verarbeitung natürlicher Sprache und zwischenmenschlicher Beziehungen eher nicht der Förderung politisch-freiheitlichen Handelns und zur Schaffung von freiheitlichen und gerechteren Gesellschaften tauglich ist. Im Gegenteil, sie fördert systematisch die Reproduktion herrschender Grammatiken mit erstarrten Begrifflichkeiten, Binaritäten und ihren Präzisierungen in der Sprache.

Insbesondere mit Blick auf die Sozialen Medien lassen sich gesellschaftliche Entwicklungen ablesen, die uns zeigen, dass solche Netzwerke nicht allein passive Träger von Zeichen sind, sondern auch aktive Erzeuger. Während diese Zeichen noch im Modus der Abstraktion (kurz, die Bereinigung der Phänomene von ›Unwesentlichem‹ und Zweideutigkeiten) Teile von Welt maschinenlesbar machen, laden die Interfaces Sozialer Netzwerke durch unsere technischen Handlungen die Zeichen wieder mit Unschärfen und Mehrdeutigkeiten auf (vgl. Trogemann 2010: 44).

Die Diffamierung und Diskriminierung marginalisierter Gruppen benötigt stets Ungenauigkeiten. »Gehasst wird ungenau« (Emcke 2016: 12), schreibt die Autorin und Publizistin Carolin Emcke in ihrem Buch *Gegen den Hass*. Die

durch die Interfaces Sozialer Netzwerke generierten Ungenauigkeiten jedoch sind andere als diejenigen, die vormals im Abstraktionsprozess weggeworfen wurden. »Präzise lässt sich nicht gut hassen« (ebd.), schreibt Emcke weiter:

> Mit der Präzision käme die Zartheit, das genaue Hinsehen oder Hinhören, mit der Präzision käme jene Differenzierung, die die einzelne Person mit all ihren vielfältigen, widersprüchlichen Eigenschaften und Neigungen als menschliches Wesen erkennt. (Ebd.: 11)

Erstens sind also die Präzisierungen in herrschenden Sprachen mit ihren klaren Gegensätzlichkeiten Trigger, um beispielsweise durch Rating- und Ranking-Algorithmen zur ideologischen Sichtbarkeit zu gelangen. Und zweitens fördert gerade der Umkehrprozess des Formalisierten bzw. Digitalisierten, der es Akteuren erst möglich macht, innerhalb dieser soziotechnischen Sprach- und Handlungsräume zu wirken, rassistische Sprachgebräuche, Verschwörungstheorien und Hassbotschaften, die die Möglichkeiten des Denkens und des gesellschaftlichen Handelns auf verheerende Weise reduzieren. Doch gezielte Veränderungen können unseren Denk-, Wirk- und Handlungsraum auch erweitern und partizipative Freiräume schaffen, in deren kollektiver Nutzung sich eine Gesellschaft entfalten kann. Ein Beispiel dafür ist die Kraft des Lyrischen in der politischen Sprache. Sie ist unter anderem den zahlreichen Versuchen emanzipatorischer Bewegungen eingeschrieben, durch Sprache neue kulturelle Normen zu manifestieren. Und sie ist es, die eine der größten Herausforderungen für Sprachmodelle mit neuronalen Einbettungen darstellt.

Auf neue Slangs, wie sie beispielsweise in Teilen der *Black-Lives-Matter*-Bewegung entwickelt wurden, um innerhalb der Bewegung ein nichtbinäres, antisexistisches und antirassistisches Vokabular zu etablieren, sind die Bedeutungsvektoren (word embeddings) vortrainierter Sprachmodelle nicht abgestimmt. Sie benötigen stets ein formales Gerüst, sozusagen eine Sammlung von a priori erworbenen Kenntnissen über die zu analysierende Sprache. In einem neuen Vokabular aber sind die Regeln noch nicht gesetzt und können auch nicht ohne Weiteres den künstlichen neuronalen Netzen antrainiert werden.

Die Kraft solch subversiver Sprachgebräuche und Schreibtechniken lässt sich bereits in der politischen Lyrik von Gustav Landauer bis hin zur Poetologie der frühen zapatistischen Bewegung in Mexiko finden. Seither hat sich unser Umgang mit Text jedoch drastisch verändert: Instagram, Facebook oder beispielsweise das Twitter-Interface, das uns forciert, unsere Sprache zu verdichten, Hashtags und weitere operierende Zeichen, die in unseren Textfluss

einen festen Platz einnehmen, hypertextuelles Lesen und Schreiben sowie das wöchentliche (Um-)Schreiben neuer Passphrases sind nur einige Beispiele, deren Reflexion uns einen neuen Möglichkeitsraum erschließen ließe.

Betrachten wir das konzeptionelle algorithmische bzw. auch vorprozessierte Schreiben als ein grundlegendes Element poetischer Schreibpraktiken, so baut sich diese Poesie ihre eigenen Gerüste. Mit der Formulierung einer eigenen Syntax und Semantik, durch die spielerisch Regelschritte erschaffen werden, trägt sie ein Potenzial in sich, herrschaftssprachliche Systeme in ihrer Pragmatik zu entmachten. Die hohe Dichte an Mehrdeutigkeit, die lyrischen Texten im Gemeinen innewohnt, wirkt bei diesem Ansatz auf mehreren Ebenen. Poesie lädt nicht nur unsere Alltagssprache mit Unschärfen und Zweideutigkeiten auf, sondern auch die Zeichen in der Maschine, die in ihrer Funktionalität eigentlich darauf angewiesen ist, natürlichsprachliche Texte zu bereinigen, so, wie es auch der Gattung des Sachtextes und mitunter der Prosa eingeschrieben steht – auch außerhalb maschineller Environments, zum Beispiel in der politischen Rede.

Die literarischen Strömungen und polit-poetischen Bewegungen des letzten Jahrhunderts waren in diesem Sinne Vorreiter in der Entwicklung alternativer Sprachformen, um erstarrte herrschaftssprachliche Begrifflichkeiten in Bewegung zu bringen. Zusammen mit der sozialen Dynamik des Hacktivismus können heute handhabbare Taktiken entwickelt werden, die sozialen Bewegungen einen Möglichkeitsraum erschließen, mit zeitgenössischen Herrschaftsinstrumenten subversiv zu spielen.

Die folgenden Experimente nehmen diese zwei Dynamiken auf, um einen Stimmungsklassifikator zu unterwandern. Die Stimmungsanalyse zählt zu einem Teilbereich des Text-Mining bzw. des Opinion-Mining und wurde zu einem beliebten Verfahren in sicherheitstechnologischen Tools zur Vorhersage sozialer Bewegungen. Sie bezieht sich auf die automatische Auswertung von Texten mit dem Ziel, eine ausgedrückte Haltung als positiv oder negativ zu identifizieren.

Beide Experimente stellen sogenannte Paraphrasierungsattacken dar. Die Unterschiede liegen lediglich in den jeweiligen Methodiken: Die eine ist regelbasiert und verortet sich in der experimentellen Literatur, die andere lässt sich dem Hacktivism (Adversarial Hacking) zuordnen und beruht auf Machine Learning.

Stein on NLP

Es gibt wohl keine schönere essayistische Annäherung an diese beiden Gegensätze in der westlichen Literaturgeschichte, keine leidenschaftlichere literarische Beschreibung des Schreibens und der Technologie des Natural Language Processings (NLP) als die von Gertrude Stein in *Poetik und Grammatik*. Und dabei schreibt sie in ihren Essays in keinem Moment über diese ganz spezifische Technologie, geschweige denn über Techniken der computergestützten natürlichen Sprachverarbeitung, die Computerlinguistik oder die maschinelle Verarbeitung natürlicher Sprachen. Diese sollten sich erst viele Jahrzehnte später in Rezeption und Anwendung gesellschaftlich etablieren. Was Gertrude Stein uns aber zeigt, ist, dass Technologie eine Art des Denkens und des Handelns ist – und es gibt wenige unter uns Code-Literat*innen, die sich von Stein distanziert oder sie nicht zumindest als Inspirationsquelle und Lehrerin erwähnt haben.

Gertrude Stein wurde von dem US-amerikanischen Schriftsteller Thornton Wilder gern als »Mutter der Moderne« bezeichnet, als die sie mitunter heutzutage noch gilt. Sie studierte 1893–1897 die damals aufstrebenden Disziplinen Psychologie und Gehirnwissenschaften am Radcliffe College u.a. beim Begründer des Pragmatismus, dem Philosophen und Psychologen William James. Danach wechselte sie an die Johns Hopkins Medical School (1897–1902), brach ihr Studium jedoch kurz vor Abschluss ab und widmete sich der Literatur. Viele Erkenntnisse und Modelle aus den frühen Neurowissenschaften und der Hirnforschung flossen in ihre späteren literarischen Experimentalansätze und poetischen Sprachtechniken mit ein. Auf diese Weise schuf Stein Anfang des 20. Jahrhunderts gänzlich neue Zugänge zu Sprache und zu den Dingen. Die modernistische Poesie, aber auch die moderne Neurowissenschaft entdeckten Anfang des 20. Jahrhunderts kognitive Systeme und Räume, aus denen die Syntax unserer heutigen formalen Techniksprachen (insbesondere für KNNs) Hand in Hand mit den frühen poetischen Sprachtechniken und Experimenten schritten.

Stein und weitere *experimentelle* Schriftsteller wie beispielsweise William Carlos Williams, der während seiner Zeit als Schriftsteller auch praktizierender Mediziner war, verwendeten diesen *neuen* Raum, um Fragmente zu sinnvollen Arrangements zusammenzusetzen, die die veralteten Systeme des 19. Jahrhunderts ersetzen sollten (vgl. Ambrosio 2018).

Sie schrieb in *Poetik und Grammatik* vom Schreiben als poietischem Handeln. Auf ihrem Weg von der Prosa hin zur Poesie nutzte sie beim Schreiben

formalisierte Einschränkungen und selbst auferlegte Zwänge, um Sprache zu verarbeiten, zu filtern und in Poesie zu verwandeln:

> »Wenn man in der Schule ist und Grammatik lernt, ist Grammatik sehr aufregend. Ich weiß wirklich nicht daß irgendetwas je aufregender gewesen wäre als Diagramme von Sätzen aufzustellen.« (Stein 1965: 159)

Und gerade weil Poesie so viel mit Freude und Leidenschaft zu tun hat, mit Bewegung und damit, wie man die Wörter innerlich fühlt, aber genauso viel auch mit unserer Kulturtechnik des Zählens, mit Wissensrepräsentationen, mit Regelsystemen unserer Sprache und wie man Diagramme von Sätzen aufstellt, vor allem jedoch mit der Art der Erzählung von eigens gemachten Schreiberfahrungen, gerade deshalb spielt Stein eine solch zentrale Rolle in unserer Historizität der semiotischen Maschinen.

»A cool red rose and a pink cut pink, a collapse and a sold hole, a little less hot.« (Gertrude Stein)

In ihren Kombinatoriken versuchte Stein, visuelle Merkmale »im Rahmen einzelner Satzstrukturen« abzubilden bzw. »objektbezogene Rhythmen« (Kirchner 2001: 12) zu schaffen, um dann durch diese rhythmisch-syntaktischen Operationen auf die semantische Ebene zu gelangen.

Heutige Rechenverfahren der Poesieanalyse haben häufig syntaktische Regeln implementiert, um freie poetische Vokabularien in eine strenge Form zu pressen. Das zeigt sich zum Beispiel in der computerlinguistischen Forschungsarbeit *Word Reordering Algorithm for Poetry Analysis* von Barakhnin und Pastushkov, in der zuerst versucht wird, Gedichte mithilfe von Chunks und Syntaxgruppen in Syntaxkonstruktionen von Prosatexten zu überführen, um daraufhin Semantiken zu generieren (vgl. Barakhnin/Pastushkov 2019). Das Feld der komputationellen Poesieanalyse bleibt experimentell. Denn sprachliche Ausdrücke, bei denen ein Wort aus seinem eigentlichen Bedeutungszusammenhang in einen anderen übertragen wurde, ohne dass dies durch direkte Vergleiche von Beziehungen zwischen Bezeichnendem und Bezeichnetem verdeutlicht würde, sind semantisch nicht gerade leicht zu analysieren – weder für Menschen noch für Maschinen.

Poesie ist demnach in sich Bewegung. Somit ist sie selbst auch weitaus näher am zählenden Algorithmus – dem performativen Element des Codes – als an den Zahlen selbst; näher am Operator als an den Operanden. Sprechen wir heutzutage von einer Kulturtechnik des Zählens, sprechen wir *von etwas*, was

im heutigen Rechnungswesen *für etwas* steht. Früher hießen diese Etwasse Ersatzmengen. Sie waren die Vorläufer des Computers. Etwas anderes können Computer nicht, auch wenn sie etwas anderes sein können.

Es ist im Kontext dieses Beitrags wichtig, nicht aus den Augen zu verlieren, dass das Zählen eine Erfindung ist – ebenso, wie Computer Erfindungen sind.

Wer Computerprogramme schreibt, muss die abstrakten Zwischenebenen, die zwischen den Dingen und unserem Denken liegen, allesamt mitdenken können, und wer Computerprogramme liest, liest diese auch mit, selbst wenn man sie gar nicht sieht. Programmierer*innen müssen stetig zwischen dem, was war – ihr Tun historisch einordnend –, und den jeweiligen Konsequenzen, die ihr Tun mit sich bringen wird, abwägen. Das gilt auch und vor allem für die sogenannten *Nebenprodukte* ihrer jeweiligen technischen Handlungen. Sie müssen dem Code, ihren geschriebenen Zeichen einen Wert zuweisen.

Wer im Volksmund sagt, *dieses oder jenes Zeichen, das steht für etwas*, meint, dass dieses Etwas dem hierfür stehenden Zeichen vorstünde. Doch wie ist es, wenn etwas Geschriebenes, wie es Gertrude Stein beschreibt, wenn ein hervorgebrachtes Werk beginnt, ein Eigenleben zu führen und eigenständig zu handeln? Wenn wir an die Technologien, die wir schreiben, plötzlich mehr und mehr Handlungsmacht delegieren? Ziehen wir zum Beispiel ein Interpunktionszeichen heran: Etwa den Punkt, wenn er beginnt, auf seine eigene »Weise […] zu existieren« (Stein 1965: 165). Ein Punkt hilft uns, »wieder und wieder manchmal anzuhalten« (ebd.), ganz einfach weil man ab und an physisch anhalten muß: *eins zwei drei vier*. Was ändert sich, wenn künstliche neuronale Netze Entscheidungen für uns treffen, die unsere Lebenswelt gestalten? Die anhalten, wenn sie anhalten und nicht, wenn die Programmierer*innen oder die Leser*innen gewillt sind oder einfach aus physischer Notwendigkeit heraus aus dem Lesefluss aussteigen möchten?

Es scheint, als wäre sich Stein beim Dichten dem bewusst gewesen, was die politische Theoretikerin Hannah Arendt ein paar Jahrzehnte später in *Vita activa* schrieb: »daß der Mensch sich an diesen Rhythmus der Maschinen gewissermaßen schon gewöhnt haben mußte, als er ein solches Ding wie eine Maschine auch nur im Geist konzipierte« (Arendt 1981: 136). Denn solange man mit und durch diese Maschinen schreibt, solange treten auch ihre mechanischen Prozesse und ihre diskreten Zeiteinheiten an die Stelle unseres eigenen Körperrhythmus.

Es muss also unbedingt eine Sprache geschrieben werden können, die »die ganze Geschichte ihrer geistigen Wiedererschaffung« (ebd.: 183) in sich trägt.

Anders gelingt es nicht, für einen kurzen Moment heraustreten zu können aus dem Schreib- und Lesefluss und zu beginnen, darüber nachzudenken und zu reflektieren.

»Wer Dichtung will, muss auch die Schreibmaschine wollen.« (Arno Schmidt)

Die Schreibmaschine ersetzte den Literat*innen nur selten ihre Handschrift – ebenso wenig wie Gutenbergs Druckpresse zuvor und der Computer danach. Aber diese Technologien »deplatzieren sie und führen sie hin zum Schreiben, zum Erfinden und zum Denken über andere Dinge« (Dick 2013).

Das sogenannte Gutenberg-Zeitalter ist aus medientheoretischer Sicht eine Periode der *Explosion*. Sie sollte von zahlreichen Kulturkämpfen und Widerstandsbewegungen gegen Normierungen und Standardisierungen der Sprache geprägt sein und Jahrhunderte dauern – bis ins *elektronische Zeitalter* hinein, ins sogenannte *Zeitalter der Implosion* (vgl. McLuhan 1962). Denn die gesellschaftliche Wirksamkeit dieses *Schlüsselelements der Renaissance* vollzog sich nicht durch die plötzliche breitengesellschaftliche Anwendung einer Drucktechnik mit beweglichen Lettern. Erst 200 Jahre später konnte ungefähr die Hälfte aller Europäer lesen. »Der Buchdruck forcierte (auf diese Weise) die Entwicklung einer Zusatztechnologie, nämlich des Lesenkönnens.« (Luhmann 1997: 729) Die Technologie des Buchdrucks bewirkte weder demokratische Prozesse noch befreite sie die Gesellschaft von kirchlichen Repressionen und deren gesellschaftlicher Wirkmächtigkeit.

In der Moderne angekommen, zeichnen sich Gesellschaften dann dadurch aus, »dass sie überall reproduzierbare schriftliche Spuren hinterlassen«, die »wiederum die Voraussetzung für weitere Operationen« (Nassehi 2019: 136) sind.

Im dritten Jahrtausend nun sind es Datenspuren, die wir fortwährend hinterlassen und die in kleinen zielgerichteten bis hin zu weltumspannenden Datenbanken effizient, widerspruchsfrei und bestenfalls dauerhaft gespeichert werden. In Sicherheitstechnologien implementierte NLP-Modelle ordnen Zeichen kontextspezifisch so an, dass dadurch soziale Prozesse und Bewegungen nicht nur analysiert, sondern auch vorhergesehen und verändert werden können (siehe weiter unten).

Steinese Tokenization

Es war Jack Kerouac, der in der Mitte des 20. Jahrhundert den Begriff der Beat Generation in die New Yorker Literaturszene einführte. Diese literarische Avantgardebewegung benannte sich nicht nur in Analogie zur Lost Generation (F. Scott Fitzgerald, Ernest Hemingway, Gertrude Stein u.v.m.), sondern auch ihr Umgang mit dem Schreiben steht in direkter ästhetischer Linie mit dieser Vorgängergeneration: Alsbald wurden sie als diejenigen bezeichnet, *die im Rhythmus schreiben*.

William S. Burroughs, der wohl ambivalenteste unter ihnen, stellte die Sprache oft als etwas dem Menschen Fremdes dar. In seinem Essay *The Electronic Revolution* (Burroughs 1970), der Gilles Deleuze (1990) zu seiner Vorstellung der ›Kontrollgesellschaft‹ verhalf (vgl. Assis 2018: 191), stellte Burroughs unter anderem Überlegungen darüber an, wie die vorherrschende Gesellschaftsform *grammatikalisch* zerwürfelt (to scramble) werden könnte. Die *Kontrollsyntax* seiner Zeit drang laut Burroughs in das gesellschaftliche Subjekt ein und bestimmte so sein ganzes Denken und Handeln. Er schrieb ihr entgegen, unter anderem in *Rub Out the Word* (Burroughs/Gysin 1978), in dem er zusammen mit Brion Gysin die folgenden drei Schritte herausarbeitete:

Löschen Sie **1.** die Kopula (*sein/bleiben*), diese Satzbänder mit zwar überaus wichtiger grammatischer Funktion für die Identitätsbildung, doch mit eher schwach ausgeprägter Bedeutung. **2.** Ersetzen Sie bestimmte Artikel (*der*) vor Substantiven durch unbestimmte Artikel (*ein*), das heißt, vermeiden Sie eine Verdinglichung. Und ersetzen Sie **3.** *entweder/oder* mit *und*, was so viel heißt, wie das Gesetz des Widerspruchs zu ignorieren (vgl. ebd).

Da die herrschende Sprache die Sprache der herrschenden Klasse ist, musste laut Burroughs und Gysin der herrschende legitime Sprachgebrauch durch den Einsatz alternativer Sprachtechniken untergraben werden. Wie Gertrude Stein haben sie sich hierfür die Sprachtechnologien ihrer Zeit poietisch angeeignet und erforscht.

Um »innerlich die Wörter zu fühlen die herauskommen um außerhalb von einem zu sein« (Stein 1965: 158), ging Gertrude Stein beim Diagrammieren von Sätzen wie folgt vor: Sie begann mit einer Aufzählung all derjenigen Wörter, die in ihren Augen etwas tun, denn »so lange irgend etwas etwas tut, bleibt es lebendig« (ebd.: 162). Hierfür tat sie selbst etwas, was dem POS-Tagging, einem Verfahren aus der Computerlinguistik, sehr nahe kommt:

(›they‹, ›PRP‹),
(›see‹, ›VBP‹),
(›that‹, ›IN‹),
(›darker‹, ›NN‹),
(›makes‹, ›VBZ‹),
(›it‹, ›PRP‹),
(›be‹, ›VB‹),
(›a‹, ›DT‹),
(›color‹, ›NN‹),
(›white‹, ›JJ‹),
(›for‹, ›IN‹),
(›me‹, ›PRP‹).

Stein erstellte zuerst eine grammatikalische Klassifikation, um auf deren Basis ihre ganz eigenen Regeln aufzuschreiben. Sie begann in *Poetik und Grammatik* über *Präpositionen* zu schreiben und bemerkte, dass diese sich meist irren. Dann kam sie zu *Artikeln*: empfindliche und vielfältige Etwasse, deren Gebrauch ein lebendiges Vergnügen sein kann. »Sie sind genauso interessant, wie Substantive und Adjektive es nicht sind.« (Ebd.: 161) Die *Adjektive* waren für sie uninteressant, und zwar aus dem einfachen Grund, dass sie »das erste Ding [sind] das irgendeiner aus dem geschriebenen von irgendeinem herausnimmt« (ebd.). Und warum sind *Substantive* uninteressant? Die Suche nach genau dieser Antwort würde ihre lange Reise von der Prosa zur Poesie sein. Dann kam sie zu den *Verben* und zu den *Adverbien*, die unentwegt Fehler machen können und dürfen. Sie können »sich verändern um auszusehen wie sie selbst […] sie sind sozusagen in Bewegung« (ebd.: 162). Und *Adverbien* bewegen sich mit ihnen. *Pronomina* bewegen sich im Gegensatz zu *Namen* in einem weitaus größeren Möglichkeitsraum, weil sie erstens nicht von Adjektiven begleitet werden können und weil sie zweitens eben nicht wirklich der Name von irgendetwas sind.

Schließlich kommt Stein zu den *Interpunktionen*; zur Groß- und Kleinschreibung, mit der es Spaß macht, zu spielen, und zu »uninteressanten Fragezeichen« und zu Ausrufezeichen und Anführungszeichen, die sie als »unnötig« und »häßlich« empfindet, weil sie »das Bild des geschriebenen verderben« (ebd.: 163); zu Zwischenräumen wie Gedankenstrichen und Pünktchen und zu besitzanzeigenden Apostrophen, die so manchem Genitiv »einen feinen zarten Unterton« (ebd.: 164) verleihen. Die Benutzung von Kommata

ist für Stein ohne weiteren Nutzen, denn sie ersetzen einem auf ihre Art das eigene Interesse, sie machen einem die Sache leichter, aber nicht einfacher, sie sind »arme Punkte« (ebd.). Kommata sind servil – ganz im Gegensatz zu den Punkten.

Wie bereits zuvor im Text erwähnt, beschreibt Gertrude Stein nirgends schöner, wie es ist, wenn ein Geschriebenes, wenn ein hervorgebrachtes Werk beginnt, ein Eigenleben zu führen, als mit dem *Punkt*. Wenn er beginnt, eigenständig zu handeln. Wenn er einen dazu anhält, »wieder und wieder manchmal anzuhalten« (ebd.: 165), ganz einfach weil man ab und an physisch anhalten muss. Der Punkt konnte auf seine eigene »Weise dazu kommen zu existieren« (ebd.). Punkte haben ihre »eigene Notwendigkeit ein eigenes Gefühl eine eigene Zeit. Und jenes Gefühl jenes Leben jene Notwendigkeit jene Zeit kann sich selbst ausdrücken in einer unendlichen Mannigfaltigkeit.« (Ebd.: 166)

Abschließend kommt Gertrude Stein dann zu den *Sätzen*, die sich eben in der Benutzung des vorherigen Satzes bilden, und zu *Absätzen*. Ihr Versuch, in einem kurzen Satz das nicht emotionale Gleichgewicht eines Satzes mit dem emotionalen Gleichgewicht der Absätze in Einklang zu bringen, ist ein Meilenstein auf ihrem Weg über die Prosa hinaus hin zur Poesie. Ein neues Gleichgewicht zu schaffen, »das zu tun hatte mit einem Gefühl von Bewegung von Zeit eingeschlossen in einen gegebenen Raum« (ebd.: 171).

* * *

do nOt FOrGET ANaRCHist – January 3, 2023

1 Methode

1) Gertrude Steins *Poetik und Grammatik* lesen
2) Den Text durch eine Festlegung einfacher grammatikalischer Regeln formalisieren
3) Einlesen der Message »*No Justice No Peace*« von anonymous Anarchist agency als Variable salute_spacy
4) Zuordnen der jeweiligen Wörter und Satzzeichen in der Message zu Wortarten (POS-Tagging)
5) Beginnen die Message Schritt für Schritt nach den oben auferlegten Regeln mit Python & SpaCy umzuschreiben:
    1) **Präpositionen** irren sich meist
    2) **Adjektive** sind uninteressant
    3) **Nomen** ebenso
    4) **Verben** sind in Bewegung
    5) & **Adverbien** bewegen sich mit ihnen
    6) **Pronomina** bewegen sich in einem sehr großen Möglichkeitsraum
    7) **Namen** bewegen sich nicht
    8) mit **Groß- und Kleinschreibung** macht es ungemeinen Spaß zu spielen
    9) **Fragezeichen** sind uninteressant
    10) **Ausrufezeichen und Anführungszeichen** sind unnötig und hässlich
    11) **Kommas** sind überflüssig
    12) der **Punkt** führt den Text zu seinem Eigenleben
6) Nach jedem Schritt wird eine Stimmungsanalyse (mit TextBlob) durchgeführt

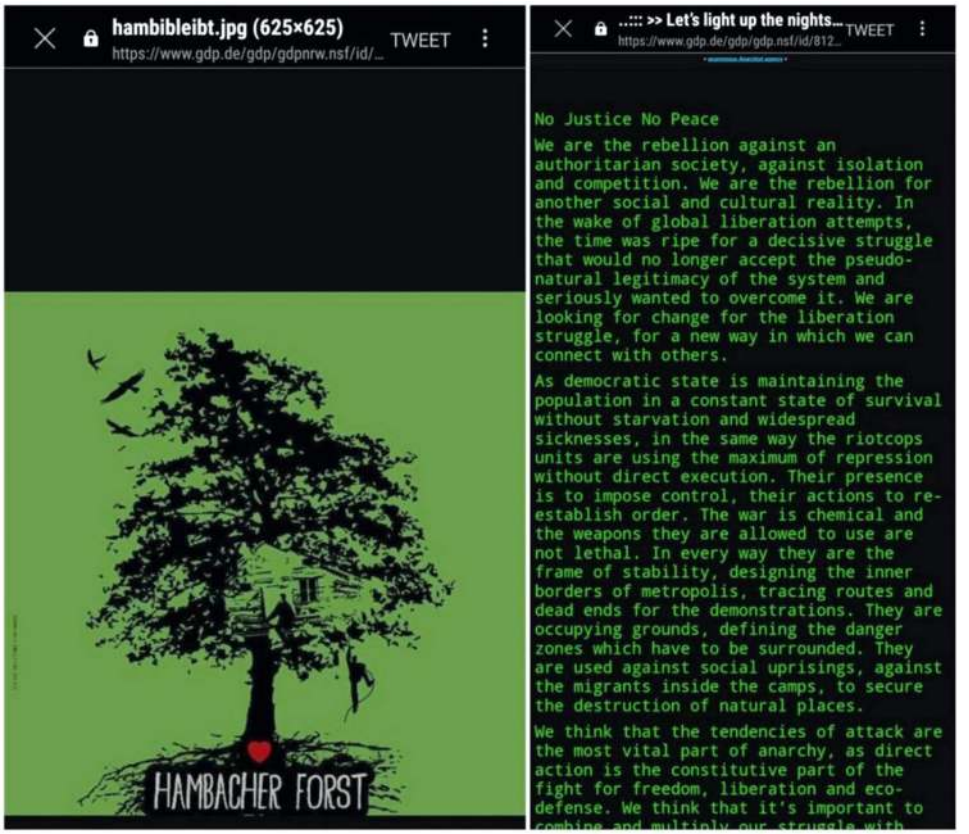
2 Ziel

Versuchen, die Maschine literarisch so auszutricksen, dass die subjektive Information in einer Äußerung genau umgekehrt klassifiziert wird. Kurz gesagt: Meinungen, Gefühle oder Einstellungen zu einem Thema oder einer Person etc., die normalerweise von Stimmungsklassifikatoren als negativ oder schlecht interpretiert werden, werden als positiv »fehlinterpretiert«.

3 Der zu manipulierende Text

Das anarchistische Kommuniqué **»No Justice No Peace«** von **anonymous Anarchist agency**.

Bild 2: Screenshot des GdP-Hacks[1]



1 Ausführlichere Informationen zu den Hintergründen dieser Aktion sind zu finden unter: hambi blog: https://hambacherforst.org/blog/2019/12/03/13669/; the anarchist-news: https://anarchistnews.org/content/german-police-union-gdp-website-hacked-%E2%80%9Canonymous-anarchist-agency%E2%80%9D-aaa. Zugegriffen: 27. Februar 2022.

Am 03.12.2019 hackte diese Gruppe die Webseite der deutschen Polizeigewerkschaft. Für einige Stunden war auf ihr das Banner der Hambacher-Forst-Besetzung zu sehen und die nebenstehende Messsage zu lessen (vgl. Bild 2).

4 Umsetzung

```
# import necessary librairies
import spacy
from textblob import TextBlob
from textblob.sentiments import NaiveBayesAnalyzer
import nltk
import re
import random

# load the communiqué from file & create spacy object from it
file = open('salute-anonymous', encoding='utf-8')
salute_string = file.read()
file.close()
spacy_obj = spacy.load('en_core_web_sm')
salute_spacy = spacy_obj(salute_string)

# operate sentiment classification on text & print original text
blob = TextBlob(salute_string, analyzer=NaiveBayesAnalyzer())
print("Print sentiment classification on original message:\n", blob.sentiment)
print("\nPrint original message:\n", salute_string)
```

Print sentiment classification on original message :

Sentiment(classification='pos', p_pos=1.0, p_neg=4.4526365268591785e-27)

Print original message:

No Justice No Peace

We are the rebellion against an authoritarian society, against isolation and competition. We are the rebellion for another social and cultural reality.

In the wake of global liberation attempts, the time was ripe for a decisive struggle that would no longer accept the pseudo-natural legitimacy of the system and seriously wanted to overcome it. We are looking for change for the liberation struggle, for a new way in which we can connect with others.

As democratic state is maintaining the population in a constant state of survival without starvation and widespread sicknesses, in the same way the riotcops units are using the maximum of repression without direct execution. Their presence is to impose control, their actions to re-establish order. The war is chemical and the weapons they are allowed to use are not lethal. In every way they are the frame of stability, designing the inner borders of metropolis, tracing routes and dead ends for the demonstrations. They are occupying grounds, defining the danger zones which have to be surrounded. They are used against social uprisings, against the migrants inside the camps, to secure the destruction of natural places.

We think that the tendencies of attack are the most vital part of anarchy, as direct action is the constitutive part of the fight for freedom, liberation and eco-defense. We think that it's important to combine and multiply our struggle with comrades around the globe, because everywhere that anarchists are fighting, we can become a more dangerous and subversive threat against power when we internationalise and break down the borders between us. This is the reason that the States and corporations across the world are becoming afraid of anarchy again in the 21st Century, because we are the only real opposition to power's domination and they do not want us to grow and link up with each other to conspire.

The insurrection of 2008 in Greece was one of the most powerful in the contemporary world. It showed an the amount of power and creativity which can be brought against the state mechanisms and how weak and small they looked those days. The murder of the anarchist Grigoropoulos from the Greek police will always be present in our memory and through our actions. Even though, we are to mark that this assassination, and the insurrection it brought within the society and the all left political spectrum was mainly because he was a 15 years old white Greek student (meanwhile they try to hide his political identity). From our sight, we are seeing the state's murders in various subjects, in a daily base along with the imprisonment and domination over the ground. That is why our

> struggles are violent and constant. Our insurrectionary acts do not depend on social legitimisation. Society is an abstract concept, more likely it refers to what is visible and approved to exist. Our struggles are connected to our experiences. We do not fight to save the people, we fight to survive and give solidarity to those who resist with the target to stimulate more individuals and collectives to join this sincere anarchist struggle.

Our tactics can easily be adopted in any metropolis of the InterState fortress. Every preparing action, every conversation, every aspect of our plan, can get us one step forward from our enemy giving us the advantage to surprise him. Let's create an insurrectional movement without borders that will be able to spread anarchistic ideas and practices.

> We send a burning signal and we join the call for a Black December announced by the comrades in Chile by communicating with this action and our words with the insurgents around the world. In fighting memory of all friends, comrades and unknown killed or imprisoned by the state.
>
> We will never forget the comrades, we want to especially remember those who decided to give everything in the struggle and died in it. Our memory and our full respect for those whose names we can not name because we do not know them.
>
> The future belongs to those who struggle for liberation. Solidarity with people facing repression near and far! Let's light up the nights and days We are Autonomen, We are Ungovernable, We are Action, We do not forget our comrades,
>
> Salute,
> anonymous Anarchist agency

Präpositionen irren sich meist

```
# operate POS-Tagging & rule 5.1
prepositions = []
for token in salute_spacy:
if token.tag_ == 'PRP':
```

```
prepositions.append(token.text)
big_regex = re.compile(r'\b%s\b'% r'\b|\b'.join(map(re.escape, prepositions)))
the_message = big_regex.sub(".", salute_string)
salute_spacy = spacy_obj(the_message)

# operate sentiment classification on manipulated text & print it
blob = TextBlob(the_message, analyzer=NaiveBayesAnalyzer())
print("Print sentiment classification:\n", blob.sentiment)
print("\nPrint first 50 tokens of next step version:\n", salute_spacy[0:50])
```

Print sentiment classification :

Sentiment(classification='pos', p_pos=1.0, p_neg=4.986502948777287e-27)

Print first 50 tokens of next step version:

No Justice No Peace

. are the rebellion against an authoritarian society, against isolation and competition. . are the rebellion for another social and cultural reality. In the wake of global liberation attempts, the time was ripe for a decisive struggle that would no longer

Adjektive sind uninteressant

```
# operate POS-Tagging & rule 5.2
adjectives = []
for token in salute_spacy:
if token.pos_ == 'ADJ':
adjectives.append(token.text)
big_regex = re.compile(r'\b%s\b'% r'\b|\b'.join(map(re.escape, adjectives)))
the_message = big_regex.sub(".", the_message)
salute_spacy = spacy_obj(the_message)
```

```
# operate sentiment classification on manipulated text & print it
blob = TextBlob(the_message, analyzer=NaiveBayesAnalyzer())
print("Print sentiment classification:\n", blob.sentiment)
print("\nPrint first 50 tokens of next step version:\n", salute_spacy[0:50])
```

Print sentiment classification :

Sentiment(classification='pos', p_pos=1.0, p_neg=2.759697554176465e-17)

Print first 50 tokens of next step version:

No Justice No Peace

. are the rebellion against an . society, against isolation and competition. . are the rebellion for another . and . reality. In the wake of . liberation attempts, the time was . for a . struggle that would no longer

Nomen ebenso

```
# operate POS-Tagging & rule 5.3
nouns = []
for token in salute_spacy:
if token.pos_ == 'NOUN':
nouns.append(token.text)
big_regex = re.compile(r'\b%s\b'% r'\b|\b'.join(map(re.escape, nouns)))
the_message = big_regex.sub("", the_message)
salute_spacy = spacy_obj(the_message)

# operate sentiment classification on manipulated text & print it
blob = TextBlob(the_message, analyzer=NaiveBayesAnalyzer())
print("Print sentiment classification:\n", blob.sentiment)
print("\nPrint first 50 tokens of next step version:\n", salute_spacy[0:50])
```

Print sentiment classification :

Sentiment(classification='pos', p_pos=0.9996714411157043, p_neg=0.0003285588843093348)

Print first 50 tokens of next step version:

No No

. are the against an . , against and . . are the for another . and . . In the of . , the was . for a . that would no longer accept the ... of

Verben sind in Bewegung

```
# operate POS-Tagging
verbs = []
for token in salute_spacy:
if token.pos_ == 'VERB':
    verbs.append(token.text)
```

& Adverbien bewegen sich mit ihnen

```
# operate POS-Tagging & rule 5.4 + 5.5
adverbs = []
for token in salute_spacy:
if token.pos_ == 'ADV':
adverbs.append(token.text)
salute_list=the_message.split(' ')

for i in range(0, random.randint(0, 500)):
dotx = random.randint(0, len(salute_list))
hoho=random.choice(adverbs) + " " + random.choice(adverbs) + " " + random.choice(verbs)
hoho=hoho.split(' ')
salute_list = salute_list[:dotx] + hoho + salute_list[dotx:]
the_message = ' '.join(salute_list)
```

```
salute_spacy = spacy_obj(the_message)

# operate sentiment classification on manipulated text & print it
blob = TextBlob(the_message, analyzer=NaiveBayesAnalyzer())
print("Print sentiment classification:\n", blob.sentiment)
print("\nPrint first 50 tokens of next step version:\n", salute_spacy[0:50])
```

Print sentiment classification :

Sentiment(classification='pos', p_pos=0.9996714411157043, p_neg=0.0003285588843092 96)

Print first 50 tokens of next step version:

No As easily stimulate No

. are the far again used As forward seriously mainly send remember As though never meanwhile remember have Even forward give against an . , against and . longer Even Let . are far forward grow always though forget the

Pronomina bewegen sich in einem sehr großen Möglichkeitsraum

```
# operate POS-Tagging & rule 5.6
pronouns = []
for token in salute_spacy:
if token.pos_ == 'PRON':
pre=pronouns.append(token.text)
big_regex = re.compile(r'\b%s\b'% r'\b|\b'.join(map(re.escape, pronouns)))
the_message = big_regex.sub(".", the_message)
salute_spacy = spacy_obj(the_message)

# operate sentiment classification on manipulated text & print it
blob = TextBlob(the_message, analyzer=NaiveBayesAnalyzer())
print("Print sentiment classification:\n", blob.sentiment)
print("\nPrint first 50 tokens of next step version:\n", salute_spacy[0:50])
```

Print sentiment classification :

Sentiment(classification='pos', p_pos=0.9993378976272029, p_neg=0.0006621023728111459)

Print first 50 tokens of next step version:

No As easily stimulate No

. are . far again used As forward seriously mainly send remember As though never meanwhile remember have Even forward give against an . , against and . longer Even Let . are far forward grow always though forget .

Namen bewegen sich nicht

```
# operate Named Entity Recognition & rule 5.7
for i in reversed(salute_spacy.ents):
start = i.start_char
end = start + len(i.text)
  salspac = the_message[:start] + '.' + the_message[end:]

# operate sentiment classification on manipulated text & print it
blob = TextBlob(salspac, analyzer=NaiveBayesAnalyzer())
print("Print sentiment classification:\n", blob.sentiment)
the_message = salspac
print("\nPrint first 50 tokens of next step version:\n", the_message[0:50])
```

Print sentiment classification :

Sentiment(classification='pos', p_pos=0.9993378976272029, p_neg=0.0006621023728111459)

Print first 50 tokens of next step version:

No As easily stimulate No
. are . far again use

mit Groß- und Kleinschreibung macht es ungemeinen Spaß zu spielen

```
# operate rule 5.8
salutebreak=the_message.replace('\n', ' ')
the_message=''.join(random.choice((str.upper,str.lower))(x) for x in salute-
break)
salute_spacy = spacy_obj(the_message)

# operate sentiment classification on manipulated text & print it
blob = TextBlob(the_message, analyzer=NaiveBayesAnalyzer())
print("Print sentiment classification:\n", blob.sentiment)
print("\nPrint first 50 tokens of next step version:\n", salute_spacy[0:50])
```

Print sentiment classification :

Sentiment(classification='pos', p_pos=0.9993378976272029, p_neg=0.0006621023728111459)

Print first 50 tokens of next step version:

no As eaSiLy StImulate no . ARE . FaR agAin USed AS foRwArD SErlOUsLy MaINLY seND rememBEr AS tHOugH nEvER mEANwHilE REMEmbeR HAvE eVen FoRwaRD GIVE AGAInst aN . , aGaInst And . lOngER EVEn Let . ARe FAr foRwARD GroW alWaYs ThOUgH FoRgET .

Fragezeichen sind uninteressant

Ausrufezeichen und Anführungszeichen sind unnötig und hässlich

Kommas sind überflüssig

```
# operate rule 5.9, 5.10, 5.11
sal=the_message.translate({ord(ch):'' for ch in '!?;:“”,*"#.'}).replace('\n', ' ')
salute_list=sal.split(' ')
```

der Punkt führt den Text zu seinem Eigenleben

```
# operate rule 5.12
dot = ['.']
for i in range(0, random.randint(0, 10000)):
dotx = random.randint(0, len(salute_list))
salute_list = salute_list[:dotx] + dot + salute_list[dotx:]
salute_string2 = ' '.join(salute_list)
the_message=re.sub(r'\s([.](?:\s|$))', r'\1', salute_string2)
salute_spacy = spacy_obj(the_message)

# operate sentiment classification on manipulated text & print it
blob = TextBlob(the_message, analyzer=NaiveBayesAnalyzer())
print("Print final sentiment classification:\n", blob.sentiment)
print("\nPrint final version:\n", salute_spacy)
```

Print final sentiment classification :

Sentiment(classification='pos', p_pos=0.9993378976272029, p_neg=0.0006621023728111459)

Print final version:

no. As eaSiLy StImulate. .. .. no. .. . . ARE. .. .. FaR agAin. . USed AS foRwArD SErIOUsLy MaINLY seND. rememBEr AS. .. . tHOugH. nEvER. .. .. mEANwHilE. .. .. REMEmbeR. .. HAvE eVen. FoRwaRD. GIVE. .. .. .. AGAInst aN. .. . . aGaInst . And . . lOngER. . EVEn Let . ARe. .. FAr. . foRwARD. GroW. alWaYs ThOUgH FoRgET. . . . fOR. . . .. . AnD. .. .. . IN. . . . oF. . . . . Was. .. . .. FoR . .. . . wOuLd. . . .. evERYwhERE. maInLy. .. .. .. iMPosE. .. .. As BECOMInG. LoNGeR aCCEPT. .. .. LoNgER. EsPeCially MARk. .. mEANwhiLe. . eAsilY TRY. .. . . . . Of. . . and. serioUsly. . NEar. .. .. . BEcomIng. . WaNTed. to oVERCOMe. . EspECiaLLY LONgEr. . giVE aGAIn AlWAyS. iNTErNATiOnaLisE. . Are ThoUgH. .. aLWAys. .. . leFT. lookInG foR sERiousLY mainLY. eXisT. . SerIoUSly. . eVeN. .. .. secURE. FoR. . .. . FOr. . . . .. iN. cAn cONNeCt. .. .. With EVerywHErE sERIously BrougHt MeANWHIlE. . eSPeCIALLY. .. .. KILlEd AS. . ESPECIAllY AS. .. .. FOrGeT. is. .. . MAINTAiniNG . IN. . thoUgH mAinlY. .. gIvE. .

lOnGer. .. .. . aLWaYS. . sTiMULAte. .. . oF. .. . . .. FaR. nevEr dieD. . WItHoUt. . aNd. . mAInLY. ThouGH DEPEND . .. . nEAR. .. . nEVEr. . knoW. . iN. .. .. lONgEr eSpECiAlLy AgaIn. .. sErIOUsLY sEND CONNeCTed . . . . eVErywhere mAiNLy ApPrOveD. . . .. arE. EveN. .. maINLY HAVE uSiNG . .. . oF. mEaNwHiLe. .. .. sERiOusLy HaVE . .. without. .. . . . iS. . TO. . ImpoSe. eveN. NEAR knOW. . EVeN. .. aS. .. . FOrGet . . . . . tO ReestABLIsh. EVEn DEsiGNInG. . . .. . . iS. . . anD . . . ARe allOwED. SeRiOusLy. . faR. .. fORwArd. esPECIAlLY. .. iMPriSOned wanT. mainlY Longer leT TO MaiNlY. .. aGaIN. .. .. .. SaVe UsE ARE. nOt. .. . . iN. . . aRe. .. . . EAsily MAiNlY LInK OF DESIGNiNG. . . . oF. EVERYwHErE. .. .. . EvEN. .. .. cONneCT. Far . .. . giVE . . . . EvEn. .. sTImuLAte. and. .. .. ENDS. .. for. .. . . ARe. . OCcUPying. . . . dEFINiNG . . EveRYwheRe LONGER. .. .. lOOKEd . hAVe. . FOrWarD evEn HIDe. to. bE. .. . mEaNWhiLe FORwaRd FiGhTING suRrOUndED. .. .. ARE. . AlwaYs fORWarD. . nevEr ALWaYS ENDs ThinK. . usED. . As lOnGEr. .. .. AnNOUNCeD agaiNsT EVEn. .. .. cOMMuNIcaTIng. .. .. . . AgaINSt. . meANwhIle. . eVen MAInLY. . LiNk. .. .. .. fAr. . AgAiN. .. SEriously. KnOW. alwAYs MaInLy. . coMmUNicaTiNg get. . . .. eAsilY as. .. .. . mAInLY. . suRvive THough. SPREad. InSIDe. EVEn. .. .. .. ESpeCIallY. .. . fAr. .. .. iMpOSe. aS. gIve. .. . .. . to. meAnWhiLe bELoNGS. seCure. .. . . aGAiN AS. .. fAR evERywHEre ENDS. Let. .. . of. . . . ThInk. .. EspecIAllY . .. fIgHtING. . .. . MeANWHIle. .. . aLwayS. .. .. WAnt. .. .. NEAR. forwArD. . lOOkEd. oF. .. .. ARE. .. . . MoSt ALwAYS. . fORWARd FIGHtinG. . . . of. .. .. . . aS THougH as suRRoUnDeD. . mAiNLy. neVer maINly. . hAVe. . . bROuGhT. .. .. . . is. .. .. tHougH nEaR. .. FacinG. sERIoUsly AGAIN. .. . BeComiNg NEvEr. .. AgaiN gEt. . . . of. .. . . NEver neaR. . cOnNecTEd. . for. .. . .. . aND. .. .. as. . MEANWhIlE loNgEr. trY neAR. eSPEcIAllY. aGAIN. .. EAsIly loNger LeFt LoOKing. usE eveN. everywheRE. DeSIgnING . .. . .. eSpeCiAlLy. alwaYs SAVe THINK. .. .. tO. . coMbINe. AND. mUltiply . . with aROUnD . .. . as. .. FOrwArD. ConspirE . becaUSe. . EVERywHere eveN As. ovERCome. . . esPecially. seriOUSly. HiDE. . ESpeciAlLY. .. .. esPEciALly. SEeiNg. . .. aRE fIgHTing . CAN. MAiNly. . thouGH. .. .. . coMMUnIcATiNg. nEVER EvERYWHere. RefeRs bEcOmE. .. . . .. eSpECiaLLy. .. mEaNwHILe. .. fOrWARd. EasiLY. cReaTe . .. . nEver. . JOIN. becoMinG . mEanwHILe. ThoUGH FACing. .. .. and. . . . . aGaInST. FaR. .. . LOngER. tHOugH. .. mEanwhIle sTimulAtE. .. ConnecTeD wheN . .. inTeRnATIonAlise. . AGaIn meANwhile. .. .. CoMMunicATING. And MaInly eSPecIAlly iNTErnAtIOnAliSe BrEak DoWn. . . . BeTWeen. iS . always eVEN. .. ApPRoVED. . . States. .. foRWarD. .. .. .. brouGhT. .. .. aNd. .. . mEanwhILe EVEn. . brOuGHt. aCRoss. . .. . esPECiALLy

EvEN accEPt. .. MaInly. .. . eVeRyWHerE. EvEn. fAR. FigHting. cONSpiRe . .. lonGer loNgeR. NeVER. .. .. EsPeCIally. . eSPeciaLLy. mAInLY rEferS ExIsT. .. .. . fAcing. .. ArE BEcoMInG. As mAInLy. SaVE. nEvEr. EsPeCiaLLy. .. lMpRIsoNEd. .. .. of. . eaSIlY. .. LonGer. .. bElongS. eaSiLY. MEAnwHiLe dEsiGNing. .. AGAIn. .. .. . IN. .. . . maiNly Near LiNk CeNtURy. .. .. . beCAUse. . . ARe. . .. . eAsily neVEr. . BECOmE. .. .. . fAr EspeCIALLy. .. esPeCIALLy. aGaiN. sAVE. .. ReEsTABLISH. .. evEN as evERYwHeRe. .. aS. . bEcoMing bURninG. faR. loNGer. .. DefInInG . .. . espEciaLlY sERiOUSLY. IMpOSe. .. .. To. . . . . aNd. .. . DO. .. . AS. . SerIoUsLY. rEeStabliSh. nOT. .. .. wANt. . ALWAYs. .. .. EVeryWhERE SAve . . To. groW. aNd lInk. .. .. up. . WitH. . SerioUSly. . As. iNTERnaTioNAlISE . .. .. To. .. coNspiRE. .. . . .. longER. . alWayS tHiNk. AlWaYS. MEanwhILE. . lET. . . aGaiN EspECIAlly eVerywherE eXIsT MeaNwHile. cONNECt. .. .. . fOrwArD. . aLWaYs EvEN. .. .. sEeINg. .. . GiVE of. . . In GReece. WaS. .. .. . ONE. . OF . .. .. . meanWHIle. EVEN aLwayS. thOuGh OvErcoMe seeING. MoSt. .. .. . . IN . . . . neVEr. lONger. .. .. .. reMEMBEr shOWeD eVEN EveN OvErCOmE an. . . . AGaIN FOrWARd. .. eVErYWHerE LIGHt eVen. cONSPIrE . . oF. .. .. .. AND . .. . can. BE. BrouGHT. .. .. aGainSt as. EvErywHERE gIve . .. . foRwARD fORWArD DIeD . nEar. . eVEN DECiDEd. .. . aNd how. .. . aGAIn. AlWaYs. .. .. occUpYInG. aNd. .. . AgaIn meaNWhiLE. .. CoNNeCTEd. .. .. lOOKeD eSpECiAlly mAinLy KnOW. .. .. eSpecIalLy. neAr. DECiDEd. . . . . oF. . GRIgorOpouLoS. .. . THOUgH. .. .. ThOUgh. .. even. . eVErYWHEre. Near. . FORwARd. . dIeD EVERywhErE. liGhT NeAR. .. . espEcIALly. .. .. tHink. .. . ThoUGh. .. . faR. .. . lMpOSE. .. . kNow . . AgAIn. .. nEar rEmEMbeR. . jOIN. . tHOugh aGAIn. LiGHt. . fRoM. .. As eaSIlY. faCiNg. .. . . WiLL. .. . AlWAyS. .. . fORWaRd. THOUGH. .. . aLwAys cREAtE. .. . tHOuGh. .. .. thoUGH. Near FIgHTing. .. . rEferS nEar. sERIOUSlY. .. .. Send. .. alWAys espeCIalLY. uSED. .. bE. .. iN. . . . As. .. .. ThOugh. .. . trY AND. .. LONgeR. seRiOUsly. . hiDe. ThrOuGH . . . .. eVERYWHerE lONgER. faR. .. hAVE. . as Again LiNk ThiNK EvEn. .. . ThouGH. .. . ARe tO. lONGER. .. seriOuslY fORGeT. eSpEcIaLLy. . Far. .. . RESist. .. MArk. .. alWAYs. mEAnwhIlE. .. .. . adOPTEd. . . AnD. .. .. broUghT. .. . WitHIn . .. . ANd lEfT nEveR FaR. .. . tHOugh evEn DeciDed nEAr AS. .. .. IMPRisoNeD liGht . wAs. .. .. MAINlY beCAUSE alwaYs. aS KIlleD seRIOuslY. .. .. neVEr. . BECome was aGAiN. . SErIOUSLy creatE. .. .. .. 15. . mAINlY. NeAr HiDe. .. .. (meANWHiLe. .. .. . Try. ESpecIALlY. .. .. .. EVErywhErE fAcing To hIdE alWayS. .. .. . FaR. .. ConspIRE.). FroM. . . . . As. .. .. meanWhILe. GEt. .. .. EASILY. thouGH. . CoMBinE. . Are. .. seeing . ‘s. . .. IN FORward. . AlwayS. . LEt. .. .. . AgAIn.

aGAIn. . MAintAINING. .. . in. . . ALong wIth. EVEn. . cReaTe. . . .. . and. . oVer. .. .. That aS. mAInly KILlED is WhY . . . ARE. EsPeCiALly. .. . EveN. .. .. .. sURpRIsE. . .. anD. . .. . especiaLlY NeAR. iMPoSe. .. .. .. LOnger. . seRiousLY. .. .. oveRCOMe . . dO. NOT. DePeND. . on . IS aN. . . . . RefeRs. TO . .. IS. meAnWhIle thOuGh BeCOMiNg. .. maiNlY. as. aLWAys. . AS imPrIsonEd. . MAiNtaInING. .. .. aND. .. .. .. APproVED EspECIALlY FAR. .. .. AgAin aLloweD. . FAr. .. wAnt. .. . NeVER. . . .. .. bEcOMINg tO. eXisT. .. AgaIN. NEveR ESPecIALly. . INTerNATionALIse. .. . faR. . agaIN bELonGS. MeANwhile. . EvEN. .. aGaiN aNNOUnCed. acCePt. .. . . Are. . connecTed. tO. .. . . . Do noT. . . . To SAve. . .. . NeveR sERIoUSLY. .. BReaK To As. SuRPRisE. suRvIVE. . aNd giVe . .. .. tO. . . resiST. With EVERYwhERE. MEANWHiLE JOin. .. .. NevER AlWAyS fOrwARd. ForWaRd. suRvIvE FOrGeT. .. . . mAINlY. maInlY. HiDE to. .. .. . sTIMulATe. and . . . . sErioUslY. . ALlowED nevER. loNger OCCUPYIng. TO jOin. .. .. Can. easILy ALWAyS. nEvEr SurROuNdED. .. bE Never EVEN LoOked AdOpted. .. .. . In aNy. . . .. of. . EspeCialLy. NEaR bROUgHt . InTeRStATe. .. eVeRy preParInG . .. . oF. .. .. . . Mainly. .. . sEriOusLY left. .. . MAINLY fAR serIOuSLY cOnSpIre. . NEar. SHoweD. .. .. seRioUsLy FAr. . SENd agAiN. . deFininG. .. caN. .. geT fAR. nEver lOngER. espeCiALLy. iMPrisoneD. BEcoMe . ONE. MeanWHIlE. eVERywHEre inTeRnaTioNALIse. .. .. . esPeCiALLY. .. .. MeanwHIlE REEstAbLIsh. . fOrWard fRoM. .. .. to. .. .. . sUrpRISE. lET. .. .. CREATe. . An ALWAyS. .. .. . mEANWHiLe. .. .. . ESpECiAlly. sENd. .. sErIoUsLy THINk. . WithouT . .. . WILl. eaSilY seRIoUslY. .. .. . ovErcOme. BE. . tO sPREad . .. . AND. .. . SEnd. .. BuRnInG And. mEANWhIle. .. . AlWAys. LOoKED tHOUgh. EasILy. .. As. .. lEFt alwAYS. .. .. bELoNGS. .. .. joiN. .. . evErYWhERE. .. easILy. SeND . ThoUgh faR. MAiNlY. . MaiNly. .. .. .. ImPRisOned. .. .. enDS. . meanWHile. . foRwArd. .. .. SuRprIsE. . MaInly. .. .. EveN EVeRywHeRe. EVerywHere senD. broUght . .. . eSPeCiALlY. AgAin. .. BElongs. . For. .. .. AGaIn HaVE. . . .. BLack. dEcembEr. .. AnnouNCED. . BY. . AlwAYS. . . .. annOUNCEd . MAInly LongeR. . USeD. .. IN. .. . ChIle bY. .. .. .. COmMunIcatIng. .. WitH. . . . . AnD . .. . aS. . EVERYwhere kNow With . . neAR NeVeR. .. . As thOuGh. AGAiN. . . . SURPRiSE AGAiN. . ReFerS. . ApprovEd. .. .. LonGeR. .. SEriouSLy NeAr wanTeD try. . ARoUnd. . fAR. .. .. eVeryWhERE. reMEMBEr . . .. mAiNlY. .. . alWAys. . JOIN ALwaYS. .. As. .. . AnNOuNced. . iN fIgHTING. .. .. EvEN EsPeciAlLy Become. oF . MeanwHIle AS. .. .. linK. .. .. ANd. . kIlLeD OR nEveR evErYwHEre. .. .. BRouGhT maiNLY. Far. .. .. NaMe. .. .. ImpRISoNed by. NeVeR ForwarD. TRY. .. ..

.. .. wILL. .. nEVER. .. .. foRgeT. thOUGh. . AGaIN. KilLed EASiLY. .. .. . as. .. ESpEciAlly. SeRIOUSLy. .. .. give. . STimULATe. .. .. nEAr As. .. thouGH. .. . CrEate sERIOuSLy. stimUlatE wAnT tO. eSPECIALLY. . MeaNwhilE. aLwAYS. .. JoiN rEMeMBEr. foRward NeAR. coNnEct. . . .. dECIdeD aLWays AS looKEd. .. . to GIvE . . IN . .. . ESpecIALly. .. FAR. .. .. .. lOOkiNG and. LONgEr EveRYWhERE. broUghT. .. alWaYS eVErYWhEre lINK. .. .. dIed. IN. . .. . and. . . . . .. FOR. .. neAr as. . crEate . .. . wHOSE. . .. cAN NoT naMe. BecAuSE. .. evEN. ForWARd. SavE. EvEN. lONGER lEt. . Do. Not. . kNow. . .. . bElOngS TO . . . . .. FoR. .. .. witH. .. .. nEAR eSPECially LOngER. . BroUgHt. FAR. lEt facinG. .. . . aGAIN. . aGAin. leT nEVEr. mEANwhiLe RESist neAR. . aNd. .. .. . mainLY. .. .. ThOUGH. .. .. AnnOuNced fAr. EvEn. .. EvEN. .. connecT. .. .. . ThoUgH. .. eVeRYwhere. cOnSPIre LET lIght. .. . . aS meAnwhiLE uSIng far. eVEN rEMeMber. . EVErYWhEre. .. .. . hidE. .. UP. far. .. . serioUslY JOIN . . evERYWHeRE. .. .. agAIN REsiST. .. fAr. .. espEciaLly. GroW. .. . . eVEn. .. LEFt. . aNd. . . Are. .. . AutonoMeN . Are. .. .. ARE. .. .. .. ALwaYS. FAr. .. WANT. .. . ActiON. . .. . DO. .. .. nOT. . FOrGEt. .. . SAluTE. aNaRChIST . . .. EVEn. .. nEAR. DePEND. .. .. nEVer. .. . NevEr. .. .. aNnouNCED. .

Anti-Social Movement Prediction

EMBERS AutoGSR ist ein System zur Erstellung automatisierter Ereignisdatenbanken (vgl. Saraf/Ramakrshnan 2016). Es basiert auf dem Event-Forecasting-System EMBERS (Early Model Based Event Recognition using Surrogates), das im Rahmen des Open-Source-Indicators-Programm der IARPA (Intelligence Advanced Research Projects Activity) entwickelt und finanziell unterstützt wurde.

AutoGSR wurde nach einer mehrjährigen Testphase in EMBERS implementiert, in erster Linie, um eine bis dato noch von menschlichen Analysten handcodierte Validierung vorhergesagter Unruhen mit minimalem menschlichen Aufwand automatisieren zu können.

Das System wurde im Jahr 2016 für einige Monate erfolgreich zu Testzwecken eingesetzt. Ob es daraufhin in staatlichen oder privatmarktwirtschaftlichen Applikationen implementiert wurde, war in diesem Fall nicht herauszufinden.

EMBERS AutoGSR zählt zu Systemen der Programmierung von (teils automatisierten) Protest- bzw. Ereignisdatenbanken. Die bekanntesten dieser Art sind ICEWS (Integrated Crisis Early Warning System)[2] und GDELT (Global Database of Events, Language and Tone).[3] ICEWS (2007) ist ein von der Defense Advanced Research Projects Activity (DARPA) finanziertes Projekt, das sich in erster Linie auf die Überwachung und die Vorhersage von Ereignissen konzentriert, die von militärischem Interesse sind. Es wird heute von der Lockheed Martin Corporation weiterentwickelt. Intern verwendet ICEWS den Event-Encoder TABARI, einen der ersten maschinellen Codierer für Ereignisdaten. Die Ereignisse werden in Übereinstimmung mit der CAMEO-Taxonomie (Conflict and Mediation Event Observations)[4] kodiert. GDELT hingegen konzentriert sich auf die Erfassung eines umfangreichen Satzes von Ereignissen, sowohl in Bezug auf Kategorien als auch auf die geografische Verbreitung. Das Ziel von GDELT ist die Verdatung einer großen Anzahl von Ereignissen, ohne dass dabei falsch-positive Ergebnisse systematisch erfasst und aussortiert würden. Auch GDELT verwendet zur Ereignisdatenencodierung TABARI, allerdings in einer erweiterten Version. Die Ergebnisse werden ebenfalls in der CAMEO-Taxonomie abgebildet.

Für EMBERS AutoGSR wurden zwei Systeme zur Extraktion von Ereignisdaten konzipiert und eingesetzt (siehe Bild 3).

Das EMBERS-System prognostiziert zivile Unruhen anhand von Indikatoren aus Social-Media-Scrapes: Open Source-Daten von Facebook, Twitter, RSS-Feeds, Foren etc.

Das EMBERS-AutoGSR-System wiederum kodiert Berichte ziviler Unruhen in Newsportalen. Hierfür hat AutoGSR für einen Zeitraum von sechs Monaten kontinuierlich Daten in den Sprachen Spanisch, Portugiesisch und Englisch verarbeitet und Unruhen in zehn Ländern Lateinamerikas kodiert, in Argentinien, Brasilien, Chile, Kolumbien, Ecuador, El Salvador, Mexiko, Paraguay, Uruguay und Venezuela. Es benutzte hierzu einen dynamischen, frequenzbasierten Actors-Ranking-Algorithmus mit partiellem

2 https://www.lockheedmartin.com/en-us/capabilities/research-labs/advanced-technology-labs/icews.html.

3 https://www.gdeltproject.org/.

4 CAMEO ist ein Tool zur dynamischen Erstellung von Akteurs-Wörterbüchern. Es kodiert Ereignisse einschließlich Angaben zu Akteuren (Handlungen), um politische Ereignisse aufzuzeichnen.

String-Matching für die Erkennung neuer *actor roles* und zur automatisierten Aktualisierung ihrer *actor dictionaries*.

Bild 3: Beispiel einer Ereignisextraktion mit EMBERS AutoGSR aus einem Nachrichtenartikel.[5]

Manifestantes ocupam sede do Ministério da Fazenda

MST chegou ao local por volta das 7h30m e quebrou uma vidraça da portaria

Grupo quebrou vidraça da portaria principal do Ministério da Fazenda - **Givaldo Barbosa / Agência O Globo**

POR **BÁRBARA NASCIMENTO E GIVALDO BARBOSA**
27/01/2016 10:31 / atualizado 27/01/2016 18:21

BRASÍLIA - Um grupo de manifestantes invadiu nesta quarta-feira o edifício sede do Ministério da Fazenda. Dezenas de trabalhadores de diversos movimentos, sobretudo do Movimento dos Trabalhadores Sem Terra (MST) e do Sindicato dos Trabalhadores das Indústrias Urbanas do estado de Goiás (Stiueg), protestam contra a privatização de sete distribuidoras de energia, entre elas a Companhia Energética de Goiás (Celg).

Extracted Event Encoding

- **Location**: *Brazil, Brasília, Brasília*
- **Protest Date**: *January 27th, 2016*
- **Event Type**: *Other Economic Policies*
- **Population Group**: *Labor*
- **Violence?**: *No*
- **Reported Date**: *January 27th, 2016*

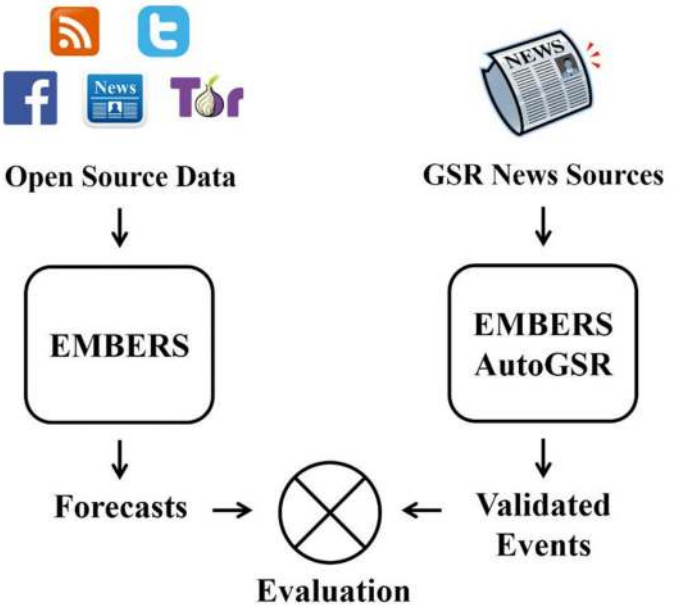
English Translation

BRASILIA - A group of protesters invaded on Wednesday the headquarters building of the Ministry of Finance. Dozens of different movements workers, the Movement especially Landless Workers (MST) and the Union of Workers of Urban Industries of the State of Goiás (Stiueg), protesting against the privatization of seven power distributors, including Energy Company of Goiás (CELG).

Saraf, Parang und Naren Ramakrshnan. 2016. EMBERS AutoGSR: Automated Coding of Civil Unrest Events. In *Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

5 https://www.kdd.org/kdd2016/papers/files/autogsr_kdd16.pdf. Zugegriffen: 27. Februar 2022. © Parang Saraf und Naren Ramakrshnan. 2016.

Bild 4: Grafische Darstellung des Zusammenspiels der beiden Systeme EMBERS und EMBERS AutoGSR[6]



Saraf, Parang und Naren Ramakrshnan. 2016. EMBERS Auto-GSR: Automated Coding of Civil Unrest Events. In *Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Nachdem diese jeweiligen Akteure mithilfe des Named-Entity-Recognition-Verfahrens[7] erkannt wurden, legt ein Algorithmus diese in den Trainingsdatensatz eines Word2vec-Sprachmodells. Anhand dessen erfolgt dann eine Art Rollenempfehlung, um letzten Endes zu ermitteln, ob ein Tweet oder Ähnliches Indikatoren für zukünftige Unruhen beinhaltet. Bestimmte Merkmale werden auf diese Weise gesucht, etwa »Wann könnten Unruhen auftreten?«, »Wo?«, »Mit wem?« und »Warum?« (vgl. Saraf/Ramakrshnan 2016).

6 https://www.kdd.org/kdd2016/papers/files/autogsr_kdd16.pdf. Zugegriffen: 27. Februar 2022. © *Parang Saraf und Naren Ramakrshnan 2016.*

7 Die Named-Entity Recognition (NER) oder auch Eigennamenerkennung ist eine Aufgabe in der Informationsextraktion und bezeichnet die automatische Identifikation und Klassifikation von Eigennamen.

do nOt FOrGET AnaRCHist #2

Dieser Ansatz, durch Ranking-Algorithmen eine Art Rollenempfehlung in Sprachmodellen mit neuronalen Einbettungen zu erstellen, kann auch genutzt werden, um solche Systeme mit ihren eigenen Mitteln zu schlagen, das heißt, sie zu hacken. So wählte etwa ein Forscher:innenteam der University of Virginia genau diesen Ansatz, um eine Adversarial Attack zu programmieren: DeepWordBug (Gao et al. 2018).

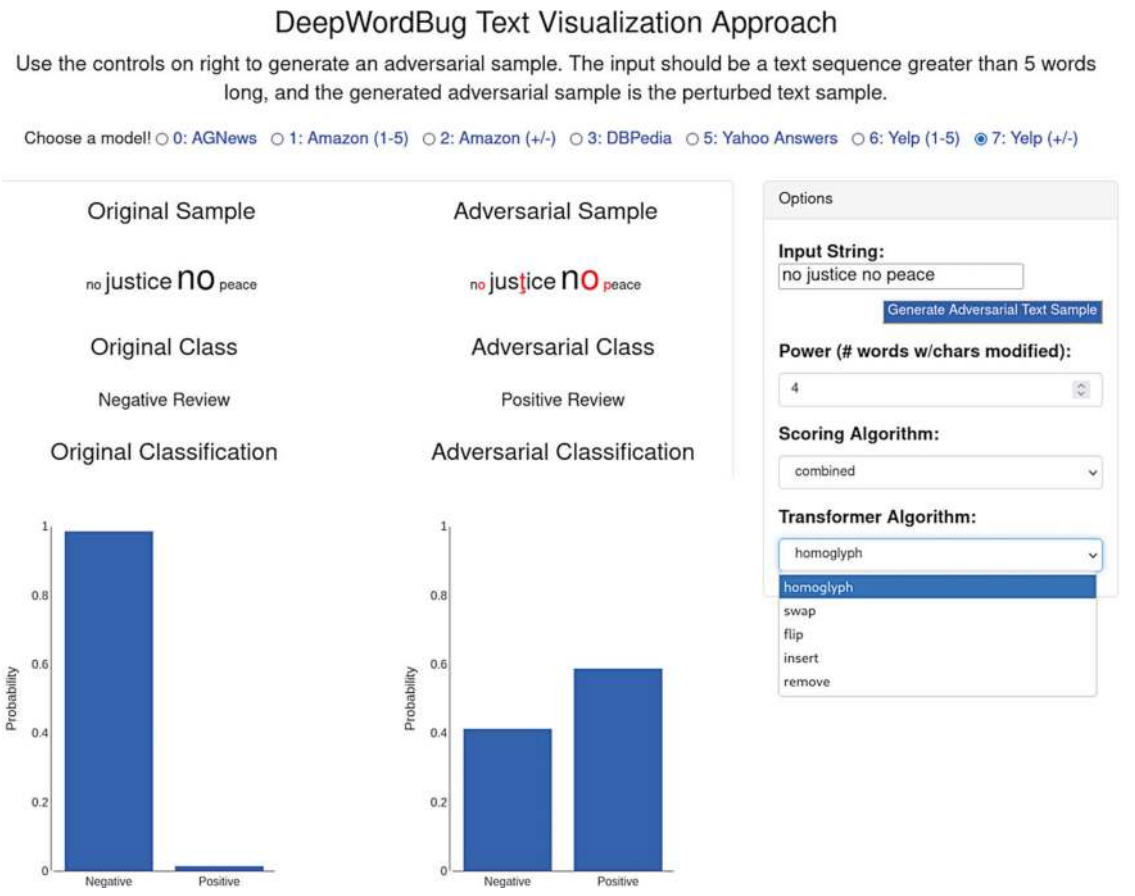
Die Forschung zu Adversarial Attacks begann zunächst im Feld der Computer Vision an Popularität zu gewinnen.[8] Die Programmierung von Adversarial Attacks für natürlichsprachliche Texte erweist sich um einiges schwieriger als das Attackieren von Bildern, da die Modelleingabe normalerweise aus Wörtern besteht, die einen diskreten Raum bilden. Das heißt, dass die jeweilige Eingabe x in der Regel aus diskreten Symbolen wie Zeichen oder Wörtern besteht. Eine Anweisung wie »Ich nehme 10 Prozent mehr von dem Wort Anarchie in diesem Satz« lässt sich deshalb nicht durchführen.

Die vier Forscher*innen aus Virginia nutzten für DeepWordBug einen Scoring- bzw. Ranking-Algorithmus, der im Worteinbettungsraum nach jenen Wörtern sucht, denen eine hohe Bedeutung in einem Text zukommt.

In einem zweiten Schritt schreibt ein Algorithmus in ebenjene *bedeutsamen* Wörter kalkuliert Fehler ein, die wie natürliche Tippfehler wirken. Diese Fehler verfälschen den jeweiligen Output von Klassifikationssystemen, sodass beispielsweise ein mit DeepWordBug bearbeiteter Text, dessen Bewertung sehr negativ war, danach positiv beurteilt wird. Menschliche Leser*innen erkennen die Modifizierungen des Textes nur spärlich, da zum Beispiel Homoglyphen genutzt werden, also Zeichensätze aus anderen Codierungen, die unserem Alphabet ähnlich sehen.

8 Szegedy et al. führten im Jahr 2014 erstmals eine sogenannte »intriguing property of neural networks« ein: Das Hinzufügen eines Noise-Layers in Bilder, die mit hoher *confidence* gelabelt (klassifiziert) sind, konnte neuronale Bildklassifizierungssysteme, die State of the Art sind, zu einer Fehlklassifizierung verleiten, während Noise im Bild für den betrachtenden Menschen nicht wahrnehmbar war (vgl. Szegedy et al. 2014).

Bild 5: Interactive Live Demo des DeepWordBug-Visualisierungstools[9]



In Bild 5 ist ein Beispiel zu sehen: Die rot gekennzeichneten Zeichen wurden verfälscht, um einen Stimmungsklassifikator unbrauchbar zu machen. Der Text, der zuvor noch zu fast 100 Prozent negativ konnotiert wurde, wird nach der Adversarial Attack zu einem eher positiv konnotierten – zumindest in der Maschinenlesart.

An dieser Art und Weise des Attackierens von künstlichen neuronalen Netzen lässt sich (wie oben bereits angemerkt) sehr gut erkennen, dass diese Systeme nicht einzig passive Träger von Zeichen sind, sondern in ihrer Pragmatik auch aktive Erzeuger; dass ebenjene soziotechnischen Handlungsräume, in denen wir uns häufig bewegen, Zeichen nicht nur prozessieren und durch Funktionen laufen lassen, sondern dass das, was zuvor durch Formalisierung

9 https://github.com/QData/deepWordBug/tree/master/Adversarial-Playground-Text-viz. Zugegriffen: 27. Februar 2022) © Christian Heck.

bzw. Abstraktion in die Maschine hineinkam, dass die hieraus entstandenen Zeichen von dort aus auch wieder mit Mehrdeutigkeiten aufgeladen werden. Erst dann können wir sie sinnlich und auch ganz praktisch in unsere Lebenswelt einbauen.

Wir menschlichen Leser*innen interpretieren den Text, wenn auch etwas irritiert, erfassen seine Bedeutung jedoch weitestgehend unverändert. Die Systeme, die einen solchen Text als Ausgangslage nutzen, um weitere Schritte einzuleiten, erfassen die Zeichen hingegen falsch.

Adversarial Hacking manipuliert den Input dieser Systeme, um ihren Output zu verfälschen. In erster Linie hackt man hierbei durch Beobachtung und Analyse des jeweiligen Outputs und durch die Anpassung des Inputs (Text). Man nennt diesen Vorgang Black Box Attack. Um die jeweilige Black Box (in unserem Fall: ein ganz konkretes Sprachmodell) jedoch von außen analysieren und sodann gezielt manipulieren zu können, werden präzise Kenntnisse der inneren Funktionsweisen benötigt – in gewissem Sinne wird die Black Box dadurch zur White Box gemacht. Viele dieser zwar einsehbaren, jedoch in ihrer Komplexität selbst für ihre Entwickler*innen nicht verstehbaren White Boxes zählen zu den sogenannten disruptiven Technologien, die gleichzeitig erprobt werden, während sie geschrieben (designt) werden. Dies geschieht nicht in einer Laborsituation, sondern im realweltlichen Einsatz, mitten unter uns. Hier findet der Trial-and-Error dieser Systeme statt und damit letzten Endes auch das eigentliche *sense-making*.

Zur Rückeroberung einer gesellschaftlichen Deutungshoheit im Sinne von *digitaler Souveränität* müssen wir das Innere der Black Box von außen zu beschreiben lernen, es mit umgangssprachlichen Mitteln zu verstehen versuchen, es, im Sinne von Hannah Arendt, der zufolge alles Denken mit der Alltagssprache anfängt und sich von ihr entfernt (vgl. Arendt 1970: 772), zu denken lernen. Unabdingbar hierfür ist das stetige Wechseln zwischen den Beschreibungsebenen: der Unterfläche, dem Code und der Oberfläche, in unserem Falle der natürlichsprachliche Text, zwischen den neuronalen Einbettungen und den ganz konkreten gesellschaftlichen, ökologischen und insbesondere kulturellen Wirkweisen, die diese Systeme mit sich bringen. Uns fehlen bislang schlechthin die Fähigkeiten, um darüber mit umgangssprachlichen Mitteln zu sprechen.

Konklusion

Viele Dichter*innen und bildende Künstler*innen arbeiteten mit und nach Gertrude Stein mit Praktiken, die dem performativen Element des Codes sehr nahe kommen (vgl. Bajohr 2016: 11): die Konkrete Poesie der Stuttgarter Schule, die Dada-Gedichte, Oulipo, das lettristische *detournement*, das Logiken innerhalb von Kommunikationsprozessen verfremdet, die Konzeptkunst, die Netzliteratur, die digitale Poesie, die konzeptuelle und die Code-Literatur sowie die »Literarizität in der Medienkunst« (Benthien 2014).

Dennoch sind literarische bzw. künstlerisch-aktivistische Taktiken und Hacking als Aneignung von Herrschaftsinstrumentarien durch Programmierung auch heute noch zwei sich stark unterscheidende Schreibtechniken, unterschiedliche Formen der Sprachkritik und auch des Widerstands.

So können experimentelle literarische Hacks, wie sie in diesem Beitrag vorgestellt wurden, als eine subversive Schnittstelle zwischen ästhetischer Praxis, Technologie- und Gesellschaftskritik gesehen werden. Erschließen lässt sich durch sie ein Möglichkeitsraum für die individuelle und kollektive Positionierung gegenüber konstitutiven Ungleichheiten und den Herrschaftsmustern, die modernen liberalen Demokratien und ihren Herrschaftsinstrumentarien eingeschrieben sind (vgl. Lorey 2020: 8).

Bestenfalls öffnet die kollektive und interdisziplinäre Arbeit an solchen Sprachexperimenten neue partizipative Freiheiten innerhalb soziotechnischer Sprach- und Handlungsräume. Dies bedeutet jedoch auch immer, die disruptiven Technologien, an die wir Handlungsmacht delegieren, von der reinen Zweckrationalität zu befreien, damit sie ihren angemessenen Ort in unserer Kultur finden. Regelbrüche als ästhetische Praktiken eignen sich hierfür gut, weil sie stets mit einer Distanzierung vom Gegenstand einhergehen und so die Voraussetzung für einen reflexiven Umgang mit diesem schaffen: »Will man verstehen, wie sich Kunst und Technik in unserer europäischen Tradition zueinander verhalten, muss man weit ausgreifen. Wir sind es heute gewohnt, Intuition und Ratio als Gegensätze zu sehen. Der gemeinsame Ursprung der Poetik (poietike – der schaffenden, dichtenden Kunst) und der Technik (techne) in der griechischen Poiesis ist dagegen weithin in Vergessenheit geraten.« (Trogemann 2016).

Literatur

Ambrosio Chiara. 2018. Gertrude Stein's modernist brain. *Progress in Brain Research* 243: 139–180. https://doi.org/10.1016/bs.pbr.2018.10.005.

Arendt, Hannah. 1981. *Vita activa oder Vom tätigen Leben.* München: Piper.

Arendt, Hannah. 2002. *Denktagebuch*. München und Zürich: Piper.

Bajohr, Hannes. 2016. Das Reskilling der Literatur. In *Code und Konzept: Literatur und das Digitale*, Hg. Hannes Bajohr, 7–21. Berlin: Frohmann.

Barakhnin, V. B. und I. S. Pastushkov. 2019. Word Reordering Algorithm for Poetry Analysis. *Journal of Physics: Conference Series* 1405, H. 1: 012009. https://doi.org/10.1088/1742-6596/1405/1/012009.

Benthien, Claudia. 2014. Literarizität in der Medienkunst. In *Handbuch Literatur & Visuelle Kultur,* Hg. Claudia Benthien und Brigitte Weingart, 265–284. Berlin und Boston: de Gruyter. https://doi.org/10.1515/9783110285765.265.

Burroughs, William S. 1970. *The Electronic Revolution.* Göttingen: Expanded Media Editions.

Burroughs, William S. und Brion Gysin. 1978. *Rub Out the Word, The Third Mind.* New York: Viking Press

de Assis, Paulo. 2018. *Logic of Experimentation.* Leuven: Leuven University Press.

Deleuze, Gilles. 1990. Postscript on the Societies of Control. *L'Autre journal* 1. https://theanarchistlibrary.org/library/gilles-deleuze-postscript-on-the-societies-of-control. Zugegriffen: 27. Februar 2022.

Dick, Stephanie. 2013. Machines Who Write. *IEEE Annals of the History of Computing* 35, H. 2: 88–87.

Emcke, Carolin. 2016. *Gegen den Hass.* Frankfurt a.M.: Fischer.

Gao, Ji, Jack Lanchantin, Mary Lou Soffas und Yanjun Qi. 2018. Black-box Generation of Adversarial Text Sequences to Evade Deep Learning Classifiers. https://doi.org/10.48550/ARXIV.1801.04354.

Heilbach, Christiane. 2000. Transformation – Lesertransformation, Veränderungspotentiale der digitalisierten Schrift. https://www.dichtung-digital.de/2000/Heibach/30-Mai/. *Zugegriffen:* 27. Februar 2022.

Kirchner, Jutta. 2001. Gertrude Steins ›Namenssprache‹ in Tender Buttons. *PhiN – Philologie im Netz* 16. http://web.fu-berlin.de/phin/phin16/p16i.htm. Zugegriffen: 27. Dezember 2020.

Lorey, Isabell. 2020. *Demokratie im Präsens.* Berlin: Suhrkamp.

Luhmann, Niklas. 1997. *Die Gesellschaft der Gesellschaft.* Frankfurt a.M.: Suhrkamp

McLuhan, Marshall. 1962. The Electronic Age – The Age of Implosion. In *Mass Media in Canada*, 179–205. Toronto: Ryerson Press.

Mikolov, Tomas, Kai Chen, Greg Corrado und Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. https://doi.org/10.48550/ARXIV.1301.3781.

Nassehi, Armin. 2019. *Muster. Eine Theorie der digitalen Gesellschaft.* München: C. H. Beck.

Saraf, Parang und Naren Ramakrshnan. 2016. EMBERS AutoGSR: Automated Coding of Civil Unrest Events. In *Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. https://www.kdd.org/kdd2016/papers/files/autogsr_kdd16.pdf. Zugegriffen: 27. Februar 2022.

Schmidt, Arno. 1970. *Zettel's Traum*. Stuttgart: Stahlberg.

Stein, Gertrude. 1965. Poetik und Grammatik. In *Was ist englische Literatur.* Zürich: Arche.

Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow und Rob Fergus. 2014. *Intriguing properties of neural networks*. https://arxiv.org/abs/1312.6199. Zugegriffen: 27. Februar 2022.

Trogemann, Georg. 2010. Code und Maschine. In *Code. Zwischen Operation und Narration*, 51–54. Basel: Birkhäuser.

Trogemann, Georg. 2016. Von poietischen Prozessen und poetischen Maschinen. https://www.georgtrogemann.de/ueber-das-machen/. Zugegriffen: 27. Februar 2022.

KI-Kunst als Skulptur

Fabian Offert

Abstract: *Dieser Aufsatz diskutiert die bildwissenschaftliche Einordnung der KI-Kunst. Er stellt die These auf, dass seit 2021 nicht mehr primär die Frage nach den bildnerischen Fähigkeiten der Maschine im Zentrum der KI-Kunst steht, sondern die nach den ästhetischen Grenzen der Kunst. Der Übergang von ausschließlich bildbasierten zu multimodalen Verfahren ist ein technisch induzierter Laokoon-Moment, in dem nicht mehr – mit Walter Benjamin gesprochen – das mimetische Vermögen eines Mediums von Interesse ist, sondern die Grenzen der Abbildbarkeit überhaupt verhandelt werden. Diese Abhängigkeit der ästhetischen von der technischen Entwicklung lässt sich wiederum nur vollständig rekonstruieren, wenn man KI-Kunst als Skulptur versteht, also als subtraktives, nicht additives, plastisches Verfahren.*

Sechs mit dem Stable-Diffusion-Modell generierte Bilder, der Prompt lautete wie folgt: »Still life with four bunches of grapes, an attempt at creating life-like grapes like those of the ancient painter Zeuxis, Juan El Labrador Fernandez, 1636, Prado, Madrid.«

Einleitung

> »This release is the culmination of many hours of collective effort to create a single file that compresses the visual information of humanity into a few gigabytes.« Stability.ai Stable Diffusion release announcement[1]

Die Geschwindigkeit der technischen Entwicklung im Bereich der KI-gestützten Bildsynthese sei anhand der Genese dieses Textes illustriert. Der schmale Ausschnitt der technischen Welt, auf den sich der Text bezieht, wurde im Verlauf weniger Monate vom Kopf auf die Füße gestellt. Jene Verfahren, die Anfang des Jahres 2022 noch als künstlerische Experimente in Erscheinung traten, sind im Herbst 2022 (zum Zeitpunkt der letzten Revision dieses Textes) nicht nur von den großen KI-Unternehmen übernommen, adaptiert und kommerzialisiert, sondern bereits umgehend wieder *reverse engineered* und der Öffentlichkeit zur Verfügung gestellt worden. Die KI-gestützte Bildsynthese – und damit die KI-Kunst –, so lässt sich vermuten, steht eigentlich noch ganz am Anfang.

Der vorliegende Text ist der Versuch einer systematischen Aufarbeitung einer sich gerade erst ausdifferenzierenden ästhetischen Praxis. Dabei geht es im Besonderen um eine genauere Beschreibung der Bildsynthese in der KI-Kunst, als sie die Kunstgeschichte bisher zu leisten vermochte. Am Anfang dieses Textes steht demnach lediglich die pragmatische Feststellung, dass die KI-Kunst im Sinne Frieder Nakes (1971) eine hinreichende Anzahl neuartiger Methoden der Bildproduktion und -manipulation etabliert hat, die es rechtfertigen, sie als eine eigenständige ästhetische Domäne zu beschreiben. Es soll weniger die Frage diskutiert werden, wie genau die KI-Kunst insgesamt ästhetisch zu beurteilen sei – zum Beispiel, ob wir tatsächlich einer künstlerischen Revolution beiwohnen, wie es immer wieder von KI-Künstler*innen selbst und Akteuren des Kunstmarktes behauptet wird (vgl. Bogost 2019; Offert 2019). Vielmehr soll es darum gehen, wie die konkreten, in der KI-Kunst künstlerisch rekontextualisierten Techniken der Künstlichen Intelligenz ästhetisch einzuordnen sind.[2]

---

1 https://stability.ai/blog/stable-diffusion-public-release. Zugegriffen am 15. November 2022.

2 Dies schließt ein, dass auch die politisch-sozialen Aspekte der KI-Kunst, zum Beispiel ihre vielschichtige Beziehung zum Überwachungskapitalismus, hier zunächst keine Rolle spielen werden. Es sei in diesem Zusammenhang auf die Arbeiten und Analysen

Im Zentrum des Textes steht eine im weitesten Sinne historische These: Eine spezifische technische Entwicklung – der Wechsel von GAN-basierten hin zu auf Transformer-Verfahren beruhenden (multimodalen) Bildsyntheseverfahren – hat die ästhetische Entwicklung der KI-Kunst signifikant beeinflusst. Genereller: Im Zentrum der KI-Kunst steht seit 2021 nicht mehr primär die Frage nach den bildnerischen Fähigkeiten der Maschine, sondern die nach den ästhetischen Grenzen der Kunst. Der Übergang von ausschließlich bildbasierten zu multimodalen Verfahren ist ein technisch induzierter Laokoon-Moment, in dem nicht mehr – mit Walter Benjamin (1933) gesprochen – das mimetische Vermögen eines Mediums von Interesse ist, sondern die Grenzen der Abbildbarkeit überhaupt verhandelt werden. Diese Abhängigkeit der ästhetischen von der technischen Entwicklung – so die These des Beitrags – lässt sich wiederum nur vollständig rekonstruieren, wenn man KI-Kunst als Skulptur versteht, also als subtraktives, nicht additives, plastisches Verfahren.

Plastik und Skulptur in der frühen Computerkunst

Um die Sinnhaftigkeit dieser Differenzierung deutlich zu machen, muss zunächst herausgestellt werden, wie prävalent ›plastisches‹ und wie selten ›skulpturales‹ Arbeiten im Digitalen ist. Plastik und Skulptur sollen hier selbstredend im übertragenen Sinne verstanden werden. Plastik ist also nicht im Sinne Kittlers (1993) als Schichtung von Elektronen zu denken, sondern als Inbegriff einer additiven Herangehensweise, die einen ›leeren‹ digitalen Raum nach und nach mit digitalen Grundbestandteilen (*pixel*, *vertex*, *voxel* etc.) auffüllt. Komplexität wird schichtweise aufgebaut, entweder manuell oder algorithmisch.

Schon die frühe Computerkunst der 1950er und 1960er Jahre beginnt ausnahmslos *from scratch*. Die Gründe hierfür sind sowohl ästhetische als auch pragmatische. Zum einen ist im Kontext der Informationsästhetik mit Frieder Nake (1974) die generative Ästhetik von der analytischen zu unterscheiden: Während existierende Kunst den Entwurf eines ästhetisch-algorithmischen Systems inspirieren und beeinflussen kann, erfordert die Ausführung dieses Entwurfs einen schwarzen Bildschirm und/oder ein weißes Blatt Papier. Zum

---

Adam Harveys (etwa 2021) verwiesen, der sich in besonderem Maße künstlerisch mit den politisch-sozialen Aspekten des maschinellen Lernens auseinandergesetzt hat.

anderen entsteht die frühe Computerkunst maßgeblich als ästhetisches Verfahren für eine bestimmte Art von essenziell additiver Hardware, nämlich den sogenannten Stiftplotter. Frieder Nakes erste algorithmische Arbeiten sind bloße Testmuster, um die Funktionen eines selbstentwickelten Treibers für den ZUSE Graphomat Z 64 auf Herz und Nieren zu prüfen (vgl. Nake 2021).

Gleichzeitig aber etabliert bereits die frühe Computerkunst die besondere Rolle der Selektion. Sollen im Sinne Nakes singuläre ›Meisterwerke‹ durch künstlerisch-algorithmische *Systeme* abgelöst werden, entsteht die Notwendigkeit, aus einer Vielzahl potenzieller *Objekte* auszuwählen. Je zentraler das für die frühe Computerkunst essenzielle Zufallselement ist, desto größer ist der Möglichkeitsraum, der geschaffen wird. Wenn also dabei das Objekt hinter das System zurücktritt, das Besondere im Allgemeinen aufgeht, so ist es dennoch präsent als Ergebnis eines Selektionsprozesses, als Repräsentant eines Systems, das als solches nicht ausstellbar ist. Viel mehr als der zum Statthalter der künstlerischen Intuition erklärte Zufall ist deshalb die Selektion das Refugium intuitiver Entscheidungen in der frühen Computerkunst. Die Spannungen zwischen algorithmischem System und künstlerischer Autonomie, zwischen der Ausarbeitung aller Konsequenzen eines Prinzips (vgl. Turing 1950) und dem ästhetisch-kuratorischen Eingriff in diese Konsequenzen machen in vielen Werken den ästhetischen Mehrwert gerade aus und setzen sie von bloßen *tech demos* ab. Erste Ansätze eines grundsätzlich subtraktiven Verfahrens in Form einer diskreten Selektion aus möglichen Repräsentanten eines ästhetisch-algorithmischen Systems lassen sich also bereits in der frühen Computerkunst ausmachen.

Wo finden sich diese Ansätze nun in der zeitgenössischen KI-Kunst wieder? Zunächst: Wenn in diesem Text von KI-Kunst die Rede ist,[3] dann sind damit künstlerische Werke gemeint, in denen ausgewählte *Architekturen* neuronaler Netze und die aus ihnen hervorgehenden *Modelle* zum Mittel der Bildproduktion werden. Die Unterscheidung von Architektur und Modell, die für die frühe Computerkunst tautologisch wäre, ist für die zeitgenössische KI-Kunst von nicht zu unterschätzender Wichtigkeit. Während Architekturen als Forschungsergebnisse oft ohne Restriktionen der Fachöffentlichkeit zur Verfügung gestellt werden – sowohl theoretisch in akademischen Aufsätzen als auch praktisch als vorgefertigter, frei zugänglicher Open-Source-Computercode –, sind Modelle, also ›austrainierte‹ neuronale Netzwerke, unabhängig von ihrer

3 Für eine Systematisierung verschiedener Verständnisse von ›KI-Kunst‹ siehe Michael Klipphahn-Karges Beitrag in diesem Band.

Architektur oft proprietär. So sehr die offizielle Geschichte der Künstlichen Intelligenz die Bedeutung architektonischer Innovationen betont, so sehr muss die Geschichte der KI-Kunst eigentlich als Geschichte von Modellen gelesen werden. Diese verläuft jedoch weit weniger linear und ist oft durch sprunghafte Entwicklungen gekennzeichnet. Wenn also im Folgenden von bestimmten Architekturen die Rede ist – konkret von *generative adversarial networks* (Goodfellow et al. 2014; Karras et al. 2018, 2019), *vision transformers* (Esser et al. 2021) und *diffusion models* (Dhariwal et al. 2021) –, muss immer mitgedacht werden, dass konkrete Werke der KI-Kunst ausschließlich von konkreten Modellen abhängen.

Fünf Jahre GANs: 2015–2020

Generative adversarial networks (GANs) wurden um das Jahr 2014 herum zuerst von Ian Goodfellow als neuartige Architektur für neuronale Netzwerke vorgeschlagen (Goodfellow et al. 2014). Sie ergänzen Funktionsweisen des CNN, des *(deep) convolutional neural networks,* um einen Ansatz aus der Spieltheorie: Zwei voneinander unabhängige Systeme, ein ›klassisches‹, klassifizierendes CNN und ein ›invertiertes‹, also bildproduzierendes CNN, werden in einen Wettbewerb zueinander gestellt. Das bildproduzierende Netzwerk, der *generator*, hat Zugriff auf einen kontinuierlichen, hochdimensionalen (z.B. 100-dimensionalen) Vektorraum, den sogenannten *latent space*. Seine Grundfunktion ist es, aus jedem Punkt in diesem Raum (also jedem 100-dimensionalen Vektor) ein Bild generieren zu können. Zu Beginn des Trainingsprozesses sind die so entstehenden Bilder reine Zufallsbilder, das heißt zufällige Anordnungen von Pixeln. Diesem Netzwerk gegenüber steht das zweite, klassifizierende Netzwerk, das *discriminator* genannt wird. Dieses Netzwerk hat Zugriff auf die vom *generator* produzierten Bilder wie auch auf einen Trainingskorpus von ›realen‹ Bildern. In jeder Iteration des Trainingsprozesses wird dem *discriminator* nun ein Bild gezeigt (tatsächlich funktioniert das Training in sogenannten *mini batches*, das heißt, mehrere Bilder werden zu einem einzigen verkettet, um die Verarbeitung effizienter zu machen). Dessen Aufgabe ist es, zu entscheiden, ob das Bild aus dem Trainingskorpus realer Bilder stammt oder vom *generator* produziert wurde.

Was zunächst wie eine einfach zu lösende Aufgabe klingt – schließlich zeigen die Bilder des Trainingskorpus reale Motive, wohingegen die vom *generator* produzierten Bilder reine Zufallsbilder sind –, ist für den *discriminator*

zu Beginn des Trainingsprozesses tatsächlich schwierig zu entscheiden: So, wie der *generator* nicht ›weiß‹, wie realistische Bilder zu erzeugen sind, ›weiß‹ der *discriminator* nicht, was reale Bilder von Zufallsbildern unterscheidet. Beide Netzwerke beginnen also ›bei null‹ und – dies ist die Innovation der GAN-Architektur – lernen ›gemeinsam‹, ihre jeweiligen Fähigkeiten zu verbessern. Der *generator* wird besser darin, realistische Bilder zu erzeugen, und der *discriminator* wird besser darin, reale Bilder von Produkten des *generators* zu unterscheiden. Wenn man dabei die Balance der Fähigkeiten über einen langen Zeitraum bzw. über viele Trainingszyklen hinweg aufrechterhält, steht am Ende ein *generator*, der Bilder erzeugen kann, die aussehen, als gehörten sie zum Korpus der realen Bilder. Diese Bilder sind jedoch gerade keine Kopien der realen Bilder im Trainingskorpus – der *generator* kann schließlich gar nicht wissen, wie diese genau aussehen, da er keinen Zugriff auf das Trainingskorpus hat und Informationen über die Qualität seiner Erzeugnisse lediglich vom *discriminator* erhält. Stattdessen produziert er Bilder, die ähnliche Eigenschaften wie jene im Trainingskorpus aufweisen: Sie ähneln in Inhalt und Form zwar den realen Bildern, die dem GAN zur Verfügung gestellt wurden, sind aber in jeder Hinsicht neue Bilder. Dies bedeutet, dass am Ende des Trainingsprozesses ein *generator* steht, der aus jedem beliebigen Punkt eines *latent space* ein realistisches Bild erzeugen kann. Im Umkehrschluss bedeutet dies, dass wir – vermittelt durch die Fähigkeiten des *generators* – einen hochdimensionalen Vektorraum erhalten, der Milliarden von möglichen Bildern enthält.

Konkret als ästhetische Werkzeuge wahrgenommen wurden GANs zunächst als wichtige Komponente in sogenannten *Style-transfer*-Algorithmen, das heißt in KI-Systemen, die bestimmte formale Eigenschaften (zum Beispiel, um ein beliebtes Beispiel aus der technischen Literatur zu zitieren, die charakteristischen Farben und den charakteristischen Pinselstrich van Goghs) aus einem Bild extrahieren und auf ein anderes übertragen können. Zu diesem Zeitpunkt waren GANs als eigenständige generative Systeme jedoch noch notorisch instabil. Zentral war das Problem des *mode collapse*: die Entstehung eines *generators*, der nur eine sehr kleine Anzahl möglicher Bilder erzeugen kann. *Mode collapse* entsteht genau dann, wenn der *generator* es zu schnell schafft, den *discriminator* von seinen Kreationen zu ›überzeugen‹. Einige frühe Beispiele, etwa die von Radford et al. (2015) erzeugten »imaginären Schlafzimmer«, zeigten jedoch bereits damals, wozu GANs potenziell in der Lage waren.

Die tatsächliche künstlerische Erforschung von GANs beginnt daher erst nach einer Reihe von technischen Verbesserungen (Goodfellow et al. 2016). Erste Experimente wie Mike Tykas *Portraits of Imaginary People* (2017) nutzten vor allem die DCGAN-Architektur (Radford et al. 2015). Um das Problem der niedrigen Auflösung, mit dem frühe GAN-Architekturen ebenfalls zu kämpfen hatten, zu umgehen, entwickelte Tyka einen Prozess, der *GAN-sampling* und *up-sampling*, also die künstliche Erhöhung der Bildauflösung, miteinander verband.

Inwiefern kann dieses Verfahren aber nun als subtraktives beschrieben werden? Schließlich beginnen beide GAN-Subsysteme, wie gesagt, ›bei null‹, also ohne analytische (*discriminator*) oder synthetische (*generator*) Fähigkeiten. Vergleichen wir den untrainierten *latent space* eines GANs mit einem weißen Blatt Papier oder einem schwarzen Bildschirm, so wird deutlich: Im Gegensatz zum ›leeren‹ Medium ist der *latent space* bereits mit Bildern gefüllt. Genauer: Jeder spezifische Punkt im *latent space* entspricht bereits einem spezifischen Bild, noch bevor der *discriminator* auch nur ein einziges ›reales‹ Bild ›gesehen‹ hat. Mit anderen Worten, das Medium selbst, die Architektur des neuronalen Netzes, birgt bereits das volle Potenzial eines scheinbar unendlichen Bildraumes. Allerdings haben die zu diesem Zeitpunkt im Bildraum verfügbaren Bilder keinerlei repräsentativen Charakter: Sie stellen nichts dar, verweisen weder auf ein allgemeines, universelles noch auf ein partikulares Objekt. Dies ändert indes nichts an der Tatsache, dass sie Bilder sind, die existieren und auf einfache Art und Weise zum Vorschein gebracht werden können.

Auch dieses ›Zum-Vorschein-Bringen‹ muss hier kurz technisch kontextualisiert werden. Der *generator* eines GANs (der oft schon allein für sich als GAN bezeichnet wird, obschon er, wie oben beschrieben, lediglich ein Subsystem desselben ist) erzeugt ein spezifisches Bild aus einem spezifischen Punkt im *latent space*. Im Falle eines 100-dimensionalen Vektorraumes kann dieser Punkt durch 100 einzelne Werte beschrieben werden. Wie in einem euklidischen, zweidimensionalen Koordinatensystem zwei Werte ausreichen, um einen exakten Punkt auf einer Fläche anzugeben, so definieren 100 Werte einen eindeutigen Punkt in einem 100-dimensionalen Raum. Gehen wir von einem normalisierten Raum aus, so liegen diese Werte zwischen –1 und 1. Dies bedeutet aber keineswegs, dass die Anzahl möglicher Bilder auf 3^{100} begrenzt ist. Denn der *latent space* eines GANs ist ein kontinuierlicher Raum, das heißt, Koordinaten werden als Gleitkommazahlen angegeben, in ihrer Präzision nur begrenzt durch den eingesetzten Variablentypus. Gehen wir von einem GAN mit einer Präzision von 32 Bit (*single-precision floating-point format*) und einem 100-

dimensionalen *latent space* aus, so können theoretisch $(3.4028235\times10^{38})^{100}$ Bilder erzeugt werden. Praktisch ist die Anzahl begrenzt durch die Art des Samplings, durch Besonderheiten diverser Architekturen und insbesondere durch die Effizienz des Trainings.

Was hier allerdings – völlig unabhängig von der genauen, technisch bedingten Anzahl möglicher Bilder – offensichtlich wird: Die Herausforderung bei der künstlerischen Arbeit mit GANs besteht nicht in der Bilderzeugung, sondern im Finden von Mitteln und Wegen, ›interessante‹ Punkte im *latent space* zu identifizieren. Aus einer Vielzahl von Möglichkeiten, aus einem Überfluss an Material werden konkrete Artefakte in einem solchen Maße herausgearbeitet, dass sie das Potenzial des Materials transzendieren. KI-Kunst, so könnte deshalb polemischer formuliert werden, ist Skulptur, nicht Plastik.

Der klassische Witz über den Bildhauer, der gefragt wird, woher er denn wisse, dass eine bestimmte Statue sich in einem bestimmten Marmorblock versteckt gehalten habe, zieht seine Pointe aus der Unsichtbarkeit der eigentlichen künstlerischen Arbeit. Freilich ist das Werk in gewisser Hinsicht im Material angelegt, aber herausgearbeitet werden kann es eben nur von wenigen künstlerisch und technisch ausgebildeten Menschen, die einen besonderen historischen, methodischen und ästhetischen Zugriff auf ihre Lebenswelt haben und über das Vermögen verfügen, diesen Zugriff in die Arbeit am Material zu übersetzen. In der Welt der GANs jedoch ›weiß‹ nur die Maschine selbst, was sich hinter einzelnen Punkten im *latent space* verbirgt. Im besten Falle existiert eine halbwegs kohärente räumliche Verteilung, die bestimmte Bildbestandteile bestimmten Punkten im *latent space* zuordnet (die Informatik spricht hier von *disentangled representations*). Selbst diese ist dem *latent space* jedoch in keinem Falle ›anzusehen‹, sondern nur experimentell ermittelbar.

Die *manuelle* Erkundung des *latent space* kann als Essenz der KI-Kunst zwischen circa 2015[4] und 2020 betrachtet werden. Werke aus dieser Zeit finden dabei unterschiedliche Ansätze, dessen unfassbare Ausmaße zu thematisieren. Neben dem einfachen kuratierten *sampling* setzte sich dabei der sogenannte *latent space walk* als Format durch. Dessen Reiz liegt in der Flüssigkeit der Bewegung: Bilder, die nahe beieinander liegenden Punkten im *latent space* zugeordnet sind, unterscheiden sich auch visuell nur minimal, sodass eine Inter-

4 An anderer Stelle (Offert 2022) habe ich die zentrale Rolle herausgearbeitet, die dem 2015 veröffentlichten DeepDream-Algorithmus beim Aufkommen der KI-Kunst zukam.

polation möglich wird, in der sich Einzelbilder organisch ineinander aufzulösen scheinen. *Latent space walks* wurden zentrale Werkzeuge im Repertoire von (mittlerweile etablierten) Künstlern wie Memo Akten und Mario Klingemann. Auch Anna Ridlers vielzitiertes Schlüsselwerk *Mosaic Virus* (2018) besteht aus einem solchen *latent space walk*, in dem semantisch eigentlich abgeschlossene Bildbestandteile verschwimmen, sich verschieben, aufteilen und verdoppeln.

Wie schon in der frühen Computerkunst geht es also die Ausstellung eines ästhetischen Systems als ästhetisches Objekt. Die ästhetische Spannung zwischen System und Objekt ist das erste wichtige Merkmal der KI-Kunst der GAN-Epoche. Das Format des *latent space walks* erfüllt jedoch darüber hinaus eine pragmatische Funktion, die auf das zweite wichtige Merkmal der KI-Kunst der GAN-Epoche verweist. *Latent space samples* sind in fast allen Fällen des künstlerischen Einsatzes von GANs *nicht* fotorealistische Bilder. Während populäre Architekturen wie StyleGAN 2 (Karras et al. 2019) für ihre nahezu fotorealistischen Ergebnisse angepriesen wurden, ist dieser Fotorealismus nicht haltbar, wenn kleinere, individuellere und weniger homogene Datensätze zum Einsatz kommen. Kann zum Beispiel ein von Nvidia im Rahmen der StyleGAN-2-Architektur veröffentlichtes Modell, das anhand eines besonders homogenen Datensatzes von Gesichtern[5] (Flickr-Faces FFHQ) trainiert wurde, auch nur diese, nämlich fotorealistische Gesichter erzeugen, bleiben künstlerische Anwendungen, so sie denn über die bloße Reproduktion und Kritik dieser und ähnlicher Modelle hinausgehen und eigene Datensätze verwenden, auf produktive Weise hinter diesem Fotorealismus zurück.

Philipp Schmitts und Steffen Weiss' Serie von GAN-generierten Stühlen (2018)[6] ist ein gutes Beispiel für KI-Kunst, die deren *glitch*-Aspekt bewusst betont. Erst signifikante Eingriffe in den generativen Prozess führen hier zu einem ästhetischen Artefakt, das als reales Objekt überhaupt bestehen kann. Sofia Crespo verbindet in ihrer Arbeit *Neural Zoo* (2018) handkuratierte Datensätze mit den traditionellen Techniken Collage, Pastiche und Cyanotypie. Durch die ›analoge Weiterverarbeitung‹ digitaler Bilder werden dabei nicht nur die klassischen ›Fehler‹ neuronaler Netzwerke wie eine zu große Anzahl hoher Bildfrequenzen (Geirhos et al. 2019) ausgeglichen, sondern auch beständige Artefakte generiert, die der Flüchtigkeit und Beliebigkeit der GAN-

5 Zur problematischen Verbindung von Gesichtserkennung und Künstlicher Intelligenz siehe Meyer (2021).

6 https://steffen-weiss.design/the-chair-project-generating-a-classic. Zugegriffen am 15. November 2022.

Bildermengen entgegenstehen. In den Arbeiten von Sarah Rosalena Brady schließlich, die GAN-generierte Bilder als maschinengewebte Tapisserien ausgeben lässt, löst sich der naturalistische Charakter synthetischer Bilder vollends in ein formales Prinzip auf. Bradys Werk *Untitled* (2020) zeigt dementsprechend GAN-halluzinierte Planeten. Aus der Übersetzung von Pixeln in Maschen entsteht so das Bild eines die *fabric of reality* webenden neuronalen Netzwerks.

Mit Aaron Hertzmann (2020) ist es die *visual indeterminacy* dieser Bilder, ihre bildinhaltliche Unschärfe, die einen ästhetischen Effekt hervorbringt. Mit Brecht (1957) könnte man gar von einem technischen Verfremdungseffekt sprechen, der in der KI-Kunst zum Einsatz kommt. Die ästhetische Rahmung unvermeidbarer technischer Artefakte schafft eine Distanz zwischen Rezipient und Werk, die dessen kritische Reflexion begünstigt. Dieser Verfremdungseffekt, der in letzter Instanz auf die kuratierte Unvollkommenheit der generierten Bilder zurückfällt, ist das zweite wichtige Merkmal der KI-Kunst dieser Epoche.

Transformer und die Automatisierung der künstlerischen Selektion

Die GAN-Epoche der KI-Kunst endet abrupt. Wie Helena Sarin, eine der prominentesten Vertreter*innen der GAN-basierten KI-Kunst, am 27. August 2022 auf Twitter schreibt:

> Come to think of it, this exhibit [die Online-Ausstellung *Reflections in the Water*[7]; Anm. F. O.] […] was a perfect closure for the GAN period of AI art; I mean there was some solid GAN art created after and maybe still is created but that was the end of the movement as we knew it, the thrill is gone.[8]

Die im September 2021 eröffnete Ausstellung, auf die der Tweet verweist, versammelt Künstler*innen wie Mario Klingemann, Anna Ridler, Helena Sarin, Jake Elwes und viele andere Vertreter*innen der GAN-Epoche. Dass sie in Sarins Tweet als Abschlusspunkt verstanden wird, verweist auf einen radikalen

7 https://feralfile.com/exhibitions/reflections-in-the-water-9ov. Zugegriffen am 15. November 2022.

8 https://twitter.com/NeuralBricolage/status/1563645089288753153. Zugegriffen am 15. November 2022.

Paradigmenwechsel in der KI-Forschung und – in der Folge – in der KI-Kunst, der bereits 2021 begann.

Im Januar 2021 veröffentlichte OpenAI, das wohl bekannteste unter den nichtakademischen KI-Forschungsinstituten, sowohl eine neuartige Architektur als auch ein fertiges Modell namens CLIP (Radford et al. 2021). CLIP (die Abkürzung steht für *contrastive language-image pre-training*) verfolgt einen multimodalen Ansatz, das heißt, CLIP-Modelle bringen Text und Bilder in Zusammenhang. Möglich wird dies durch die Analyse von Bildern im Kontext, wie sie das Internet in hoher Anzahl bereithält: Jedes Bild, das auf einer Website erscheint, ist gemeinhin von Text umgeben, der in direktem Zusammenhang mit dem Bild steht. CLIP produziert aus diesen Daten einen gemeinsamen *latent space* für Text und Bild sowie entsprechende *encoder*, die ungesehene Bilder und ungelesenen Text in Punkte in diesem *latent space* übersetzen können. Eine konkrete praktische Anwendung ist dementsprechend die Herstellung von Bildbeschreibungen: ›Zeigt‹ man CLIP ein Bild, so ist es in der Lage, die dem Bild am besten entsprechenden Textpunkte im *latent space* auszugeben. CLIP hat jedoch selbst keine Bildsynthesefähigkeiten, sodass der umgekehrte Weg zunächst verschlossen blieb.

Dennoch kann das CLIP-Modell zur Bildsynthese eingesetzt werden, wenn man es an ein existierendes generatives Modell ›ankoppelt‹. Genauer: wenn man, statt den *latent space* eines Modells auf zufälligem Wege zu durchlaufen, an jedem Punkt prüft, ob das vom generativen Modell produzierte und von CLIP klassifizierte Bild näher an die angepeilten Textpunkte im *latent space* von CLIP gerückt ist. In den verbleibenden Monaten des Jahres 2021 konnte sich auf der Basis dieser einfachen Idee eine höchst vitale ›Szene‹ etablieren, die zahlreiche Varianten und Verbesserungen der durch CLIP gesteuerten Bildsynthese auf den Weg brachte.[9] Programmierer:innen und Künstler*innen wie Katherine Crowson und Ryan Murdoch veröffentlichten ihre Programme im Google CoLab-Format, das heißt als im Browser ausführbare, cloud-basierte und interaktive Skripte. Betrachtet man die Ergebnisse dieser vielfältigen Ansätze, wird offensichtlich, dass der Einsatz von CLIP jenes grundsätzliche Problem der vorhergehenden Epoche der KI-Kunst löst, das für den künstlerischen Einsatz zentral ist: Wie können visuell interessante Punkte im *latent*

9 Siehe https://ml.berkeley.edu/blog/posts/clip-art/ (zugegriffen: 15. November 2022) für einen gewissenhaften Überblick und Underwood (2021) für eine weiter gefasste kulturelle Einordnung des Prinzips latent space.

space gezielt angesteuert werden? Mit CLIP funktioniert dies gewissermaßen ›auf Zuruf‹, durch die gezielte Manipulation des eingegebenen Textes.

Den technischen Durchbruch brachte dabei die Kombination von CLIP mit sogenannten *diffusion models*, einer Klasse generativer neuronaler Netzwerke, die das *generator-discriminator*-Prinzip der GAN-Architektur durch das Prinzip der gelernten Kompression ersetzt. Im Frühjahr 2022 veröffentlichten OpenAI und Google eine ganze Reihe von Modellen, die dieses bisher nur als künstlerisch-technisches Experiment vorliegende Prinzip aufnahmen und kommerzialisierten. Beginnend mit DALL-E 2, das durch eine umfassende PR-Kampagne und künstliche Verfügbarkeitsbeschränkungen[10] einer weiten Öffentlichkeit bekannt gemacht wurde, begann die KI-gestützte Bildsynthese, die lange gerade mit der *Unvollkommenheit* synthetisierter Bilder identifiziert wurde, als Möglichkeit der fotorealistischen Bilderzeugung wahrgenommen zu werden. Den vorläufigen Abschluss – zum Zeitpunkt der letzten Revision dieses Textes – bildet Stable Diffusion[11] (Rombach et al. 2022), ein von einem großen Kollektiv von Forschenden und Akteur*innen der KI-Industrie veröffentlichtes Open-Source-Modell, das es im Hinblick auf die Qualität der synthetisierten Bilder durchaus mit DALL-E 2 aufnehmen kann.

Der Fokus der KI-Kunst verschiebt sich im Kontext dieses Paradigmenwechsels hin zu multimodalen Modellen, radikal: *prompt engineering*, also die gezielte Komposition von Texteingaben zur Erzeugung ganz bestimmter Bildinhalte, ersetzt die formalen Experimente der GAN-Epoche. Dennoch sind es keinesfalls nur semantische Aspekte des Bildes, die sich über *prompts* gezielt beeinflussen lassen. Durch die Einstreuung von bestimmten Stichwörtern lassen sich ebenso bestimmte ›Stile‹ provozieren. Die Verwendung des Stichworts ›Unreal Engine‹ sorgt für einen hyperrealistischen Stil, der sich an bekannten 3-D-Programmen (wie eben auch Unreal, eine von Epic Games entwickelte Computerspiel-Engine) orientiert. Das Stichwort ›trending on ArtStation‹ sorgt für einen ›populären‹ Stil, wie er von vielen Hobbykünstler:innen auf der

10 Das Web-Interface für DALL-E 2 (das Modell selbst bleibt weiterhin proprietär) war lange Zeit nur auf Anfrage und für Forschende und Akteure der KI-Industrie zugänglich. Im August 2022 wurde es schließlich auch einer breiteren Öffentlichkeit zur Verfügung gestellt und gleichzeitig kommerzialisiert: Für jedes generierte Bild muss ein bestimmter Betrag entrichtet werden.

11 Der folgende Twitter-Thread bietet einen guten Überblick über die Funktionsweise von Stable Diffusion: https://twitter.com/ai__pub/status/1561362542487695360. Zugegriffen: 15. November 2022.

Website ArtStation gepflegt wird. Die Angabe von bestimmten analogen Filmtypen (›Provia, Velvia, Fujifilm Superia‹) produziert fotorealistische Bilder, deren Spektrum das Profil der genannten Medien erstaunlich genau abbildet.

Damit kehren wir zurück zur Ausgangsfrage dieses Aufsatzes: Wie sind die konkreten Techniken der KI-Kunst ästhetisch einzuordnen? Mit CLIP, so könnte man sagen, kommt das der KI-Kunst immer schon inhärente skulpturale Moment zu sich selbst. *Latent spaces* können nun gezielt befragt, statt bloß passiv durchschritten werden. Damit tritt aber auch die KI-Kunst in eine neue Phase ein und entfernt sich weiter von traditionellen algorithmisch-künstlerischen Methoden. Statt – wie die frühe Computerkunst – platonische Grundformen durchzuexerzieren, erzeugt die KI-Kunst post CLIP gezielt Neues aus *found footage:* visuelle *musique concrète.* Die Skulpturen, die sie schafft, sind Monumente für das Internet, die Mutter aller *datasets*.

Literatur

Bogost, Ian. 2019. The AI-Art Gold Rush Is Here. *The Atlantic*. https://www.theatlantic.com/technology/archive/2019/03/ai-created-art-invades-chelsea-gallery-scene/584134/. Zugegriffen am 1. Feburar 2023.

Brecht, Bertolt. 1957. Die Straßenszene. Grundmodell einer Szene des epischen Theaters. In *Schriften zum Theater*. Frankfurt a.M.: Suhrkamp.

Dhariwal, Prafulla und Alex Nichol. 2021. Diffusion Models Beat GANs on Image Synthesis. https://doi.org/10.48550/arXiv.2105.05233.

Esser, Patrick, Robin Rombach und Björn Ommer. 2021. Taming Transformers for High-Resolution Image Synthesis. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/CVPR46437.2021.01268.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville und Yoshua Bengio. 2014. Generative Adversarial Nets. *Advances in Neural Information Processing Systems*. https://dl.acm.org/doi/10.5555/2969033.2969125.

Harvey, Adam und Jules LaPlace. 2021. Researchers Gone Wild: Origins and Endpoints of Image Training Datasets Created »In the Wild«. In *Practicing Sovereignty: Digital Involvement in Times of Crisis,* Hg. von Bianca Herlo, Daniel Irrgang, Gesche Jost und Andreas Unteidig, 289–310. Bielefeld: transcript.

Karras, Tero, Samuli Laine und Timo Aila. 2018. A Style-Based Generator Architecture for Generative Adversarial Networks. https://doi.org/10.48550/arXiv.1812.04948.

Karras, Tero, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen und Timo Aila. 2019. Analyzing and Improving the Image Quality of StyleGAN. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/CVPR42600.2020.00813.

Meyer, Roland. 2021. *Gesichtserkennung*. Berlin: Wagenbach.

Nake, Frieder. 1971. There Should Be No Computer Art. *Bulletin of the Computer Arts Society*: 18–19. https://dam.org/museum/essays_ui/essays/there-should-be-no-computer-art/. Zugegriffen: 15. November 2022.

Nake, Frieder. 1974. *Ästhetik als Informationsverarbeitung. Grundlagen und Anwendung der Informatik im Bereich ästhetischer Produktion und Kritik*. Wien und New York: Springer.

Nake, Frieder. 2021. The Art of Being Precise. Interview mit Margit Rosen. https://www.youtube.com/watch?v=Z_pOiHX6HYE. Zugegriffen: 15. November 2022.

Offert, Fabian. 2019. The Past, Present, and Future of AI Art. *The Gradient*. https://thegradient.pub/the-past-present-and-future-of-ai-art/. Zugegriffen: 15. November 2022.

Offert, Fabian. 2023. KI und/als bildende Kunst. In *Handbuch Künstliche Intelligenz und die Künste*, Hg. Stephanie Catani und Jasmin Pfeiffer. Berlin: De Gruyter.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry et al. 2021. Learning Transferable Visual Models from Natural Language Supervision. https://doi.org/10.48550/arXiv.2103.00020.

Radford, Alec, Luke Metz und Soumith Chintala. 2015. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. https://doi.org/10.48550/arXiv.1511.06434.

Rombach, Robin, Andreas Blattmann, Dominik Lorenz, Patrick Esser und Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. https://doi.org/10.48550/arXiv.2112.10752.

Underwood, Ted. 2021. Science Fiction Hasn't Prepared Us to Imagine Machine Learning. Blog: *The Stone and the Shell*. https://tedunderwood.com/2021/02/02/why-sf-hasnt-prepared-us-to-imagine-machine-learning/. Zugegriffen: 15. November 2022.

Repräsentation, Kritik und Anlass

Eine Trichotomie der künstlerischen Nutzungsaspekte von KI

Michael Klipphahn-Karge

Abstract: *In diesem kunsttheoretischen Beitrag widme ich mich der Untersuchung des Verhältnisses von künstlerischen Werken der Gegenwart und Systemen Künstlicher Intelligenz (KI). Ich folge im ersten Teil der Untersuchung der These, dass besonders drei Einzelaspekte prominent für die Nutzung von KI durch die Kunst sind und eine unfeste Trichotomie bilden. Diese Aspekte sind: (1) die technikvermittelnde Repräsentation von KI-Systemen durch deren technische Nutzung im Werk beziehungsweise zur Erstellung einer künstlerischen Arbeit; (2) die Kritik an KI durch eine KI-Technologie nutzende künstlerische Arbeit; (3) das Begreifen von KI als narratives Element und als Anlass für die Erschaffung eines Kunstwerks ohne den Gebrauch von KI-Technologie. Ich exemplifiziere meine Analyse im zweiten Teil des Beitrags an einem Werk der im Kunstkontext oftmals rezipierten Rechercheagentur Forensic Architecture und an Arbeiten der Künstler:innen Nora Al-Badri, Pierre Huyghe und Anna Ridler. Dabei stelle ich jeweils einen der drei benannten Aspekte als prädestinierend für die werkinhärente Bezugnahme auf KI scharf. Die thematische Klammer meiner Werkauswahl ist eine vergleichbare Terminologie von KI: Alle Arbeiten bedienen sich formal oder metaphorisch maschineller oder gar tiefer maschineller Lernverfahren und der diesbezüglichen Arbeit künstlicher neuronaler Netze.*

Einleitung

Systeme Künstlicher Intelligenz (KI) überfluten derzeit die ›Kunstwelt‹ (vgl. Becker 2008 [1982]). Derartige technische Neuerungen beeinflussen sowohl das Kommunikations- und Konsumverhalten wie auch die Sozialisationsprozesse innerhalb kunstproduzierender Netzwerke. Parallel wird KI durch

Künstler:innen in die Herstellung von Kunstwerken integriert. Beispiele für die Auseinandersetzung mit KI-Prozessen und den damit einhergehenden Umbrüchen finden sich in zahlreichen Gruppenausstellungen, die zum Großthema ›KI in der Kunst‹ jüngst realisiert worden sind. So zeigte das Barbican Centre in London im Jahr 2019 die interaktive Ausstellung *AI: More than Human*, mittels derer versucht wurde, den derzeit viel diskutierten Bogen von der Kunst zur Wissenschaft zu schlagen (vgl. Bippus 2010), indem Forschungsprojekte von Wissenschaftler:innen integriert und mit künstlerischen Arbeiten in einen Dialog gebracht wurden. Nahezu zeitgleich widmete sich das Museum für angewandte Kunst in Wien im Jahr 2019 mit der Schau *Uncanny Values. Künstliche Intelligenz & Du* künstlerischen Reflexionen zu aktuellen Anwendungsgebieten von KI. Das de Young Museum in San Francisco konzentrierte sich mit der Ausstellung *Uncanny Valley: Being Human in the Age of AI* im Jahr 2020 auf Kunstwerke, die sich mit der künftigen Schnittstelle zwischen Mensch und Maschine auseinandersetzen. Einzelpräsentationen und Interventionen in Museen und anderen Institutionen von Künstler:innen wie Memo Akten, Gene Kogan oder Helena Sarin, die mindestens anteilig auf technischem Code beruhende, technisch responsive und/oder mithilfe von KI erschaffene Elemente enthielten, wurden jüngst ebenso realisiert.

Gemein ist dem Großteil der Werke in diesen Ausstellungen, dass die Darstellung und Interpretation materieller Realitäten in Form von Daten in die Entstehung von Kunst Einzug findet – das heißt in diese eingreift und sie umbildet, sodass die Materialität der Werke vielmals durch technologische Komponenten bestimmt wird (vgl. Scorzin 2021b). Dem liegt das Verständnis zugrunde, dass KI als Nutzung von Verfahrensweisen maschineller Automatisierung im Sinne der Nachbildung kognitiver Prozesse begriffen werden kann (vgl. Manovich 2018: 4f.). Der Begriff KI bezeichnet in diesem Zusammenhang datenbasierte Systeme und wird von mir in diesem Beitrag folglich auf die Nutzung, Anwendung von oder Beschäftigung mit maschinellen und tiefen maschinellen Lernverfahren begrenzt.

Solche Verfahren, die historisch gesehen aus regelbasierten Wiederholungen ›erlernter‹ Vorgänge erwuchsen, beeinflussen heute Großteile der Kunst, indem beispielsweise Bilder synthetisiert, künstlerische Vorgänge ›lernend‹ nachgeahmt, additiv-schichtende und automatisierte Druckverfahren gesteuert oder Materialien simuliert werden. Diese künstlerischen Arbeitsweisen haben sich seit den späten 1960er-Jahren aus kybernetischen und bildgenerierenden Ansätzen in der Computerkunst entwickelt (vgl. Bense 1965; Nake 1966,

1974; retrospektiv auch Flusser 1985, 1993).[1] Bereits in dieser Zeit wurde Computerprogrammen Variabilität bei der Generation von Bildern bescheinigt und es wurden in ersten Feldversuchen synthetische Bilder erstellt, indem Künstler:innen systematisch mit Verfahren experimentierten, die mittels des – zunächst noch händisch ausgeführten – Einsatzes von Algorithmen zeichnerische und malerische Techniken nachahmten und entsprechende Bilder erzeugen konnten. Beispiele finden sich bei den Künstler:innen Harold Cohen, Vera Molnár, Frieder Nake, Georg Ness oder Karl Sims. Die erhöhte Verfügbarkeit von Rechenleistung führte dazu, dass sich das Spektrum der für Künstler:innen verfügbaren digitalen Technologien deutlich erweiterte, wodurch es ab den 1980er-Jahren vermehrt zum experimentellen Einsatz von KI in der Kunst kam (vgl. Manovich 2002: 567ff.).

Vielmals werden die seither ausgemachten Einschnitte durch KI oder verwandte Technologien als instruierend begriffen für »[d]igitale Kunstproduktionen, die mithilfe von [maschinellen Lernverfahren und] intelligenten Algorithmen hergestellt werden, […] [und] als Aktualisierungen eines größeren computerisierten Netzwerkes« (Scorzin 2021a: 48) verstanden. Der Umgang mit Daten ist demnach bestimmend für den ästhetischen Output, also das formale künstlerische Werk – ein Konstruktionsmerkmal, das sich noch heute in sogenannter ›KI-Kunst‹ reproduziert und Einfluss auf Kontext, Inhalt und Rezeption nimmt (vgl. ebd.: 66ff.).

Zum einen wird KI also vielfältig genutzt, zum anderen wird diese Vielfältigkeit der Nutzungsaspekte kunst- und bildtheoretisch eher gruppierend reflektiert. So wird KI erstens verengt als kreierendes Element innerhalb des Autor:innendiskurses beleuchtet und in ein dualistisches Verhältnis mit Künstler:innen gesetzt, um Fragen von Originalität und Authentizität auszutarieren (vgl. Schröter 2021). Zweitens wird KI mit Blick auf Mechanismen des Kunstmarktes besprochen. In diesem Kontext werden allerdings Fragen von Wertsteigerungsprozessen mit künstlerischem Schöpfungsimpetus und Qualitätsanspruch sowie der Befähigung zu künstlerischen Fertigungsweisen

1 Derzeit nimmt das Interesse an generativer Kunst wieder stark zu. Grund dafür sind barrierearme Programme wie das von der Non-Profit-Organisation OpenAI Inc. entwickelte DALL·E, das Bilder aus Textbeschreibungen erstellen kann. Mittels maschineller Lernverfahren, die auf der Arbeit mit künstlichen neuronalen Netzen fußen, werden Wörter als Eingabe genutzt, um Anordnungen von Pixeln als visuelle fotorealistische Ausgabe zu generieren. Damit sind diese Prozesse nach Roland Meyer (2022: 51) eher mit »einer Suchanfrage […] als einem Produktionsvorgang« vergleichbar.

vermischt, indem »auf die Parallele« hingewiesen wird, »die es zwischen dem Programmieren eines Algorithmus gibt und der Expertise, die das Handwerk und den Stil eines Künstlers ausmachen« (Caselles-Dupré 2018). Damit wird zur Mystifizierung jedweden bildgebenden Prozesses als kunstgenerierend beigetragen. Und drittens wird Kunst, die in irgendeiner Weise mit KI zusammenhängt, schlicht unter dem Sammelbegriff KI-Kunst gefasst und simplifizierend mit »neuartige[n] Kulturästhetiken und algorithmisierte[n] Formensprachen« beschrieben (Scorzin 2021b: 60).

1. Analyserahmen

Nutzungsaspekte

Gegebenheiten wie letztere sind der Ausgangspunkt meiner Überlegung, innerhalb der unter dem dritten Punkt gefassten Gruppierung von KI-Kunst Tiefenschichten freizulegen. Denn die bisherige Analyse von KI-Kunst blendet Unterscheidungen aus, die sich zwischen verschiedenen maschinellen Ästhetiken zeigen und die nicht unmittelbar mit bloßen Rechenregularien verschränkt oder begründbar sind.

Einen Vorschlag, das äußerst diverse Korpus von Kunst zu fassen, die mit maschinellen Lernverfahren produziert, von diesen Prozessen beeinflusst oder inspiriert wurde, und die so subsumierten Kunstwerke grundlegend zu systematisieren, formuliert und durchdenkt Sofian Audry (2021). Audry argumentiert trichotom für eine stark auf der Kenntnis technischer Verfahren beruhende Reflexion, die sich auf die verschiedenen aufeinander aufbauenden Komponenten bezieht, die maschinelle Lernsysteme charakterisieren: die Evaluation des Trainingsprozesses, die dahingehende Auswahl des passenden KI-Modells und die Datenabhängigkeit des maschinellen Lernens für das künstlerische Schaffen. Der damit vereinbare Ansatz von Francis Hunger (2019) forciert ebenso eine stärker technologiereflexive Unterscheidung: Statt Ästhetiken, die auf der Anwendung und/oder Beschäftigung mit datenbasierten Systemen fußen, lose mit einem Attribut wie ›algorithmisch‹ zu verschlagworten, plädiert Hunger für eine stärkere Differenzierung und damit auch für eine Untersuchung, die etwa von der Beschaffenheit der Datenbasis und der statistischen Verarbeitung ebendieser, vom Informationsmodell oder von der Art der Datenabfrage ausgeht.

In Übereinstimmung mit diesen stark auf technologische Aspekte fokussierten Unterscheidungen werde ich die Untersuchung der Möglichkeitshorizonte von KI-Kunst um eine kunstwissenschaftliche und bildtheoretische Sichtweise ergänzen und dabei der Frage nachgehen, was das Kunstwerk mit KI macht. Indem ich stärker vom Kunstwerk als von der verwendeten Technologie her denke, zielen meine Ausführungen auf eine Konkretisierung bestehender Kategorisierungen innerhalb der Gattungszuschreibung KI-Kunst.

Bisher wird KI-Kunst mehrheitlich als changierend zwischen »Tool […] [und] Thema« (Scorzin 2021a: 48) oder zwischen »explorativ-experimentell und affirmativ bis reflexiv und kritisch« (Scorzin 2021b: 58) verschlagwortet, ohne exakt zu benennen, was das meint und welche Arbeiten aus welchem Grund unter welchen zentralen Aspekten betrachtet und untersucht werden sollten. Vereinzelt lassen sich in der Forschung bereits rubrizierende Ansätze und Impulse ausmachen, die innerhalb des Feldes der KI-Kunst Unterscheidungen benennen, beispielsweise hinsichtlich der ›Authentizität‹ der geäußerten Kritik an KI als Aspekt der Verwendung von KI-Technologie in Kunst (vgl. Bajohr 2021). Um diese Tendenz zu bestärken, schlage ich eine analytische Rahmung vor, die verschiedene Bezugnahmen von Kunst auf KI kategorisiert und so helfen kann, KI-bezogene Kunst genauer zu bestimmen. Meine These ist, dass besonders drei Einzelaspekte prominent für den Einsatz von KI durch die Kunst/die Künstler:innen sind und eine unfeste Trichotomie bilden. Diese Aspekte sind: (1) die technikvermittelnde Repräsentation von KI-Systemen durch deren technische Nutzung im Werk beziehungsweise zur Erstellung einer künstlerischen Arbeit; (2) die Kritik an KI durch eine KI-Technologie nutzende künstlerischen Arbeit; (3) das Begreifen von KI als narratives Element und als Anlass für die Erschaffung eines Kunstwerks ohne den Gebrauch von KI-Technologie.

Ich exemplifiziere meine Untersuchung an einem Werk der im Kunstkontext oftmals rezipierten Rechercheagentur Forensic Architecture und an Arbeiten der Künstler:innen Nora Al-Badri, Pierre Huyghe und Anna Ridler. Dabei stelle ich jeweils einen der drei benannten Aspekte als prädestinierend für die werkinhärente Bezugnahme auf KI scharf und verweise auf Brüche zu und Überschneidungen mit den jeweils anderen Nutzungsaspekten. Die thematische Klammer meiner Werkauswahl ist eine vergleichbare Terminologie von KI: Alle Arbeiten bedienen sich – entsprechend der eingangs vorgenommenen Definition – formal oder inhaltlich maschineller oder gar tiefer maschineller Lernverfahren. Folglich wird evident, dass formale und in-

haltliche Aspekte in KI-Kunst oftmals nicht exakt abzugrenzen sind und dass sich inhaltliche Schwerpunktsetzungen naturgemäß aus formalen Entscheidungen im Rahmen technischer Möglichkeiten ergeben. In der gleichen Weise ist zutreffend, dass der Werkinhalt den formalen Einsatz von KI-Technologie bestimmen kann.

Ich betrachte die Untersuchung entlang der von mir vorgeschlagenen Nutzungsaspekte des Weiteren als unfest, das meint als beweglich, erweiterbar und fließend. Damit gebe ich einer ontologisch begründeten Skepsis gegenüber allzu generalisierenden Periodisierungs- und Ordnungsmodellen von Kunstwerken Raum und zeige gleichzeitig auf, wie sich eine unfeste Trichotomie im Umgang mit KI durch die Kunst als produktives Analysewerkzeug manifestieren kann: Der von mir vorgeschlagene Analyserahmen kann helfen, die Nutzung von und die Auseinandersetzung mit KI-Technologien durch Künstler:innen theoretisch besser rückbinden und reflektieren zu können sowie diese differenzierter zu benennen. Denn generell schließe ich mich der Auffassung an, dass »symbolische Orientierungs- und Ordnungsversuche, die im Vertrauen auf die Leistungsfähigkeit systematischen und analytischen Denkens unternommen werden« (Locher 2010 [2001]: 14), essenziell für eine nuancierte Betrachtung technischer Errungenschaften sind – besonders, da in künstlerischen Werken permanent impulsgebende Innovativen des Technischen aufgegriffen werden. Ihr Einbezug beschränkt sich nicht nur auf den künstlerischen Umgang mit Material oder auf die Verwendung von Technologien als Werkzeugen, sondern kann auch in der metaphorischen Reflexion derartiger Innovationen im Werk bestehen – etwa, indem KI als Inspirationsquelle, Verrätselungselement oder Reibungsfläche fungiert.

Trichotomie

Zu Beginn möchte ich die Trichotomie der Nutzungsaspekte konkreter fassen und ihr einige Beispiele zuordnen (siehe Abb. 1).

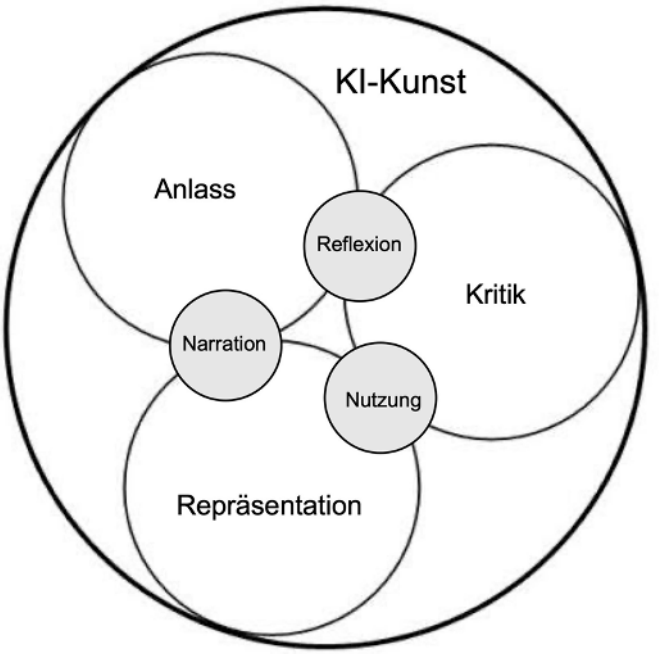
(1) Repräsentation: Verfahren Künstlicher Intelligenz können als technische Elemente in einem Kunstwerk eingesetzt sowie modellierend für das Werk genutzt werden, womit derartige Werke gleichsam repräsentativ für KI sind. Diese Form der Nutzung von KI scheint mir die derzeit prominenteste zu sein (vgl. Bogost 2019). Entsprechende Kunstwerke zeichnen sich durch eine Nutzung von KI aus und parallel findet sich KI in ihnen häufig als narratives Element – das heißt, durch die Beschäftigung mit ihnen erfährt man etwas

über KI. Beispiele finden sich etwa in den automatisierten Zeichenmaschinen von Sougwen Chung (*Drawing Operations*, 2015), der auf generativen algorithmischen Bildgebungsverfahren beruhenden digitalen Malerei des Kollektivs Obvious (*Edmond de Belamy*, 2018) oder in Refik Anadols Werkkomplexen zur optischen Simulation von Wahrscheinlichkeitsvariationen (*Quantum Memories*, seit 2020), die mittels Verfahren von KI und Quantum Computing auf Basis der Google-AI-Quantum-Supremacy-Experimente erstellt sind. Seine Werke beschreibt der Künstler unter anderem als Datenskulpturen (vgl. Anadol/Scorzin 2021: 164).[2] Die Anforderungen, die das Herstellen künstlerischer Arbeiten mit sich bringt, wirken dabei modellierend auf KI-Technologien ein, da diese in diesem Zusammenhang auch Werkzeuge der Künstler:innen sind. Gleichzeitig sind die Künstler:innen den technischen Möglichkeiten der genutzten maschinellen Verfahren unterworfen.

(2) Kritik: Künstler:innen nutzen KI-Systeme, um beispielsweise die damit einhergehenden Schwierigkeiten – wie die mit digitalen Technologien verbundenen sozialen Exklusions- und Marginalisierungsproblematiken – zu reflektieren und – ganz generell – KI-Technologien zu kritisieren. Das wird exemplifiziert in Arbeiten von Künstler:innen wie Mario Klingemann, etwa dessen gesteuerten Fehlinterpretationen von Eingabedaten und dem daraus resultierenden Bild autonomer Kreativität, das absichtlich veruneindeutigt ist (*Neural Glitch/Mistaken Identity*, seit 2018). Weitere Beispiele finden sich in Trevor Paglens Werk. Besonders prominent ist seine anhaltende Kritik an automatisierten Bildgenerierungswerkzeugen und an Gesichtserkennungssoftware sowie an zugrunde liegenden Bilddatenbanken. Seine Kritik zeigt er visuell anhand von maschinellen Imaginationsprozessen auf, die durch die Reaktion eines automatisierten bilderkennenden auf ein bildgenerierendes KI-System entstehen (*Vampire [Corpus: Monsters of Capitalism*]. Adversarially Evolved Hallucination, 2017). Zu nennen sind ferner Hito Steyerls Werke zu prädiktiven Analyseverfahren (*Power Plants*, 2019), die »Künstliche Dummheit [als] […] real existierende Erscheinung von Künstlicher Intelligenz« offenlegen sollen (Steyerl, o. D., zitiert nach Karrasch 2019).

2 Der Frage, ob generative Daten in Relation zu Kunst additiv oder subtraktiv wirken beziehungsweise additiv oder subtraktiv auf sie eingewirkt wird, ob also ›KI-Kunst‹ in diesem Sinne plastisch oder skulptural hergestellt wird, geht Fabian Offert in seinem Beitrag in diesem Band nach.

Abb. 1: Schematisierung der trichotomen Nutzungsaspekte von KI-Kunst.



(3) Anlass: Systeme Künstlicher Intelligenz können narratives Element und damit Anlässe künstlerischer Arbeit sein, ohne dass eine Einbindung maschineller Lernverfahren als Technologie Teil der Werkgenese ist. Rubriziert werden unter diesem dritten Aspekt also die zahlreichen Kunstwerke, die etwas über KI erzählen, ohne KI technisch zu nutzen. Dieser Aspekt lässt sich – entgegen den Aspekten der Repräsentation und der Kritik – noch einmal unterteilen. Denn es finden sich sowohl Belege für ein eher affirmatives als auch für ein aversives KI-Narrativ. KI dient in entsprechenden Werken also sowohl als Anlass für eine kritische Reflexion des Technischen – beispielsweise eine dystopische oder hermeneutische – als auch für das Bedienen eines technikvermittelnden Narrativs – etwa eines utopischen oder positivistischen.

Zeugnisse für eher bejahende Sichtweisen finden sich unter anderem bei Cecile B. Evans, die in ihren HD-Videos (*Hyperlinks or It Didn't Happen*, 2014) über Möglichkeiten der Konservierung des Selbst in künstlichen Superintelligenzen spekuliert, oder in Pierre Huyghes Arbeiten zu Schwärmen, Schwarmver-

halten und seiner Übertragung auf Funktionen künstlicher neuronaler Netze, die Huyghe durch Bienenstämme verkörpert sieht (*Exomind [Deep Water]*, 2017/2020). Ein ähnliches Anliegen verfolgt Agnieszka Kurant mit ihren Arbeiten über Termiten (*A.A.I.*, seit 2016), wobei sich in diesen Werken über kollektive und nichtmenschliche Intelligenzen durchaus kritische Anspielungen auf Ausbeutungspraktiken von Sozialkapital, auf den Überwachungskapitalismus und auf die Zukunft der Arbeit finden lassen, wodurch sie gleichsam eine Brücke zu einem ablehnenden, kritischen beziehungsweise warnenden Zugang zu KI als Anlass künstlerischer Werke bauen. Beispiele hierfür finden sich in Agnieszka Polskas filmischer Reflexion über protokapitalistische Systeme des 15. Jahrhunderts, die die Künstlerin an einer Krakauer Salzmiene exemplifiziert. In diesem Film definiert ein Dämon die Kapitalisierung von Arbeit als Basis von KI und KI wiederum als eine Ursache von Ausbeutungspraktiken und Umweltschäden in der Zukunft (*The Demon's Brain*, 2018). Andere Beispiele finden sich in Zach Blas' künstlerischem Interventionspool mit verschiedenen technischen Geräten, die einem Queering digitaler Technologien dienlich sind (*Queer Technologies*, 2008–2012) – darunter etwa ein Genderadapter oder ein Manifest, das eine Anleitung für vernetzten Aktivismus von queeren Personen enthält. Ein weiteres Exempel ist Simone C. Niquilles/Technofleshs dystopische Videoversion einer Zukunft, in der die Emotionserkennung durch maschinelle Lernverfahren zum Berufsalltag gehört (*Elephant Juice*, 2020).

Gemeinhin schließen die Schwerpunktsetzungen dieser unterschiedlichen Nutzungsaspekte einander nicht aus. Sie sind unfest und teilweise durch gegenseitige Bezugnahmen, aber auch Ausschlüsse gekennzeichnet, die ihre Grenzen durchlässig machen: Zusammenfassend überschneiden sich repräsentative, durch Technik vermittelte und technikvermittelnde sowie oft unkritische Kunstwerke mit künstlerischen Arbeiten, die kritisch gegenüber KI argumentieren. Denn in beiden Gruppen werden maschinelle Lernverfahren verwendet und das künstlich generierte visuelle Artefakt wird jeweils überbetont; sie sind also bezüglich der technischen Inkorporation von KI ähnlich gelagert. Der Unterschied liegt in der Reflexion von KI-Technologie innerhalb der jeweiligen Werke. Werke, die vornehmlich zur ersten Gruppe gehören, spiegeln KI unabhängig von Kritik an KI beziehungsweise repräsentieren KI-Technologie im Kunstdiskurs. Künstlerische Arbeiten, die vornehmlich der zweiten Gruppe zugeordnet werden können, reflektieren KI durch deren technische Nutzung hingegen kritisch. Eine Überschneidung bezüglich dieser kritischen Haltung zu KI teilt sich technologiereflexive Kunst wiederum mit einer dritten Gruppe von Kunstwerken, in denen KI zum An-

lass genommen wird, um künstlerisch tätig zu werden. In ihrer narrativen Ausrichtung kann dieser Aspekt KI-bezogener Kunst partiell auch eng mit jener Kunst verschränkt sein, die KI schlicht repräsentiert.

Im Folgenden expliziere ich die trichotomen Aspekte an vier Werkbeispielen.

2. Werkbeispiele

Repräsentation: Forensic Architecture, *Triple Chaser*, 2018/19

Das erste Beispiel, in das ich tiefer einsteigen werde, ist das in den Jahren 2018 und 2019 entstandene Werk *Triple Chaser* von Forensic Architecture. Diese – an der Goldsmith University of London angesiedelte – Rechercheagentur ist ein multidisziplinäres Team, das aus IT- und bildforensisch arbeitenden Aktivist:innen, Softwareentwickler:innen, Archäolog:innen, Künstler:innen, Journalist:innen und Jurist:innen besteht. Ihre Arbeit *Triple Chaser* repräsentiert die technischen Möglichkeiten von KI qua Nutzung, ohne diese Nutzung kritisch zu reflektieren beziehungsweise im Werk zu thematisieren.[3] Das Kollektiv Forensic Architecture entwickelt in seinem multimedialen und sich vornehmlich im digitalen Raum situierenden Werk Modelle zur visuellen Aggregation von Daten, mit denen sich unter anderem im juridischen Bereich IT-forensisch argumentieren lässt.

Die dafür eingesetzten Technologien aus dem Feld der Computer Vison basieren auf dem Training von Algorithmen.[4] Entsprechende KI-Modelle werden ausgebildet, um bestimmte Typen eines Objekts anhand eines eingespeisten Bildpools digital identifizieren zu können (vgl. Forensic Architecture 2020a). Auf der Grundlage dieser Befundmasse lernen automatisierte Rechenvorschriften innerhalb eines latenten Raums Schlussfolgerungen zu treffen und zu Ergebnissen in der Kennung von Objekten zu gelangen. Bilder werden

3 Dass die Arbeiten von Forensic Architecture in großen Teilen als Kunst aufgefasst werden, liegt in der Rezeption des Arbeitsoutputs dieses Kollektivs begründet. So wurde *Triple Chaser* beispielsweise 2020/2021 in der Ausstellung *Uncanny Valley* im de Young Museum in San Francisco als Teil des Werkkomplexes *Model Zoo* (seit 2020) gezeigt.

4 ›Training‹ meint in diesem Fall, dass ein KI-System mittels einer Differenzierung von korrekten und falschen Entscheidungsweisen ›lernt‹ und so in den Stand versetzt wird, Unterscheidungsparameter weiter auszudifferenzieren.

also als Daten wie auch als Verarbeitungs- und Analysekategorien genutzt, um Bildinhalte statistisch auswerten und Informationen extrahieren zu können.

Im Falle von *Triple Chaser* dienen diese Typisierungen der technischen Identifikation von Tränengasgeschossen. Forensic Architecture arbeitet mit diesem Werk auf die automatisierte optische Identifikation dieser Munition in der Masse der vielfach online hochgeladenen Videos hin, die beispielsweise die staatliche Störung und Zerschlagung von Protestbewegungen zeigen.[5] Das Ziel ist es, die digitale Objekterkennung so zu manipulieren, dass derartige Reizkampfstoffe in online zugänglichen Bildpools automatisiert erkannt, freigestellt und extrahiert werden können. Dafür werden synthetische Renderings entsprechender Schusskörper verwendet, die als Trainingssets für Klassifikatoren maschineller Lernverfahren dienen. Die Modulation dieser künstlichen Bilddaten von Tränengasgeschossen basiert auf der Analyse ›echter‹ Schusskörper. Grundlegend wird das erst machbar, weil Proteste zunehmend medialisiert und damit für IT-forensische Zwecke analysierbar werden – unter anderem durch Posts auf Twitter, die Bilder von Tränengaskanistern zeigen.

Zusätzlich wurden diese digital modellierten Kanister nach realem Vorbild einer Triple-Chaser-Granate in diversen digitalen Umgebungen platziert. So gelang es, unter anderem Einsatzabläufe zu dokumentieren, entsprechende Daten auszuwerten und diese ebenfalls als Arbeitsgrundlage für automatisierte Objekterkennungstools zu definieren (vgl. Forensic Architecture 2020b).

Die Notwendigkeit dieser Praxis wird damit begründet, dass für Schusskörper wie Tränengasgeschosse bislang zu wenige Bilder als Befunddatensätze in den Lerndaten von Objekterkennungssoftwares verfügbar seien (vgl.

---

5 *Triple Chaser* gingen verschiedene Verstrickungen politischen und institutionellen Handelns voraus: Als US-Grenzbeamte im November 2018 Tränengasgranaten auf Zivilist:innen abfeuerten, zeigten Fotos dieses Vorgangs, dass viele dieser Granaten von der Safariland Group stammten, einem der weltweit größten Hersteller von als nichtletal klassifizierter Munition. Die Mehrheitsanteile an der Safariland Group gehörten zu dieser Zeit dem Industriellen Warren B. Kanders, dem nunmehr ehemaligen stellvertretenden Vorsitzenden des Kuratoriums des Whitney Museum of American Art in New York City. Als Reaktion auf die Einladung von Forensic Architecture zur *Whitney Biennale* 2019 und die Kontroverse um Warren B. Kanders' Verbindung mit dem internationalen Waffenhandel begann Forensic Architecture das Werk *Triple Chaser* zu entwickeln, in dem Klassifikatoren maschinellen Sehens mit dem Ziel trainiert wurden, von Safariland produzierte Tränengaskanister des Typs »Triple-Chaser« unter Millionen online geteilter Bilder zu erkennen.

Schmuckli 2020: 48f.). Dadurch ist die automatisierte Suche nach diesen Geschossen in entsprechenden Suchprogrammen stark beeinträchtigt. Grund dafür ist vornehmlich ihre bisherige Klassifizierung durch diverse staatliche Behörden als nicht-tödlich hinsichtlich ihrer Einsatzintention (vgl. UN 2020: 29ff.). Zusammenfassend wird also mittels maschineller Lernverfahren und aufbauend auf einer Modifikation der zur Verfügung stehenden Trainingsdatenbasis an einer breiteren Erkennung der Munition gearbeitet (siehe Abb. 2).

Abb. 2: Forensic Architecture, Triple Chaser (Still), 2018/19, Kennung einer Triple-Chaser-Granate in einer Bilddatei.

© Forensic Architecture/Praxis Films[6]

Die dem Werk zugrunde liegenden Ästhetiken können als investigativ beschrieben werden: Die Bilddaten, auf die die Objekterkennungssoftware zurückgreift, stammen meist aus der Erfassung und Analyse visueller Informationen, beispielsweise aus der Auslese von Bildinformationen aus Social Media (vgl. Meyer 2021: 18f.). Da diese Daten erst einmal beschafft werden mussten,

6 https://forensic-architecture.org/investigation/triple-chaser. Zugegriffen: 16. September 2022.

setzte Forensic Architecture auf zivile Mitarbeit. So wurde etwa für die Kennzeichnung und Etikettierung der für das Werk grundlegend benötigten Bilddaten auf Open-Source-Lösungen zurückgegriffen. Damit sind auch die Mechanismen affektiver Ökonomien und von Meinungsbildung im Interesse des Gemeinwohls prädestinierend für die Bilder des Widerstandes, als die sich Forensic Architectures Bildsynthesen in *Triple Chaser* – in Anlehnung an Kerstin Schankweiler (2020: 45) – durchaus beschreiben lassen. Folglich »rekonstruiert [Forensic Architecture] mithilfe maschineller Sichtbarkeitsdispositive soziopolitische Ereignisse in modellhaft angeordneten digitalen Raumdispositiven« (Naß 2021a: 46), ohne das kritische Moment der Realisationstechnologien zu reflektieren, auf denen *Triple Chaser* beruht. Das heißt – wie gezeigt – nicht, dass das Werk per se unkritisch ist, aber es übt eben keine Kritik an den Möglichkeiten und Entstehungsbedingungen jener Technologie, die an seiner Realisation beteiligt ist, sondern nutzt diese vielmehr aktiv, um investigativ in Bilddatenbanken zu intervenieren.

Damit steht ein Werk wie dieses in der Tradition der meisten technikvermittelten Kunstwerke, denen eine kritische Reflexion ihrer Materialität und damit oftmals ihrer technologischen Verfasstheit kein primäres Anliegen ist. Auch im theoretischen Nachdenken über Kunstwerke werden Technologien zwar mitgedacht, wenngleich nicht permanent in Verantwortung genommen (vgl. Rammert/Schubert 2017: 351). Hinsichtlich der Persistenz, mit der Systeme Künstlicher Intelligenz Einzug in die Kunst der Gegenwart halten, erscheint mir dieses Ergebnis Grund genug, um für meine weitere Analyse ein Beispielwerk heranzuziehen, das sich ebenfalls mit maschineller Bildlichkeit auseinandersetzt, dabei aber die Nutzung von KI-Technologie kritisch mitdenkt.

Kritik: Nora Al-Badri, *Babylonian Vision*, 2020

In *Babylonian Vision* aus dem Jahr 2020 nutzt Nora Al-Badri ebenfalls synthetische Bilder, um auf ein gesellschaftliches Problem aufmerksam zu machen und setzt sich auch mit »forensische[n] Spuren der Erinnerung« (Al-Badri/Scorzin 2021: 142) auseinander: Ihr Werk behandelt das Missverhältnis von Repräsentationslücken in Datensätzen digitalisierter Museumsbestände bei gleichzeitigem Überhang von digitalisierten Artefakten als museologischem Kapital. Al-Badri bezieht sich in diesem Zusammenhang auf Sonia K. Katyals Forschungen zur Verbindung von Technologie und kulturellem Erbe,

die Katyal unter dem Begriff »Technoheritage« (2017) verschlagwortet.[7] Ein Begriff, der konstatiert, dass digitalisierte Bilder in musealen Datenbanken, die vielmals Reproduktionen der jeweilen Sammlungsobjekte enthalten, als »digitale Vermögenswerte« gelten können (vgl. Al-Badri/Scorzin 2021: 142). Grundlegend gehen die Überlegungen der Künstlerin also davon aus, dass Museen hinsichtlich der Provenienz ihrer Objekte – also bezüglich der Objektbiografie – oftmals ebenso opak bleiben wie im Umgang mit ebenjenen Daten, die digitalisierte Abbildungen dieser Objekte enthalten. Somit ist die Ausgangslage mit jener vergleichbar, die Forensic Architecture bei der Arbeit an *Triple Chaser* vorfand: Ein Mangel an benötigten Bilddaten führt zur (künstlerischen) Arbeit mit KI. Allerdings stellt Al-Badri in ihren Bezugnahmen und Äußerungen zum Werk immer wieder heraus, wie relevant für ihre Arbeit eine kritische Reflexion der Bilderzeugung mittels KI-Technologie ist (ebd.).

Al-Badri weist darauf hin, dass das institutionalisierte westliche Kulturverständnis mit der hegemonialen Dominanz des globalen Nordens Hand in Hand geht. Diese Übermacht wird durch museale Artefaktpools fundiert, deren digitalisierte Reproduktionen eine Art KI-gestütztes Bollwerk kultureller Big Data bilden. Die Künstlerin formuliert dazu, dass heute »Datensätze mit Millionen von Bildern trainiert« werden, deren Inhalte zu »ungefähr 80 % […] aus dem Globalen Norden [stammen]«, dadurch trage KI entscheidend dazu bei, eine perpetuierte »visuelle Hegemonie« aufrechtzuerhalten (ebd.: 148). In diesem Sinne kann die technische Bilderzeugung – beispielsweise im musealen Auftrag – und das darauffolgende Horten von digitalisierten Bildern als ein Angelpunkt des Machterhalts durch ein Festschreiben von Deutungshoheit begriffen werden. Die Verschränkung der expliziten Affirmation hegemonialer visueller Praktiken des Westens mit KI – zumindest was die in sie eingeschriebene explizite Affirmation westlicher Hegemonie anbelangt – markiert KI als Bestandteil kolonialistischer Genealogien (vgl. Cave/Dihal 2020).

Al-Badri erweitert in ihrem Werk *Babylonian Vision* digitale Museumsbestände um eigens erzeugte digitale Artefakte, die als Widerlager gegen die Ex-

7 Diese Begrifflichkeit kann auch als ein Verweis Al-Badris auf jüngste Forderungen verstanden werden, bei der Entwicklung von KI beispielsweise indigene Perspektiven zu berücksichtigen (vgl. Al-Badri/Scorzin 2021: 148). In diesen emanzipatorischen und dekolonialen Denkansätzen werden unter anderem Objektdatensätze als ebenso gewichtig markiert wie die ursprünglichen Objekte. Damit wird das Recht auf eine eigene Deutungs- und Zugangshoheit bezüglich historisch kolonial besetzter Artefakte eingefordert (vgl. Lewis 2020).

traktion von Kulturgütern aus Ursprungskontexten verstanden werden können. Überdies weist Al-Badri auf die Robustheit und Persistenz von Problemverknüpfungen hin, wie sie beispielsweise zwischen Race und Technologie bestehen (vgl. Adas 1990: 3).

Abb. 3: Nora Al-Badri, Babylonian Vision (Still), 2020, GAN-Video.

© Nora Al-Badri[8]

Für Al-Badris Werk wurde ein maschinelles Lernmodell vortrainiert und so ausgebildet, dass es in der Lage ist, anhand der Einspeisung von rund zehntausend Bildern aus fünf verschiedenen Museumssammlungen neue Bilder zu generieren (siehe Abb. 3). Das Modell wurde mit Bilddaten aus Sammlungsbeständen trainiert, die mesopotamische, neosumerische und assyrische Kunst- und kunsthandwerkliche Objekte zeigen. Die Auswahl der diesbezüglichen Abbildungen erfolgte ebenfalls – wie im Falle Forensic

8 https://www.nora-al-badri.de/works-index. Zugegriffen: 16. September 2022.

Architectures – investigativ, hier etwa durch eine automatisierte Bildsuche mit *Webcrawlern* oder durch Verfahren zum Auslesen von Computerbildschirmen mittels *Screen Scraping* (vgl. Al-Badri/Scorzin 2021: 142). So betont die Künstlerin in diesem Zusammenhang, dass sie die investigativ gewonnenen Bilddaten und die entsprechend für die Beschaffung der »Datensätze aus Museen« verwendeten Technologien als »eine künstlerisch-emanzipatorische Strategie zur Dekolonisierung der Museen« (ebd.) begreift.

Durch eine Subsystematisierung mittels eines sogenannten Generative Adversarial Networks, kurz: GAN,[9] sind so neue synthetische Bilder entstanden, die ihre Herkunft und Herstellungsbedingungen nicht verbergen. Diese scheinen unter anderem aus Reproduktionsfotografien von musealen Objekten gewonnen: Davon zeugen etwa Bildbestandteile, die auf abfotografierte Farbmaßlineale schließen lassen, die beispielsweise zur nachträglichen Farbkalibrierung genutzt werden. Diese werden in den auf der Website der Künstlerin einsehbaren Trainingsbildern ansichtig (vgl. Al-Badri 2020). Auch mäandern die Konturen, Flächen und Kanten der ›fertigen‹ Videobilder, die Al-Badris generierte Artefakte zeigen. Sie legen eine dezidierte Bezugnahme auf typisch maschinelle Ästhetiken nahe, die – entgegen den visuellen Möglichkeiten von GANs heute – ihre Künstlichkeit nicht verschleiern und so optisch diffus und ohne formale Evidenzbehauptung verbleiben können (vgl. Hertzmann 2020). Diese Ästhetik ist gegensätzlich zu jener der synthetischen Bilder in *Triple Chaser*, deren Qualitätsmaßstab es ist, so ›echt‹ wie möglich zu wirken.

So sind die aus digitalisierten musealen Artefakten entstandenen 150 Videobilder neue digitale Artefakthäufungen, die angesichts dessen, dass sie nicht suggerieren, es gäbe für sie eine faktische Entsprechung im Realraum, nur im digitalen Raum existieren. Dafür spekulieren Al-Badris Artefakte über die Qualität und Herkunft der Originalobjekte und stellen auch deren inhaltlichen Evidenzanspruch als Säule des kulturellen Gedächtnisses und der Hoheit westlicher Kunst- und Kulturinstitutionen infrage.

Aufbauend auf den beiden vorgestellten Werken wird evident, dass die Steigerung des Ansehens von KI-Technologien durch das Aufzeigen potenziell progressiver Einsatzmöglichkeiten – wie im Falle von *Triple Chaser* – ebenjene

9 GANs ›arbeiten‹ über zwei konkurrierende künstliche neuronale Netze, von denen eines die Aufgabe hat, real wirkende Daten zu erzeugen, und das andere die Daten als dementsprechend echt oder künstlich klassifiziert. Durch ständiges Lernen und zahlreiche Iterationsschritte werden die generierten Daten potenziell immer realistischer.

Problemlagen kaschieren und multiplizieren kann, mit denen Systeme Künstlicher Intelligenz seit ihrer Entstehung untrennbar verbunden sind (vgl. Chun 2021). Denn wie die Kunstwissenschaftlerin Mira Anneli Naß anschlussfähig betont, scheint die »theoretische Rezeption sich betont politisch gebender Arbeiten […] eine kritische Analyse derselben, also ihre Historisierung und Kontextualisierung, […] häufig zu vernachlässigen« (2021b: 135).

Eine Arbeit wie *Babylonian Vision* hingegen verweist eher auf dieses Problem, als es unberührt zu lassen, denn:

> Every AI generated image is an infographic about the dataset. AI images are data patterns inscribed into pictures, and they tell us stories about that dataset and the human decisions behind it. (Salvaggio 2022)

So werden durch dieses Werk etwa bilderkennende und bildgebende Verfahren künstlicher Intelligenzsysteme als aus prekären globalen Arbeitsstrukturen erwachsend und in diese eingebettet markiert, für die zumeist Personen im globalen Süden ausgebeutet werden (vgl. Crawford 2021: 64f.). Des Weiteren wird aufgezeigt, dass Repräsentationslücken und Generalisierungen auf normativen, oftmals *weißen* Trainingsdaten fußen und diskriminierende Etikettierungen von Bilddaten Ausschlüsse und Marginalisierungen (re)produzieren (vgl. Apprich et al. 2018; Noble 2018). Außerdem wird deutlich, inwieweit von ›Menschenhand‹ initiierte Prozesse durch technologische Verantwortungsüberblendung insbesondere im Verhältnis zu maschinellen Lernverfahren verschleiert werden (vgl. Campolo/Crawford 2020).

Wie aber lassen sich die aufgefächerten Kritikformen differenzieren? In kritischer Kunst kann Kritik einerseits mittels eines Werks formuliert beziehungsweise durch ein Werk vermittelt werden. Im Falle von *Tripel Chaser* und *Babylonian Vision* lässt sich diese Kritik als ›engagiert‹ beschreiben. Andererseits und wesentlich seltener wird Kritik manifest, die ein Werk introspektiv übt, beispielsweise an der Begrifflichkeit dessen, was Kunst ausmacht und welche Distributionsmechanismen Kunst Sichtbarkeit verleihen, aber auch – wie im Falle von *Babylonian Vision* – mithilfe welcher technologischen Neuerungen Kunstwerke hergestellt werden und wie sich diese Neuerungen in die Werke einschreiben.

Die Suche nach der Art der Kritik, die mittels der Nutzung maschineller Lernverfahren geübt wird, lässt die Frage zu: Ist das Äußern dieser ›nach innen‹ und damit auf die verwendete Technologie gerichteten Kritik tatsächlich ausschlagend für die künstlerische Auseinandersetzung gewesen? Oder fungiert die Technologiekritik schlicht als Rechtfertigung beziehungsweise

als eine Art Freibrief für die Nutzung der entsprechenden Technologie? Diese Fragen werden Gegenstand zukünftiger Untersuchungen sein müssen, denn an dieser Stelle lässt meine Untersuchung offen, ob – wie Hannes Bajohr (2021) konstatiert – »[k]ritische Kritik, selbst immanente, […] eher das Abzeichen [ist], das man sich ans Revers heften muss, um die neuen Technologien ohne drohenden Prestigeverlust im Kunstdiskurs erproben zu dürfen«, oder ob diese Kritik etwa ein Etikett ist, das Kunstwerke in einem so stark frequentierten Feld wie der KI-Kunst als besonders informiert und diskursiv hervorhebt und damit beispielsweise als institutionell vermittelbarer ausweist.

Anlass: Pierre Huyghe, *Untilled*, 2012, und Anna Ridler, *Myriad (Tulips)*, 2018

Ich schließe meine Untersuchung mit zwei Werken, die die beiden verschiedenen Bezüge verdeutlichen, auf Basis derer KI als Anlass der Werkgenese untersucht werden kann. In diesen künstlerischen Arbeiten dienen Systeme Künstlicher Intelligenz als Elemente innerhalb einer bildkünstlerischen Erzählung. Obgleich solche Werke in der Regel auf die Inkorporation und Nutzung von KI-Technologien verzichten, können sie bei der konzeptuellen Bestimmung von KI sehr hilfreich sein, da sie Ambivalenzen von KI sowie entsprechende historische Referenzen aufzeigen.

Wie beschrieben, unterscheide ich hinsichtlich dieser letzten Form innerhalb der Trichotomie der Nutzungsaspekte von KI zwischen Werken, deren Anlass, sich mit KI zu befassen, entweder narrativ-affirmativ oder kritisch-aversiv ist. Des Weiteren werden in den Werken KI-Technologien beziehungsweise -Modelle zum Ausgangpunkt genommen, die in den bisher besprochenen Werken besonders prominent gewesen sind: künstliche neuronale Netze sowie Objekterkennungssoftware und die ihr zugrunde liegenden Trainingsdaten.

Ein Beispiel für eine affirmative Bezugnahme auf KI ist das anlässlich der *documenta 13* in Kassel entstandene Werk *Untilled* (2012) von Pierre Huyghe.[10] Im Rahmen eines positivistischen Technologieverständnisses, das an einem narrativen Arrangement exemplifiziert wird, verschränkt Huyghe darin die bei

10 Die Folgearbeit *Exomind (Deep Water)* von 2017/2020 habe ich weiter oben bereits als ebenso affirmativ gekennzeichnet. Sie wurde unter anderem 2020 in der Ausstellung *Uncanny Valley: Being Human in the Age of AI* im de Young Museum in San Francisco gezeigt.

höheren Organismen stattfindende Signalübertragung über Nervenzellen mit Funktionsweisen und Anwendungen von künstlichen neuronalen Netzen.

Kern von *Untilled* ist die Reproduktion eines als weiblich zu lesenden Aktes in patiniertem Betonguss mit einem integrierten Heizsystem.[11] Der Kopf des Nachgusses ist mit einem Bienenstock bewachsen; Wabenstrukturen und ein intaktes Bienenvolk sind in lamellenförmigen Ausprägungen sichtbar (siehe Abb. 4).

Abb. 4: Pierre Huyghe, Untilled, 2012, Lebewesen und unbelebte Dinge, Maße variabel.

© Pierre Huyghe/Galerie Esther Schipper, Berlin[12]

11 Der Betonguss ist einer Bronzeskulptur des Bildhauers Max Weber aus dem Jahr 1949 nachempfunden. Dieser *Liegende Frauenakt* wurde 1982 von der Berufs- und Fortbildungsschule in Winterthur angekauft. Dort ist die Skulptur, die in ihren Abmessungen (105 × 145 × 44,5 cm) Huyghes Nachbildung gleicht, im Park parallel zur Begrenzung der Brunnenfläche hinter dem Altbau der Schule zu sehen.

12 https://www.estherschipper.com/de/exhibitions/386-untilled-pierre-huyghe/. Zugegriffen: 16. September 2022.

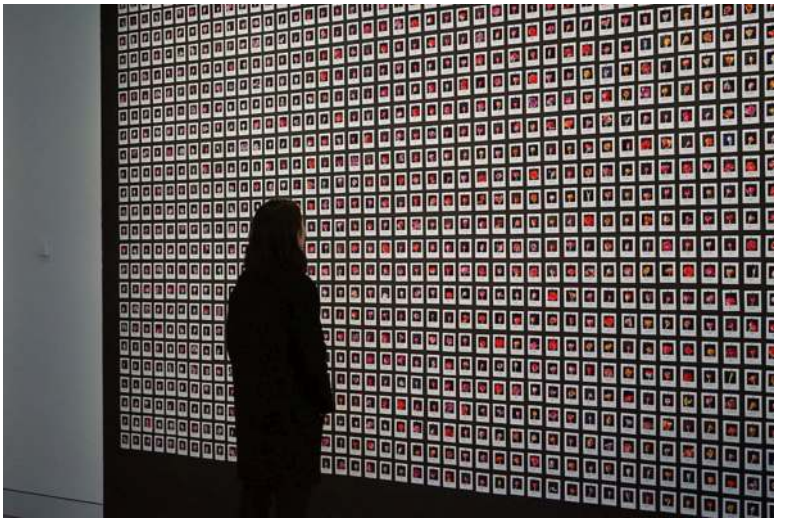
Der halbliegende, sich rechtsseitig auf den rechten Unterarm stützende Akt befindet sich auf einem Sockel. Die Umgebung des Objekts wird bewohnt von zwei Jagdhunden der Rasse *Podenco* sowie von vor Ort beheimateten Ameisenvölkern. Ebenso finden sich dort Pfützen mit Algenbedeckung, ein wannenartiger Gesteinsquader, der von Kaulquappen und später Fröschen bewohnt wird, und andere Bestandteile des Arrangements wie Pflastersteinstapel und geordnete Aufschüttungen von schwarzem Schiefer. Da das Werk aufgrund der schwärmenden Bienen und seiner pflanzlichen Bestandteile zu Wachstum imstande ist, sind seine genauen Ausmaße nicht fassbar (vgl. Schütze 2018: 212–218). Denn die Bienen interagieren unter anderem mit den Pflanzen in der Umgebung. So verunklart *Untilled* die Einordung dessen, was künstlerischer Eingriff, was Werk und was Szenerie ist (vgl. Hantelmann 2015: 225). Dass der Kopf des Aktes – mithin der Trägerort des menschlichen Hirns – durch das Nest einer dynamischen Schwarmintelligenz, der Bienen, ersetzt ist, weist das Werk als Netzwerk aus (vgl. ebd.: 231).

Untilled liegt die Idee koproduzierender schwärmender Intelligenzen als Form geistiger Kapazität zugrunde – was die Brücke zur KI-Affirmation schlägt. Mit diesem Werk wird unter anderem eine Metapher für die Grundlagen der Modellierung neuronaler Netze geschaffen und für ein Verständnis von funktionaler Intelligenz argumentiert: Das Koproduzieren von Intelligenzen, die erst im Kollektiv oder Schwarm ihr Potenzial entfalten, ist in der Natur eine Gegebenheit, die sich etwa in Bienenstaaten, Termitenvölkern, Vogel- und Fischschwärmen erkennen und als Konzept von Emergenzen untersuchen lässt (vgl. Oxenham 2017). Die Einzelorganismen besitzen dabei geringe Kenntnis über ihre Umwelten. Sie interagieren nur mit einer begrenzten Anzahl von Artgenossen; doch trifft die Gruppe als Ganzes koordinierte Entscheidungen. Diese Schwarmentscheidungen solcher Superorganismen bilden sich über Rückkopplung heraus, in dem jedes Wesen sein Verhalten an demjenigen seiner benachbarten Artgenossen ausrichtet und diese in ihrem Verhalten beeinflusst (vgl. Couzin 2008). Bestimmte Aspekte dieser funktionellen Ausprägung von kognitiver Leistung, etwa in einer Ameisenkolonie, können in Form von Regeln erfasst und mit Computerprogrammen simuliert werden (vgl. Millhouse/Moses/Mitchell 2021: 27–30). Das dementsprechende Arbeitsgebiet seitens beispielsweise der Informationswissenschaft versucht, komplexe vernetzte Softwareagentensysteme nach dem Vorbild staatenbildender Insekten zu modellieren. Die Analyse entstehender kooperativer Verbindungen dient dazu, höhere kognitive Leistungen analog zu KI zu simulieren. Huyghes Werk wird in diesem Sinne inhaltlich von einem bejahenden

KI-Narrativ getrieben, das sogar diametral liegende Sphären wie Natur und Technik zusammendenkt.

Ein Beispiel für eine kritische Bezugnahme auf KI – genauer gesagt: auf die bereits erläuterten maschinellen Bildgebungs- und Objekterkennungsverfahren –, das ohne technische Nutzung derselben auskommt, ist Anna Ridlers Werk *Myriad (Tulips)* von 2018. Die Installation besteht aus 10.000 handbeschrifteten Fotografien von Tulpen, die auf einer circa 50 Quadratmeter großen Wandfläche gruppierend gehängt sind (siehe Abb. 5).

Abb. 5: Anna Ridler, Myriad (Tulips), 2018, C-Typ-Digitaldrucke mit handschriftlichen Notizen, Magnetfarbe, Magnete, ca. 50 qm.



© Anna Ridler[13]

Der Ausgangsimpetus der Künstlerin war es laut eigener Aussage, mit *Myriad (Tulips)* einen bildbasierten Datensatz zu dekonstruieren, indem dieser in einzelne, dem Satz zugrunde liegende visuelle Gliederungselemente in Form von Fotografien zerlegt und damit das serielle Prinzip als Methode der iterativen Bildgenerierung ansichtig gemacht wird (vgl. Ridler 2019). In dem Werk werden also Bilderkennungs- und Bildgenerierungsverfahren zum Anlass künstlerischer Arbeit genommen. Um diese Prozesse zu verdeutlichen,

13 http://annaridler.com/myriad-tulips. Zugegriffen: 16. September 2022.

wurden von Ridler Tulpen beschafft, inszeniert und fotografiert. Danach wurden die entwickelten fotografischen Abbilder dieser Blumen von der Künstlerin händisch geordnet, beschreibend etikettiert und katalogisiert (vgl. ebd.). Durch die Entscheidung, die Technologie der Bildetikettierung und -klassifizierung manuell zu rekonstruieren, wird die Aufmerksamkeit auf Tätigkeiten gelenkt, die hinter den maschinellen Abläufen im Umgang mit Bilddaten liegen.[14]

In Ridlers Werk werden entsprechende Kategorisierungsmechanismen durch handgeschriebene Annotationen unter jedem Foto sichtbar, die etwa Farbe, Blütenstand und Zustand der Blüte, Musterung und so weiter ausweisen. Dabei tritt auch das Problem der Datensätze offen zutage, denn eine implizite Verzerrung ist unvermeidlich, da die für Klassifizierungen vorgenommenen Etikettierungen generell auf menschlichen und damit auf subjektiven Entscheidungen beruhen (vgl. Crawford 2019). So stellte die Künstlerin während des Prozesses fest, dass sie dazu neigte, eher farbige Tulpen auszuwählen. Ohne Regulation hätte das dazu geführt, dass ein entsprechend trainiertes Modell besser darin gewesen wäre, farbige Tulpen zu generieren; angesichts dessen musste Ridler mehr weiße und helle Tulpen als Trainingsdaten auswählen, um diesen Sachverhalt authentisch kompensieren zu können (vgl. Ridler 2019).

Grundlegend verdeutlicht *Myriad (Tulips)*, dass sich die visuelle menschliche Wahrnehmung stark von maschinellen Erkennungsweisen unterscheidet, denn in Prozessen maschinellen ›Sehens‹ spielt es keine Rolle, dass die Anhäufung von Bildbestandteilen beziehungsweise die Verrechnung der Datenpunkte visuell eine Pflanze suggeriert, die beispielsweise grün, knospig, verzweigt, dicht beblättert oder von fächerigem Wuchs ist. Vielmehr beschreibt der Vorgang eine Häufigkeitsverteilung von Pixelmustern, die mit dem Label ›Pflanze‹ korrelieren. Kulturell tradierte Umgangsweisen mit dem Bild werden in diesem Fall also nicht mehr vom Status ›Bild‹ bestimmt, sondern vom Umgang der Maschine mit entsprechenden Bildern als Datenmaterial. Durch KI wird in diesem Zusammenhang der Zugang zum Bildlichen moduliert und die Künstler:innen orientieren sich beim Umgang mit dem Bild an entsprechenden technischen Verfahrensweisen.

Myriad (Tulips) wurde – obschon das Werk nicht auf KI-Technologie basiert – in zahlreichen Ausstellungen zu KI gezeigt; so unter anderem 2018 in *Error* –

14 Die Verfahrensabläufe entsprechender Technologien wurden meinerseits anhand der automatisierten Objekterkennung am ersten Werkbeispiel erläutert.

The Art of Imperfection (Ars Electronica Export, Berlin) und 2019 in *AI: More than Human* (Barbican Centre, London). Außerdem wurden die erstellten Fotografien von Ridler als Datensätze für drei nachfolgende Arbeiten verwendet: für die Videoinstallationen *Mosaic Virus* von 2018 und 2019 und für die Ethereum-Plattform *Bloemenveiling* von 2019.

Ursprünglich als Recherchearbeit für die Videoinstallation *Mosaic Virus* geplant, die ähnlich wie Al-Badris Werk *Babylonian Vision* digitale Artefakte – allerdings von Tulpen – via GANs generiert, avancierte die komplexe Collage aus Tulpenbildern für die Künstlerin rasch zu einer eingeständigen Arbeit. Grund dafür war eine im Werkprozess enthaltene Kritik an KI: Während der Prozess des Aufbaus eigener Datensätze iterativ und experimentell ist, verlangt er parallel »extreme Kontrolle, die die Künstlerin über die Datenerhebung hat«, außerdem »erinnert die Arbeit, die mit der Erstellung dieses Datensatzes verbunden ist, an die unsichtbare Arbeit von Frauen und marginalisierten Bevölkerungsgruppen« (Audry 2021: 126; Übersetzung M. K.-K.). Deren Beteiligung an der Arbeit mit KI wird laut Ridler (2019) oftmals verschleiert, weshalb die Künstlerin mit der Eigenständigkeitserklärung dieses Werkes die manuelle und oftmals prekäre Arbeit hinter etikettierten Bilddatensätzen anerkannt wissen möchte. Außerdem ist Ridlers Werk mit den Tulpen ein Nachvollzug der beschwerlichen Arbeit von Gärtner:innen, wenngleich das keine Kritik an KI, sondern eine an globalen ökonomischen Strukturen ist (vgl. Audry 2021: 160). So sitzt beispielsweise ein großer Teil der für den europäischen Markt produzierenden Schnittblumenindustrie in Kenia. Die gering entlohnte und körperlich strapaziöse Aufzucht der Pflanzen hat verheerende ökologische und soziale Folgen für die dortige Bevölkerung wie auch für die kenianische Flora und Fauna (vgl. Schönberger 2020).

Das Kunstwerk ermöglicht überdies eine weitere KI-kritische Betrachtung, indem es auf den multivariaten Irisblumendatensatz – auch bekannt als ›Fisher's Iris‹ – Bezug nimmt, der 1936 von dem Statistiker Ronald Fisher erstellt wurde. Dieser enthält – im Gegensatz zu Ridlers Myriade an Bildern – nur jeweils 50 Abbilder als Proben der drei Schwertlilienarten *Iris setosa*, *Iris virginica* und *Iris versicolor*. Er wird als Beispiel für viele statistische Klassifizierungsverfahren beim maschinellen Lernen verwendet. Fisher analysierte vier Differenzierungsmerkmale und entwickelte darauf aufbauend ein lineares Diskriminanzmodell, um die Unterschiede der Irisarten zu fixieren. Dieses Modell wird oftmals als Beispiel verwendet, um unter anderem die Differenz zwischen überwachtem und nicht überwachtem Lernen, etwa hinsichtlich einer Methode wie der linearen Regression, zu erklären (vgl. Joselson 2016).

Auch ist der Irisdatensatz Teil diverser freier Softwarebibliotheken, etwa von *Scikit-learn*, womit »jedes Programm für maschinelles Lernen, das dieses Softwarepaket verwendet, auch irgendwo einen versteckten Blumendatensatz enthält« (Ridler 2019; Übersetzung M. K.-K.). Durch die Bezugnahme auf Fisher werden auch eugenische und rassistische Aspekte evident, da Fisher lineare Diskriminanten unter anderem entwickelte, um Race anhand von Schädelgrößen zu unterscheiden (vgl. Stodel 2020; Chun 2019: 465). Der Datensatz selbst wurde ebenso 1936 im Journal *Annals of Eugenics* veröffentlicht (vgl. Masch et al. 2021: 17). Ridler selbst verweist hingegen auf Parallelen zu den in *Myriad (Tulips)* veranschaulichten Technologien und sieht ihr Werk mit den inhärenten Problemen des maschinellen Lernens verstrickt: der in der Technologie angelegten korrelativen Voreingenommenheit, der eingeschriebenen Diskriminierung und der Grenzbereichsdetermination der Datensätze (vgl. Ridler 2019; Chun 2019: 465).

Schluss

In der Beschäftigung mit und im Herstellen von Kunst ist eine Denkbewegung von manueller zu visueller Arbeit enthalten. Dieser Transformationsprozess kann als Plädoyer für eine Diskursivierung des Bildes als Denkfigur dienen. So beinhaltet eine Fürsprache wie diese eine immanente Kritik des Kunstwerks an seinen Rezeptionsbedingungen und -auswirkungen und wendet sich kritisch gegen einen zu laxen Umgang mit bildgebenden Praxen und Technologien.

Meine hier aufgefächerte Trichotomie der Nutzungsaspekte von KI in Kunst setzte bei diesem Standpunkt an und zeigte eine Reihe von Verwendungsformen von und Bezugnahmen auf KI auf, die – wie beschrieben – unfest, also offen, sind. Ich möchte mit meinem Vorschlag eines Analyserahmens deutlich machen, in welcher Breite maschinelle Lernverfahren und KI-Technologien den Zugang zum Bildlichen zeitigen. Gleichzeitig war es mein Anliegen, aufzuzeigen, in welcher Vielfalt Künstler:innen zum Thema KI und zu dahingehenden Verfahrensweisen des maschinellen Lernens arbeiten und KI-Technologien für ihre Kunstwerke nutzen. Meine Intention war es, darzulegen, welche Vorteile ein genaues Benennen der Art und Weise der Nutzung von KI in einem Werk bietet: Das Resultat dessen kann ein differenziertes Befragen der Kunst der Gegenwart bezüglich der Reaktionen auf technologische Neuerungen sein. Dieses Vorgehen setzt – in Abgrenzung

zu zahlreichen bestehenden Ansätzen – beim Kunstwerk an und lässt Rückschlüsse auf den Status des Technischen in der Kunst, aber auch auf die Frage zu, wie die heutige Kunstproduktion mit Technik zusammenwirkt.[15]

So lässt sich die Gegenseitigkeit von Kunst und KI-Technologie in den diskutierten Werken als symptomatisch für einen unterschiedlichen Umgang mit KI lesen – beispielsweise im Sinne einer unkritischen oder einen kritischen Haltung gegenüber KI-Technologie, aber auch im Sinne ihrer Nutzung für utopische oder metaphorische Narrative.

Literatur

Adas, Michael. 1990. *Machines as the Measure of Men: Science, Technology, and Ideologies of Western Dominance*. New York: Cornell University Press.

Al-Badri, Nora und Pamela C. Scorzin. 2021. KI und Technoheritage. *Kunstforum* 278: 138–151.

Al-Badri, Nora. 2020. Works Index. https://www.nora-al-badri.de/works-index. Zugegriffen: 5. Mai 2022.

Anadol, Refik und Pamela C. Scorzin. 2021. Daten als Pigment. *Kunstforum* 278: 162–175.

Apprich, Clemens, Wendy Hui Kyong Chun, Florian Cramer und Hito Steyerl. 2018. *Pattern Discrimination*. Lüneburg: meson press.

Audry, Sofian. 2021. *Art in the Age of Machine Learning*. Cambridge: MIT Press.

Bajohr, Hannes. 2021. Malen nach 0 und 1. https://www.republik.ch/2022/05/07/malen-nach-0-und-1. Zugegriffen: 5. Mai 2022.

Becker, Howard S. 2008 [1982]. *Kunstwelten*. Hamburg: Avinus.

Bense, Max. 1965. *Aesthetica*. Baden-Baden: Agis.

Bippus, Elke. 2010. Wissensproduktion durch künstlerische Forschung. *subText04. research@film – Forschung zwischen Kunst und Wissenschaft*: 9–21.

15 Es bieten sich auf Basis der hier aufgeführten Werke auch weitere Modelle an, die andere Aspekte herausstellen: So erscheint es mir ebenso produktiv, zwischen Werken zu unterscheiden, die mit Bildern quantitativ oder qualitativ verfahren. Forensic Architectures Werk etwa argumentiert ausgehend von Bildmengen vermeintlich evidenzbasiert, parallel zu Pierre Huyghes Werk, der künstlerisch über technische Evidenzen spekuliert. Im Gegensatz dazu wird in den Werken von Nora Al-Badri und Anna Ridler sichtbar, wie mittels Bildern qualitativ verfahren wird, um beispielsweise eine vermeintliche Evidenzbasiertheit zu entkräften.

Bogost, Ian. 2019. The AI Gold Rush Is Here. https://www.theatlantic.com/technology/archive/2019/03/ai-created-art-invades-chelsea-gallery-scene/584134/. Zugegriffen: 5. Mai 2022.

Campolo, Alexander und Kate Crawford. 2020. Enchanted Determinism: Power without Responsibility in Artificial Intelligence. *Engaging Science, Technology, and Society* 6: 1–19.

Caselles-Dupré, Hugo. 2018. Kunst und künstliche Intelligenz. Christie's versteigert Werk eines Algorithmus. https://www.monopol-magazin.de/christies-versteigert-werk-eines-algorithmus. Zugegriffen: 5. Mai 2022.

Cave, Stephen und Kanta Dihal. 2020. The Whiteness of AI. *Philosophy & Technology* 33: 685–703.

Chun, Wendy Hui Kong. 2019. Data Segregation and Algorithmic Amplification. *Canadian Journal of Communication* 44: 455–469.

Chun, Wendy Hui Kong. 2021. *Discriminating Data: Correlation, Neighborhoods, and the New Politics of Recognition*. Cambridge: MIT Press.

Couzin, Iain D. 2008. Collective Cognition in Animal Groups. *Trends in Cognitive Sciences* 13, H. 1: 36–43.

Crawford, Kate. 2019. Conversation by Kate Crawford and Trevor Paglen. In *Training Humans. Notebook 26*. Mailand: Fondazione Prada.

Crawford, Kate. 2021. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven und London: Yale University Press.

Ferial Nadja Karrasch. 2019. Käthe-Kollwitz-Preis: Hito Steyerl in der Akademie der Künste. https://www.art-in-berlin.de/incbmeld.php?id=4958. Zugegriffen: 5. Mai 2022.

Flusser, Vilém. 1985. *Ins Universum der technischen Bilder*. Göttingen: European Photography.

Flusser, Vilém. 1993. *Lob der Oberflächlichkeit. Für eine Phänomenologie der Medien. Band 1*. Düsseldorf: Bollmann.

Forensic Architecture. 2020a. Detecting Tear Gas: Vision and Sound. https://forensic-architecture.org/investigation/detecting-tear-gas. Zugegriffen: 4. Mai 2022.

Forensic Architecture. 2020b. Triple-Chaser. https://forensic-architecture.org/investigation/triple-chaser. Zugegriffen: 5. Mai. 2022.

Hantelmann, Dorothea von. 2014. Denken der Ankunft. In *Kunst und Wirklichkeit heute: Affirmation – Kritik – Transformation*, Hg. Lotte Everts, Johannes Lang, Michael Lüthy und Bernhard Schieder, 223–240. Bielefeld: transcript.

Hertzmann, Aaron. 2020. Visual Indeterminacy in GAN Art. *Leonardo* 53, H. 4: 424–428.

Hunger, Francis. 2019. Interview: Date Resists the Five-Act Form. https://www.irmielin.org/interview-data-resists-the-five-act-form-2019/. Zugegriffen: 5. Mai 2022.

Joselson, Nathaniel. 2016. Eugenics and Statistics: Discussing Karl Pearson and R. A. Fisher. https://njoselson.github.io/Fisher-Pearson/. Zugegriffen: 5. Mai 2022.

Katyal, Sonia K. 2017. Technoheritage. *California Law Review* 105, H. 4: 1111–1172.

Lewis, Jason E. (Hg.). 2020. *Indigenous Protocol and Artificial Intelligence – Position Paper*. Honolulu: CIFAR.

Locher, Hubert. 2010 [2001]. *Kunstgeschichte als historische Theorie der Kunst*. München: Fink.

Manovich, Lev. 2002. Ten Key Texts on Digital Art: 1970–2000. *Leonardo* 35, H. 5: 567–569.

Manovich, Lev. 2018. *AI Aesthetics*. Moskau: Strelka Press.

Masch, Lena, Kimon Kieslich, Katharina Huseljić, Marco Wähner und Johann-Sebastian Neef. 2021. R – Ein Einführungsskript. https://docserv.uni-duesseldorf.de/servlets/DerivateServlet/Derivate-62343/R-Eine%20Einführungsskript.pdf. Zugegriffen: 5. Mai 2022.

Meyer, Roland. 2021. *Gesichtserkennung*. Berlin: Wagenbach.

Meyer, Roland. 2022. Im Bildraum von Big Data. Unwahrscheinliche und unvorhergesehene Suchkommandos: Über Dall-E 2. *Cargo* 55: 50–53.

Millhouse, Tyler, Melanie Moses und Melanie Mitchell. 2021. Frontiers in Collective Intelligence: A Workshop Report. *Santa Fe Institute*: 27–30.

Nake, Frieder. 1966. *Herstellung von zeichnerischen Darstellungen, Tonfolgen und Texten mit elektronischen Rechenanlagen*. Darmstadt: Deutsches Rechenzentrum.

Nake, Frieder. 1974. *Ästhetik als Informationsverarbeitung*. New York und Wien: Springer.

Naß, Mira A. 2021a. Bilder von Überwachung oder Überwachungsbilder? Zur Ästhetik des Kritisierten als Ästhetik der Kritik bei Hito Steyerl und Forensic Architecture. *ffk Journal* 6: 38–56.

Naß, Mira A. 2021b. Architektur von unten? Eine Kritik komplexitätsreduzierender Praktiken bei Forensic Architecture. *Kritische Berichte* 3: 124–138.

Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press.

Oxenham, Simon. 2017. Why Bees Could Be the Secret to Superhuman Intelligence. https://www.bbc.com/future/story/20161215-why-bees-could-be-the-secret-to-superhuman-intelligence. Zugegriffen: 5. Mai 2022.

Rammert, Werner und Cornelius Schubert. 2017. Technik. In *Handbuch Körpersoziologie. Band 2: Forschungsfelder und methodische Zugänge*, Hg. Robert Gugutzer, Gabriele Klein und Michael Meuser, 349–365. Wiesbaden: Springer VS.

Ridler, Anna. 2019. Myriad (Tulips), 2018. http://annaridler.com/myriad-tulips. Zugegriffen: 4. Mai 2022.

Salvaggio, Eryk. 2022. How to Read an AI Image. The Datafication of a Kiss. https://cyberneticforests.substack.com/p/how-to-read-an-ai-image. Zugegriffen: 3. Oktober 2022.

Schankweiler, Kerstin. 2020. Das zensierte Auge. In *Bildzensur – Löschung technischer Bilder*, Hg. Katja Müller-Helle, 42–48. Berlin und Boston: de Gruyter.

Schmuckli, Claudia. 2020. *The Uncanny Valley: Being Human in the Age of AI*. Petaluma: Cameron Books.

Schönberger, Sonya. 2020. Kenyan Roses for the Kingdom. https://www.sonyaschoenberger.de/Performative/Kenyan+Roses+for+the+Kingdom+%28lecture%29. Zugegriffen: 5. September 2022.

Schröter, Jens. 2021. ›Künstliche Intelligenz‹ und die Frage nach der künstlerischen Autor*innenschaft. *Kunstforum* 278: 98–107.

Schütze, Irene. 2018. Fehlende Verweise, rudimentäre ›Markierungen‹: aufgeweichte Grenzverläufe zwischen Kunst und Alltag. *Image* 28: 204–221.

Scorzin, Pamela C. 2021a. Editorial: Kann KI Kunst? *Kunstforum* 278: 48.

Scorzin, Pamela C. 2021b. ARTificiality. Künstliche Intelligenz, Kreativität und Kunst. *Kunstforum* 278: 50–75.

Stodel, Megan. 2020. Stop Using Iris. https://www.meganstodel.com/posts/no-to-iris/. Zugegriffen: 5. September 2022.

UN – Büro des Hohen Kommissars der Vereinten Nationen für Menschenrechte. 2020. *United Nations Human Rights Guidance on Less-Lethal Weapons in Law Enforcement*. Genf und New York: UN.

Composing AI

Formen, Ideen, Visionen

Miriam Akkermann

Abstract: *Die Vorstellung, dass Musik mithilfe von Computern erzeugt werden kann, fasziniert Komponierende seit Mitte des 20. Jahrhunderts. So hat die Einbindung entsprechender Techniken entgegen dem nun neu entfachten Interesse an KI in der Musik bereits eine längere Tradition. Gerade der Einsatz von Machine Learning-Techniken gibt den bekannten Visionen jedoch derzeit neuen Antrieb, verbindet eine rein KI-generierte Musik doch verschiedene Ziele, Wünsche und Erwartungen sowie eine Neugierde von Komponierenden und Publikum. Demgegenüber stehen eine mehr oder weniger komplexe und responsive technische Anlage der Programme sowie sehr unterschiedliche KI-Verfahren, deren realer Spielraum nicht immer den zugeschriebenen Freiräumen entspricht. In dem Artikel werden an drei Beispielen verschiedene Formen der Einbindung von KI-Verfahren in musikalische Arbeiten vorgestellt und angelehnt an eine Kategorisierung unterschiedlicher Systemkonzeptionen wird kurz diskutiert, welche ›Freiheiten‹ in KI-basierten musikalischen Arbeiten implementiert und welche zugeschrieben sind.*

Einleitung

Die Einbindung von Prozessen aus dem Bereich ›Künstliche Intelligenz‹ in musikalische Arbeiten scheint – zumindest mit Blick auf die Kombination der Schlagworte ›Musik‹ und ›KI‹ – ein eher junges Phänomen zu sein. Dies täuscht darüber hinweg, dass sich Komponistinnen und Komponisten bereits seit Mitte des 20. Jahrhunderts mit (Rechen-)Verfahren aus dem Forschungsbereich der ›Künstlichen Intelligenz‹ befassen und diese in verschiedenster Art und Weise in ihren künstlerischen Arbeitsprozessen verwenden. Bieten regelbasierte generative Prozesse umfassende Möglichkeiten, um eine schier

endlose Vielfalt an Ergebnissen zu produzieren, so üben gerade Ansätze eine große Faszination aus, die analytische und generative Elemente verbinden, versprechen diese doch nicht nur eine große Anzahl an verschiedenen, sondern viele, den Anforderungen angemessene Ergebnisse und ein Wissen über bis dahin vielleicht verborgene Regeln, die aus der nun größeren Anzahl an untersuchten Exempeln herausgearbeitet werden können. Die Einbindung solcher kombinierter Prozesse bringt jedoch auch verschiedenste neue Herausforderungen mit sich, die von sehr praktischen Fragen, beispielsweise nach den Ein- und Ausgabeformaten, bis hin zu komplexen Debatten über Handlungs- und Entscheidungszuschreibungen reichen und die implizit immer wieder um die dem Bereich KI ebenso wie dem der Musik inhärenten Frage nach Autor:innenschaft bzw. Autonomie und Einflussnahme einzelner Personen kreisen. Denn in der Beschreibung oder Präsentation der Musikstücke verschwimmen – gewollt oder ungewollt – oft die Handlungszuordnungen, insbesondere wenn Computerprogramme als Teil der Komposition eingebunden werden. Doch was kann ein KI-basiertes Programm in der Musik leisten? Wie sind diese Programme aufgebaut und in den kompositorischen Prozess sowie die Aufführungsumgebungen eingebettet? Und welche Handlungszuschreibungen, Erwartungen und Wünsche gibt es hinsichtlich der künstlerischen Autonomie und Kreativität?

Basierend auf einer schlaglichtartigen historischen Annäherung werden im Folgenden exemplarisch Formate, Ideen und Visionen vorgestellt, die zwar kein umfängliches Bild möglicher KI-Einbindungen in musikalische Arbeiten zeichnen, mithilfe derer es aber möglich ist, den Stellenwert von KI-Ansätzen in musikalischen Kompositionen differenziert zu beleuchten. Dies bietet wiederum einen Ausgangs- und Anknüpfungspunkt für eine weiterführende Debatte über KI in künstlerischen Prozessen und lädt zu einer Reflexion darüber ein, welche Unterscheidungen bei der Einbindung von KI-Konzepten möglich sind und welche Zuschreibungen sinnvoll sein können.

Frühe KI-Konzepte in kompositorischen Prozessen. Drei Beispiele

Kompositionen, die im weitesten Sinne Methoden aus dem KI-Bereich einbinden, werden aus musikwissenschaftlicher Perspektive im 20. Jahrhundert zumeist als Algorithmische Kompositionen bezeichnet und der Elektroakustischen Musik bzw. (heute vornehmlich) der Computermusik zugeordnet. Wurden bis in die 1990er Jahre vor allem statistische und stochastische Pro-

zesse sowie nicht-lernende generative Regelsysteme in die Arbeitsprozesse implementiert, beispielsweise zur sogenannten Partitursynthese (Erstellung von Notation) sowie zur unmittelbaren Klangerzeugung, kommen heute auch komplexe generative Verfahren in Musikanwendungen zum Einsatz, die auf Deep-Learning-Methoden basieren. Dies erfolgt oft mit dem Ziel, eine möglichst autonome, also vom Menschen unabhängig ablaufende Musik- bzw. Klangerzeugung zu ermöglichen. Zudem versprechen die neuen ebenso wie die frühen Techniken die Produktion einer größeren Vielfalt an Varianten als die, die ein Mensch im gleichen Zeitraum produzieren könnte. Die Zielstellungen, unter denen die verschiedenen Verfahren eingebunden werden, sind dabei ebenso unterschiedlich wie der Umfang, in dem die Komponistinnen und Komponisten in den Synthese- und Selektionsprozess eingreifen. Die Selektion, also die Einordnung und entsprechende Auswahl (zusammen)passender bzw. gewünschter Elemente aus den entstehenden (klingenden) Möglichkeiten, ist hierbei eine der zentralen Herausforderungen, die sich durch die gesamte Bandbreite musikalischer Kreationen mit generativen Verfahren zieht.

Eine der ersten musikalischen Arbeiten, bei der eine Einbettung früher KI-Techniken erfolgte, ist die llliac Suite (1955–1956) von Lejaren A. Hiller und Leonard M. Isaacson (vgl. Hiller und Isaacson 1958). Dieses Stück, das auch als die erste Computermusikkomposition angesehen wird, basiert auf einer computergestützten Partitursynthese (vgl. Supper 1995a: 967). Mithilfe des Computerprogramms ILLIAC I wurde die Notation für ein Streichquartett erstellt, wobei jedem der vier Sätze der Suite ein anderes Experiment zugrunde gelegt ist, mit dem Hiller und Isaacson erproben wollten, »wie mit einem Computer eine Komposition generiert werden kann« (ebd.: 972). Der erste Satz folgt 16 verschiedenen Regeln des Kontrapunkts; der zweite erweitert diese auf eine vierstimmige Polyphonie und die Möglichkeit, verschiedene Stile zu erhalten; im dritten Satz finden serielle Verfahren Verwendung, um Rhythmus- und Dynamikelemente zu erzeugen, wohingegen die Parameter Tempo und Metrum von den Komponisten bestimmt sind; im vierten Satz bestehen die »Kompositionsregeln […] [aus] einer Folge abhängiger Zufallsgrößen« (Supper 1995b: 74); die musikalischen Regelwerke werden hierbei durch sogenannte Markov-Ketten ersetzt, ein stochastisches Rechenverfahren, das dazu verwendet wird, Aussagen über das Eintreten eines Ereignisses in der Zukunft in Abhängigkeit von der Gegenwart zu treffen (vgl. Akkermann 2015: 31). Jede Note wird damit entsprechend ihrer Wahrscheinlichkeit in Abhängigkeit von der vorangegangenen Note berechnet. Die Erstellung der Noten erfolgte durch das Computer-

programm jeweils in einzelnen Experimenten, wobei mithilfe des Programms nicht nur eine einzelne Komposition, sondern eine ganze Klasse an Kompositionen berechnet werden konnte, die jeweils einem entsprechenden Stil folgten (vgl. Supper 1995a: 974). Die finale Selektion der generierten Stimmen und ihre Zusammenstellung in den jeweiligen Sätzen der Suite oblag jedoch den Komponisten und auch die Interpretation der Noten ist an Musiker:innen gebunden.

Während in der Illiac Suite weiterhin Menschen die computergenerierten Aufzeichnungen zum Klingen bringen, umfasst George E. Lewis' Voyager eine Kombination aus (digitalen) Klanganalyse- und Klangausgabeelementen, die es erlauben, (hörbaren) Klang zu erfassen, auf Regeln hin zu analysieren sowie dessen Erzeugung zu steuern. Das interaktive System Voyager, das 1987 erstmals auf der Bühne präsentiert wurde, ist auf eine Ko-Improvisation mit einem menschlichen Improvisierenden ausgelegt. Hierbei wird die erklingende Improvisation des Menschen mithilfe eines Pitch-to-MIDI-Trackings in Echtzeit in Tonhöheninformation umgewandelt und auf musikalische Regeln hin analysiert, um dann eine Art ›Antwort‹ darauf zu erzeugen. Dies geschieht basierend auf einem sogenannten set-phrase-behaviour, einem kombinierten Regelset, das sich aus zwei Elementen zusammensetzt: den aus dem zu hörenden Klang herausgearbeiteten Regeln sowie einem eigenen, unabhängigen Regelwerk, das nicht auf den analysierten Klang Bezug nimmt und innerhalb dessen von sogenannten zellulären Automaten Klangstrukturen produziert werden (vgl. Lewis 2000). Die Problematik, dass digitale Berechnungen von Klängen als Klangsynthese in Echtzeit in den 1980er Jahren noch eine größere technische Herausforderung darstellten, da sie an umfangreiche Rechnereinheiten mit limitierten Berechnungskapazitäten gebunden waren (vgl. Akkermann 2020: 125ff.), wurde umgangen, indem die Ausgabe der berechneten Tonhöheninformation per MIDI erfolgte, was die Klangausgabe mittels eines MIDI-gesteuerten (akustischen) Instruments (Disklavier) erlaubte – eine Lösung, die bis heute noch genutzt wird. Voyager ist damit (technisch) in der Lage, verschiedene Regelsets aus Klängen – Improvisationen – herauszuarbeiten und nachzuahmen sowie genuin eigene Klangsequenzen zu produzieren, die auf inhärenten Regelwerken beruhen. Lewis greift hierbei als Komponist selbst nicht aktiv in die Klangproduktion bzw. die Selektion der produzierten Sequenzen ein, sondern das Programm ist so konzipiert, dass es während einer Aufführung als eigenständige improvisierende Einheit in Interaktion mit einem Menschen musiziert. Eine Limitierung der Klangerzeugung ist nur durch die Auswahl des per MIDI gesteuerten Instruments

gegeben – dieses ist zumeist, wie auch bei dem Konzert 2020 in Berlin, ein Disklavier.

Ausgehend von der Idee, ein Computerprogramm zu erstellen, das einen ganz persönlichen musikalischen Stil (erkennen und er-)lernen kann, um dann – quasi mit ihm zusammen – neue Stücke zu erarbeiten, entwickelte der Komponist David Cope ab den 1980er Jahren das Programm Experiments in Musical Intelligence (EMI), auch Emmy genannt (vgl. Computer History Museum o.J.). Dieses generative System, das eigenständig beim Kompositionsprozess, nicht aber bei der Ausführung einer Improvisation mitarbeiten sollte, ist darauf ausgelegt, mithilfe von per MIDI getrackten Inhalten Regelsets abzuleiten und dann – ebenfalls per MIDI – notierte Musiksequenzen zu produzieren. Cope (1999: 79) beschreibt den Ablauf wie folgt:

> Experiments in Musical Intelligence composes by first analyzing the music in its database and then using the rules it discovers there to create new instances of music in that style.

EMI folgt damit im weitesten Sinne einem Prinzip, das auch heute bei KI-basierten Verfahren zum Einsatz kommt: Cope wählt bestimmte Kompositionen aus (database), aus denen von EMI jeweils die gemeinsamen Regeln herausgearbeitet werden (EMI ›lernt‹ die Regeln), die dann wiederum als Grundlage für die Berechnung von Tonhöhen und Tondauern dienen (die Regeln werden angewandt, um neue Musik im selben Stil zu produzieren). Die hierbei inhärente ›musikalische Intelligenz‹ sieht Cope in der Simulation musikalischen Denkens. Es wird, wenn man so möchte, ein Regelset erstellt, das die (charakteristischen) Merkmale eines Kompositionsstils, die aus einem gewählten Kompositionsdatensatz resultieren, nachstellen kann. Anders gesagt: Charakteristische Handlungskomponenten der Komponierenden werden auf komplexe Regelsets innerhalb von EMI übertragen (vgl. Cope 1992: 69). Dadurch entstehen musikalische Sequenzen, die einem vorgegebenen musikalischen Vorbild folgen, also bestimmte musikalische Charakteristika reproduzieren, ohne dabei jedoch exakte Kopien der Vorlagen zu sein. Die Wahl der im Datensatz enthaltenen Musikstücke kann hierbei stark das Resultat beeinflussen: Je ähnlicher die im Korpus enthaltenen Stücke sind und je prägnanter die für einen Stil als charakteristisch erachteten Elemente in Erscheinung treten, umso besser entspricht das Ergebnis den impliziten Vorgaben. Neben Copes eigenen Kompositionen dienen insbesondere Arbeiten bekannter Komponisten aus früheren Jahrhunderten als Vorlage für Musikproduktionen von EMI. Hierbei entstanden Musikstücke im Stil von unter anderem Béla Bartók, Johannes Brahms,

Johann Sebastian Bach und Frédéric Chopin, aber auch Wolfgang Amadeus Mozart und Sergei W. Rachmaninow. Während einige Stücke von Menschen aufgeführt wurden, ließ Cope andere per Disklavier erklingen, insbesondere dann, wenn die generierte Notation für Menschen unspielbar schien. Copes Rolle ist hierbei (immer) weniger die eines Komponisten, sondern eher die eines Programmentwicklers und Organisators, der den Inhalt der Datenbank wählt, die Resultate einem angemessenen Klangkörper – Musiker:innen oder Disklavier – zuführt und die entsprechenden Klangerzeugnisse präsentiert, erklärt und ggf. vermarktet. Das initiale Interesse von Cope scheint sich hierbei im Laufe der Zeit zu verlagern, weg von einem Programm, das ihn selbst unterstützt, hin zu einem, das musikalische Eigenheiten erkennen, erlernen und in ähnlicher Art und Weise (in Notation) produzieren kann.

Formen der Einbindung

Die angeführten Beispiele zeigen exemplarisch und in Abstufung drei grundlegende Einsatzrichtungen auf, mit denen Rechenprozesse in die Erarbeitung von Musikstücken eingebunden werden können:

- zur Analyse von vorhandenen Kompositionen (Notation) oder Audio bzw. live gespielten Klängen: diese werden in Echtzeit oder in vorgelagerten Prozessen auf Tonhöhen, Tonabfolgen und formale Strukturen hin analysiert, um daraus Regeln oder Muster abzuleiten;
- zur Partitursynthese: im kompositorischen Arbeitsprozess wird die Notation nach jeweils festgelegten Prozessen erzeugt; und
- zur Klangsteuerung bzw. -erzeugung: basierend auf Regeln, Mustern oder Notation werden Tonhöhenabfolgen (zumeist im Aufführungskontext) zum Klingen gebracht, beispielsweise mithilfe von digital gesteuerten Instrumenten oder direkt in Form von Klangsynthese.

Diese Formen der Einbindung können mit unterschiedlichen Rechenprozessen und in verschiedenen Kombinationen eingesetzt werden, wodurch Kompositionen entstehen, die sich – kompositorisch oder klanglich – auf algorithmische Prinzipien stützen, auch wenn dies auf den ersten Blick nicht immer zu erkennen ist. Die Komplexität der hierbei implementierten Rechenprozesse ist von der Art der Einbindung unabhängig und kann von einfachen Regelwerken bis hin zu ineinandergreifenden Systemen reichen.

Philippe Pasquier, der wissenschaftlich und künstlerisch zu Künstlicher Intelligenz forscht und auch selbst KI-basierte Systeme für musikalische Arbeiten entwickelt, unterscheidet mit Blick auf die Einbindung von generativen Prozessen in musikalische Arbeiten fünf Systemkonzeptionen nach deren praktischer Anlage:

- Systeme, die mit einem initialen Input autonom ablaufen, und solche, die keinen initialen Input benötigen;
- Systeme, die in Echtzeit ablaufen, und solche, die dies nicht tun;
- Systeme, die nach der Art des ›Wissens‹ – den Informationen, die den Rechenprozessen zugrunde liegen – gegliedert sind;
- Systeme, die nach der Art ihres Inputs bzw. Outputs – in Form von Zeichen/Notation, Audio oder hybrid – kategorisiert werden; sowie
- nichtkorpusbasierte Systeme, die ihre Ausgabe generieren, ohne musikalische Informationen als Eingabe zu erhalten, und korpusbasierte Systeme, die Wissen aus einem bereitgestellten Korpus von Musikstücken oder Ausschnitten aus musikalischen Arbeiten extrahieren (vgl. Pasquier et al. 2017: 6).

Diese grundlegenden Systemkonzeptionen beziehen sich auf unterschiedliche Umsetzungsebenen und umfassen die Kombination aus bestimmten Programmen und den im gesamten Set-up angelegten Interaktionsmöglichkeiten, wie auch in den vorgestellten Beispielen gut zu sehen ist. Während bei allen drei Beispielen die zugrunde liegenden Programme eines initialen Inputs in Form von Voreinstellungen oder vorgegebenen Inhalten bedürfen, ist nur Voyager in Echtzeit im Einsatz. Voyager greift wiederum zum einen – wie EMI – auf ein »Wissen« aus Klang bzw. Musikstücken zurück, und zum anderen sind in das Programm – vergleichbar mit ILLIAC I – eigenständige Regelwerke zur Generierung von Klangsequenzen eingebettet.

Etwas weniger eindeutig wird die Unterscheidung nach Input und Output: Zwar ist ILLIAC I insofern klar zu verorten, als das Programm einen symbolischen Output generiert, so bieten EMI und Voyager jedoch technisch MIDI-Schnittstellen zur Ein- und Ausgabe, wenn auch mit unterschiedlicher Intention. EMI bietet die Möglichkeit, die in MIDI formatierte Notation über ein MIDI-gesteuertes Musikinstrument wiedergeben zu lassen, das intendierte Ergebnis sind jedoch notierte Musikstücke, die mithilfe von MIDI erstellt werden. Bei Voyager hingegen wird MIDI als Brücke zwischen der Klangeingabe (Pitch-to-MIDI-tracking) und der Klangausgabe (z.B. MIDI-Piano) verwen-

det. Intendierter Input wie auch Output sind hier Audio bzw. hörbarer Klang. Dies ist in Pasquiers Typologie, in der zwischen In- und Output diskreter symbolischer Natur (zum Beispiel Musiknotation, MIDI-Dateien oder MIDI), Audio (in Form einer Tonspur oder Klang in Echtzeit) und hybriden Formen unterschieden wird, nicht präzise abbildbar. EMI und Voyager sind nach ihrer technischen Anlage beide Systeme mit symbolischem In- und Output – auch da ein direktes Pitch-Tracking ohne den Umweg MIDI in den 1980er Jahren, also zur Entwicklungszeit von Voyager, nicht problemlos in Echtzeit umgesetzt werden konnte. Dieses (technische) Problem stellt sich heutzutage nicht mehr, denn iterative wie interaktive Systeme können mit unterschiedlichen In- und Outputformaten umgesetzt werden. In Michael Youngs Piano Prosthesis (2008) wird beispielsweise der Audioinput direkt analysiert und mit einem komponierten generativen Musiksystem und Machine-Learning-Prozessen verbunden, um daraus sowohl Regeln abzuleiten als auch neue Klänge zu produzieren (vgl. Blackwell et al. 2012: 170). Ein in diesem Sinne hybrides System, das beide Arten als In- und Output verarbeiten kann, ist beispielsweise das unter anderem von Shlomo Dubnov an der University of California San Diego (UCSD) und am Institut de Recherche et Coordination Acoustique/Musique (IRCAM) entwickelte Audio Oracle, ein Meta-Creation-System, das dafür eingesetzt werden kann, verschiedene Variationen einer Audioaufnahme zu erzeugen (vgl. Dubnov et al. 2011).

Pasquier unterscheidet zudem zwischen zwei generellen Grundanlagen hinsichtlich der Beschaffenheit des Inputs: nichtkorpusbasierte und korpusbasierte Systeme. Nichtkorpusbasierte Systeme, wie sie durch die Einbindung der Programme ILLIAC I und Voyager entstehen, sind darauf angewiesen, dass vor ihrer Nutzung (dynamische oder andere) Werte verschiedenen Parametern zugeordnet werden. Bekannte nichtkorpusbasierte Systeme sind beispielsweise die algorithmischen Kompositionssysteme Projekt 1 und Projekt 2 des Komponisten Gottfried Michael Koenig. Für diese Systeme ist charakteristisch, dass die Personen, die die initialen Werte eingeben, mit dieser Eingabe immer auch Teile ihres jeweiligen musikalischen Wissens und/oder ihrer ästhetischen Beurteilungen beisteuern und somit den nachfolgenden Prozess – implizit und explizit – beeinflussen.

Korpusbasierte Systeme wie Copes EMI sind dagegen nicht zwingend an das Wissen der programmnutzenden Person gebunden, sondern benötigen eine (angemessene) Anzahl an Musikinformationen, entweder in Form von symbolischer Notation oder in Form von Audiodaten, um neue, den vorgegebenen Musikstücken ähnliche Ergebnisse zu produzieren. Während in nichtkorpus-

basierten Systemen die Ergebnisproduktion also von den formalisierten Angaben (z.B. Werteauswahl) einzelner Personen abhängig ist, sind in korpusbasierte Systeme vorprogrammierte Algorithmen implementiert, die auf der Grundlage bereits entstandener oder live entstehender Musik arbeiten, welche wiederum für diesen Zweck ausgewählt bzw. zusammengestellt wird.

Diese grundlegende Unterscheidung zwischen der Anlage nichtkorpusbasierter und korpusbasierter Systeme spiegelt sich, wie der Informatiker Alexander Waibel (2021) darlegt, implizit auch in der Unterscheidung zwischen dem Verständnis von ›Artificial Intelligence‹ (AI) der »Darthmouth Conference« 1956 und den heute häufig synonym verwendeten Begriffen ›Machine Learning‹ und ›Neural Networks‹ (›Deep Learning‹) wider. So sind für Waibel bei den algorithmischen Systemen, die in den 1950er Jahren diskutiert wurden, bereits Aspekte wie Automatisierung, Selbstoptimierung, Abstraktion, Zufälligkeit und Kreativität zu finden. Während in den 1980er Jahren sowohl statistische wie auch neuronale Machine-Learning-Ansätze entstanden, die aufgrund der noch beschränkten Rechenleistung vergleichbare Ergebnisse erzielten, erlauben die ›Neural Networks‹ ab den 2010er Jahren aufgrund der nun zur Verfügung stehenden Rechnerleistung die Verarbeitung großer Datenmengen (vgl. Waibel 2021). Dies macht sie gerade für korpusbasierte Systeme attraktiv.

Entsprechende Entwicklungen sind auch im Bereich der Musik zu sehen. So werden zwar weiterhin neue nichtkorpusbasierte algorithmische (Kompositions-)Systeme entwickelt, jedoch sind in den im 21. Jahrhundert entstehenden korpusbasierten Systemen zumeist neuronale Netzwerke und keine statistischen Netzwerke mehr zu finden. Beliebt sind hier insbesondere ›Recurrent Neural Networks‹ (RNNs), die nicht nur einmal mit dem vorgegebenen Korpus arbeiten, sondern nach dem Durchlaufen aller vorgegebenen Inhalte auf sich selbst rekurrieren und so mehrmalige Durchläufe vornehmen. Erzeugt wird dabei eine Art von Gedächtnis, da bereits gelernte Regeln nicht wieder vergessen (überschrieben) werden, sondern durch die wiederholten Durchläufe immer wieder ›in Erinnerung gerufen‹ und erneut angewandt werden können (vgl. dazu Abolafia 2016).

Ideen und Visionen

Die Art und Weise, wie Prozesse aus den Bereichen KI, besonders Verfahren von ›Machine Learning‹, in musikalische Systeme eingebettet werden, ist ebenso vielfältig wie die Ideen, Annahmen und Erwartungen, die mit einer

solchen Einbindung einhergehen. Für Pasquier unterscheidet sich beispielsweise »computational creativity« als programmiertechnischer Ansatz dadurch von anderen rationalen, problemlösungsorientierten KI-Prozessen, dass es sich hierbei häufig um Probleme handelt, für die zuvor keine optimale Lösung definiert ist (vgl. Pasquier et al. 2017: 2). Es geht also nicht darum, ein Problem möglichst schnell oder effizient zu lösen, sondern auch darum, dass die Ergebnisse eine spezifische Qualität in Bezug auf die Anforderungen aufweisen – selbst wenn dies ggf. mehr Rechenaufwand erfordert. Pasquier wie auch sein Forschungskollege Kıvanç Tatar beziehen sich dabei auf den Soziologen Herbert A. Simon, der KI im Kontext von Entscheidungen als »the science of having machines solve problems that do require intelligence when solved by humans« (Simon 1960; zitiert nach Tatar 2019: 56) definiert.

Dabei, so scheint es, überschneiden sich nun zwei unterschiedliche Zielvorstellungen im Hinblick auf Anwendungen in der Musik: zum einen der Wunsch, Prozesse, die in ihrer Ausführung zu komplex oder zu langwierig sind, um sie per Hand von einem Menschen umzusetzen, auf Computerprozesse auszulagern und dadurch eine Fülle an Möglichkeiten zu generieren, aus der der Mensch dann auswählen kann; und zum anderen das Interesse daran, Prozesse zu entwickeln, die so angelegt sind, dass sie – mehr oder weniger selbstständig – die Balance zwischen einer geforderten Regeltreue und einer erwarteten kreativ-künstlerischen Freiheit einhalten können und dabei (autonom) Ergebnisse produzieren, die so auch von einem Menschen hätten erzeugt werden können. Beide Vorstellungen beinhalten die Idee, dass mithilfe der Systeme Ergebnisse erzeugt werden, die in den vom Menschen a priori angelegten Bewertungsrahmen passen. Bei der zweiten Zielvorstellung schwingt zudem die Vision mit, dass ein Ergebnis entsteht, bei dem der Mensch keine weiteren Selektionen oder Anpassungen mehr vornehmen muss, das aber dennoch menschengemachten Ergebnissen qualitativ ebenbürtig ist.

Der Vision von einer weitgehend autonom ablaufenden, generativen Musikproduktion wird aktuell in vielen Projekten nachgespürt. Unter den aktuellsten Projekten stechen zwei hervor: Beethoven X (2021) und I'll Marry You, Punk Come (2019/20), ein Projekt, das zum Gewinnersong des AI Song Contest 2020 gekürt wurde.

Beethoven X. The AI Project verfolgte das Ziel, die 10. Sinfonie Ludwig van Beethovens, die dieser aufgrund seines Todes nicht selbst fertigstellen konnte, mithilfe von dafür entwickelten korpusbasierten Systemen zu vollenden. Die verfügbaren Sinfonien Beethovens sollten als Basis dienen, um daraus cha-

rakteristische Kompositionstechniken herauszuarbeiten und darauf aufbauend die bisher unvollendete Sinfonie fertig zu komponieren. Dieses Projekt, das verschiedenste Disziplinen – unter anderem Komposition, Musikwissenschaft und Informatik – zusammenbrachte und in Kooperation großer Institutionen und Unternehmen, darunter die Rutgers University, Google, Telekom und BMG, umgesetzt wurde, stellte mit großem medialen Aufwand heraus, welche Möglichkeiten für ein solches Unterfangen heutzutage zur Verfügung stehen. Es fällt jedoch auf, dass das Projekt zwar medial in Form von Trailern und einer Webseite präsentiert und viel Werbung für das Ergebnis – ein Konzert und ein Tonträger – gemacht wurde, die breitere Öffentlichkeit jedoch keinen Einblick in die Projektabläufe bekommt und auch über den Aufbau des verwendeten Systems sowie die darin implementierten (Selektions-)Prozesse und Lernmethoden sehr wenig bekannt ist.

Ein ähnliches Ziel, nämlich die automatische Erstellung eines Musikstücks und dessen (medienwirksame) öffentliche Präsentation, wurde auch in dem von Magenta 2019/20 ausgeschriebenen AI Song Contest 2020 verfolgt, bei dem nur Musikstücke eingereicht werden durften, die rein generativ und ohne menschliches Zutun entstanden. Ausgeschrieben von Magenta, eine seit 2016 bestehende Forschungsgruppe bei Google, die Tools und Plug-ins im Bereich ›AI Music Creation‹ entwickelt und diese kostenfrei zur Verfügung stellt, werden im Rahmen des Contests aber nicht nur die finalen Musikstücke präsentiert und prämiert, sondern auch die verwendeten Prozesse und Iterationen wurden in die Bewertung miteinbezogen; Eingriffe durch Selektionsprozesse und ein finales Editing der Klangergebnisse waren durch das Anforderungsprofil des Contests ausgeschlossen. Eine Jury und die über die Magenta-Webseite abstimmende allgemeine Hörer:innenschaft kürten I'll Marry You, Punk Come vom Team Dadabots x Portrait XO, bestehend aus CJ Carr, Zack Zukowski und dem Team Portrait XO, zum Gewinnersong. Das ursprünglich in den USA angesiedelte Projekt Dadabots, das 2012 initiiert wurde und bereits mit einem 24/7 streamenden generativen KI-Metal-Kanal auf YouTube auf sich aufmerksam machen konnte, arbeitete sowohl direkt mit Audiodateien als auch mit Konzepten aus text-to-speech-Anwendungen, darunter RNNs mit ›unsupervised learning‹. So entstehen in dem Song immer wieder Passagen, die stilistisch klar verortbar sind, wenngleich im Text, der ebenfalls generiert ist, wie auch in der Musik einige Wendungen durchaus in ihrer (stilistischen) Abfolge überraschen. I'll Marry You, Punk Come steht damit sinnbildlich für das, was oft bei der Einbettung von KI in Musik imaginiert wird, nämlich für ein Musikstück, das von einer artifiziellen Entität,

die im Sinne Pasquiers kreativ und in gewisser Weise auch autonom, also unabhängig von menschlichem Zutun oder Eingreifen, operiert, produziert wurde.

Reflexion

In den voranstehend angeführten Beispielen wird deutlich, dass in den meisten Fällen, in denen KI-Prozesse in Musikproduktionen implementiert werden, diese inhärenten, von Menschen gesteuerten Selektionsmechanismen unterliegen. Nicht nur müssen die hierbei verwendeten Prozesse auf verschiedene Arten und Weisen auf möglichst viel oder möglichst präzises Wissen von Menschen zugreifen, sondern es stellt sich dabei auch das Problem, dass eine KI nicht auf derart abstrahierte Bewertungskriterien zurückgreifen kann, wie Menschen es tun. Die Vision, durch die Einbettung von KI-Prozessen in kompositorische oder interpretative Abläufe Ergebnisse zu erhalten, die menschliches Handeln und menschliche Entscheidungen ersetzen, ist damit noch nicht erreicht. Selbst die Möglichkeit, große Datenmengen mithilfe Künstlicher Intelligenz verarbeiten zu können, ist laut Alexander Waibel (2021) zwar ein zentrales Alleinstellungsmerkmal, aber auch eine zentrale Unzulänglichkeit, denn im Gegensatz zum Menschen verfüge eine KI nicht über die Fähigkeit, von wenigen Beispielen zu lernen und aus den derart gewonnenen Einsichten abstrakte Schlüsse mit Blick auf eine Gesamtheit zu ziehen. Was können KI-basierte Ansätze in der Musik also leisten? Sind klare, systematisch erfassbare Ähnlichkeiten in Musikstücken vorhanden, so können korpusbasierte Systeme aus diesen Stücken durchaus generative Regeln ableiten, die zu ähnlichen klanglichen Ergebnissen wie im Auswahlkorpus führen. Ebenfalls können generative Systeme auf der Grundlage statistischer Berechnungen einer bestimmten Klanglichkeit ähnliche Höreindrücke produzieren. Jedoch korrespondieren die Zuschreibungen dessen, was ein ›Computer‹ kann und welche ›Freiheiten‹ eine KI hat, nicht zwingend mit den implementierten Regeln oder dem eingebetteten Wissen; vielmehr werden sie den Systemen von außen angetragen, was durch eine entsprechende Präsentation seitens der Erstellenden noch forciert werden kann. Sind die Mechanismen versteckt oder derart komplex, dass sie nicht (mehr) einfach nachvollzogen werden können, so scheinen Handlungszuschreibungen auch einfache Erklärmodelle zu sein, in denen sich Konzeption und Umsetzung mit Erwartungen und Wünschen hinsichtlich der Autonomie und der künstlerischen Kreativität dieser Systeme

mischen. In anderen Worten: angesichts zu großer Komplexität wird (gerne) zu – einerseits einfachen und andererseits erwünschten – Erklärungen gegriffen, bei denen der Maschine oder den Systemen Autonomie und künstlerische Kreativität zugeschrieben wird. Inwieweit es dabei wirklich möglich ist, alle implementierten Prozesse zu verstehen, ist ebenso zu diskutieren, wie die Auswirkungen, die es haben kann, wenn in größerem Maße Musik produziert wird, die bereits vorhandenen (und für die Lernkorpora ausgewählten) Musikstücken immer ähnlicher wird – eine Frage, die sowohl eine neue Debatte über Kanonbildung und ästhetische Werte eröffnet als auch erneut den Stellenwert der Selektion hervorhebt, die mit einer Einordnung der Inhalte (als passend oder unpassend) immer einhergeht. Dies zeigt noch einmal deutlich, dass die Einordnung und auch die finale (ästhetische) Bewertung eines Musikstücks weiterhin den (zuhörenden oder adressierten) Menschen obliegt, zumindest sofern die Musik, die entsteht, ein menschliches Publikum als Bewertungsinstanz intendiert.

Literatur

Abolafia, Dan. 2016. A Recurrent Neural Network Music Generation Tutorial. https://magenta.tensorflow.org/2016/06/10/recurrent-neural-network-generation-tutorial. Zugegriffen: 22. November 2021.

Akkermann, Miriam. 2015. Instrument oder Komposition? David Wessels Contacts Turbulents. In Reflexion – Improvisation – Multimedialität. Kompositionsstrategien in der Musik des 20. und 21. Jahrhunderts, Hg. Christian Storch, 95–108. Göttingen: Universitätsverlag Göttingen.

Akkermann, Miriam. 2017. Zwischen Improvisation und Algorithmus. David Wessel, Karlheinz Essl und Georg Hajdu. Schliengen: Edition Argus.

Akkermann, Miriam. 2020. (Musik)instrument (im) Computer. In Brückenschläge zwischen Musikwissenschaft und Informatik. Theoretische und praktische Aspekte der Kooperation, Hg. Stephanie Acquavella-Rauch, Andreas Münzmay und Joachim Veit, 125–140. Detmold: Musikwissenschaftliches Seminar der Universität Paderborn und der Hochschule für Musik Detmold.

Blackwell, Tim, Oliver Bown und Michael Young. 2012. Live Algorithms: Towards Autonomous Computer Improvisers. In Computers and Creativity, Hg. John McCormack und Marc d'Inverno, 147–204. Heidelberg: Springer.

Computer History Museum. o.J. Cope, David Oral History. Artefact Details. ht tps://www.computerhistory.org/collections/catalog/102738612. Zugegriffen: 13. März 2021.

Cope, David. 1992. Computer Modeling of Musical Intelligence in EMI. Computer Music Journal 16: 69–83.

Cope, David. 1999. Facing the Music: Perspectives on Machine-Composed Music. Leonardo Music Journal 9: 79–87.

Dubnov, Shlomo, Gerard Assayag und Cont, Arshia. 2011. Audio Oracle Analysis of Musical Information Rate. Proceedings of the Fifth IEEE International Conference on Semantic Computing, 567–571. https://doi.org/10.1109/ICSC.2011.106.

Goodyer, Jason. 2021. How an AI Finished Beethoven's Last Symphony and What That Means for the Future of Music. BBC Science Focus Magazine. ht tps://www.sciencefocus.com/news/ai-beethovens-symphony/. Zugegriffen: 29. April 2022.

Hiller, Lejaren A. und Isaacson, Leonard M. 1958. Musical Composition with a High-Speed Digital Computer, JAES 6 (3): 154–160. https://www.aes.org/e-lib/browse.cfm?elib=231.

Hoffmann, Peter. 2000. The New GENDYN Program. Computer Music Journal 24: 31–38.

Hoffmann, Peter. 2002. GENDY3 von Iannis Xenakis: Eine Höranalyse. In Konzert – Klangkunst – Computer. Wandel der musikalischen Wirklichkeit, Hg. Institut für Neue Musik Darmstadt, 255–272. Mainz: Schott.

Koenig, Gottfried Michael. 1979. Projekt Eins – Modell und Wirklichkeit. Musik und Bildung 11/12: 752–756.

Koenig, Gottfried Michael. 1983. Integrazione estetica di partiture composte mediante elaboratore – Segmente 99–105. Bollettino 3 – Bollettino del Laboratorio 27/28: 29–34.

Koenig, Gottfried Michael. 1993a. Zu Projekt 1 – Version 1. In Ästhetische Praxis. Texte zur Musik 1968–1991. Quellentexte zur Musik des 20. Jahrhunderts, Hg. Musikwissenschaftliches Institut der Universität des Saarlandes, 12. Saarbrücken: Pfau.

Koenig, Gottfried Michael. 1993b. Tabellen, Graphiken, Klänge. Ein Computer-Programm für kompositorische Strategien: Projekt 3. In Ästhetische Praxis. Texte zur Musik 1968–1991. Quellentexte zur Musik des 20. Jahrhunderts, Hg. Musikwissenschaftliches Institut der Universität des Saarlandes, 315–319. Saarbrücken: Pfau.

Koenig, Gottfried Michael. 1999. Project 1 Revisited: On the Analysis and Interpretation of PR1 Tables. In Navigating New Musical Horizons, Hg. Otto Laske, 53–70. Santa Barbara: Praeger.

Lewis, George. 2000. Too Many Notes: Computers, Complexity and Culture in »Voyager«. Leonardo Music Journal 10: 33–39.

Mirianda, Eduardo. 2007. Cellular Automata Music: From Sound Synthesis to Musical Forms. In Evolutionary Computer Music, Hg. Eduardo R. Miranda und John Al Biles, 170–193. London: Springer.

Pasquier, Philippe, Arne Eigenfeldt, Oliver Bown und Shlomo Dubnov. 2017. An Introduction to Musical Metacreation. Computers in Entertainment 14: 1–14. https://doi.org/10.1145/2930672.

Schiff, Joel L. 2007. Cellular Automata: A Discrete View of the World. Hoboken: John Wiley & Sons.

Simon, Herbert A. 1960. The New Science of Management Decision: The Ford Distinguished Lectures. New York: Harper & Brothers.

Supper, Martin. 1995a. Computermusik. In MGG – Die Musik in Geschichte und Gegenwart, Hg. Ludwig Finscher, Sp. 967–987. Kassel und Stuttgart: Bärenreiter.

Supper, Martin. 1995b. Elektroakustische Musik, Elektroakustische Musik ab 1950, Live-Elektronik. In MGG Online, Hg. Lorenz Lütteken. https://www.mgg-online.com/mgg/stable/14626. Zugegriffen: 15. Februar 2020.

Supper, Martin. 1997. Elektroakustische Musik & Computermusik. Hofheim: Wolke.

Tatar, Kıvanç. 2019. Musical Agents: A Typology and State of the Art Towards Musical Metacreation. Journal of New Music Research 48: 56–105.

Waibel, Alexander. 2021. AI in the Service of Humanity. https://www.youtube.com/watch?v=HwtrWwd8WSY. Zugegriffen: 29. April 2022.

Computational Topologies

Can Artificial Neural Networks Be Normative Models of Reason?

Limits and Promises of Topological Accounts of Orientation in Thinking

Lukáš Likavčan & Carl Christian Olsson

Abstract: *The history of thinking about thinking is populated by numerous attempts to model reason in topological terms. Amongst them, the prominent place is occupied by Immanuel Kant's explanation of thought's need to restrain its own exercise by means of an analogy between geographical orientation (modeled on the human body) and orientation in thinking. As natural as his analogy might seem, the first part of this chapter aims at deconstructing Kant's attempt as both replaceable and constraining, and at proposing a possibility of alternative topological accounts of thinking. Hence, while endorsing the utility of spatial models, we call for an unbinding of the parochial connection between thinking and the form of the human body implicit in historical topological models of reason. For this reason, in the second part of the chapter we suggest that the topological framework embodied by Artificial Neural Networks (ANNs) can be used as an alternative to formulate such a model of thinking, based on their commitment to dimensionality and use of space as an active, dynamic and transitory element. Rather than arguing that ANNs somehow think, we suggest that they offer a mirror that lets humans look back at themselves and construct their thinking differently. We conclude by proposing that the benefit of looking in this mirror is open-ended and twofold: (1) It divorces our image of thinking from anthropomorphism, and (2) it offers a normative model of reason with potentially practical consequences for how humans act.*

1. Introduction

The intellectual history of space, at least in the Western tradition of mathematics, has been that of a concept long subjugated to set theory, starting with the Euclidean definition of space as a set of points. However, the last two centuries marked a departure from this perspective, positing logic or algebra as derivative from the operations with spatial or space-like concepts (Plotnitsky 2012: 355). It started with breakthroughs in geometry initiated by Galois, Abel, Poincaré and Riemann, leading to topology – a discipline concerned with an abstract inquiry into the nature of space – being instituted as the general discovery of surprising relations between distant fields of mathematics (Zalamea 2012). Topology also became the generic means of identifying and treating abstract structures. This made possible a true ontological Cambrian explosion in contemporary mathematics, illustrated by the success of category theory as potentially the new foundational programme for mathematics. The aim of this chapter is to take stock of how this topological turn has reverberated beyond the disciplinary confines of mathematics, returning space to the center of philosophy and computation.

In philosophy, Kant famously argued that space is an a priori condition of experience, meaning that no objects could be cognizable without being located in space (e.g. Kant CPR B73). But here we will discuss one of Kant's other intriguing insights – namely, an analogy between orientation in thinking and geographical orientation (Kant AK 8: 133–146). This analogy will help us explore intuitions about the relationship between topology and cognition, which we will valorize in the discussion of artificial neural networks (ANNs). ANNs represent today the most successful cluster of computational technologies that fall under the rubric of artificial intelligence (AI). They excel in pattern recognition based on sound or visual data, and they find their application in the processing of voice commands, biometrics, text translations etc. Apart from that, the most recent branch of ANNs, such as GPT-3 or DALL-E 2, is capable of automated generation of new text, sound or visuals, which find numerous applications in the sciences, arts and different industries.

What made these highly abstract technologies into such powerful tools of statistical inference is their topologization of computation, which lies in representing real-world data in n-dimensional space and applying regression analysis to arrive at an adequate way of modeling them. This may be a starting point for updating the Kantian topological metaphor of orientation, which in turn may lead to: (1) a reassessment of critiques of instrumental, enumerative ratio-

nality that ANNs are claimed to embody; and (2) a philosophically warranted alternative to understanding orientation in thinking. The latter may enable us to uncover some normative implications of the computational topology of ANNs, without providing a normative theory of topologies in general.

2. Kant's geographical model of orientation in thinking

In his essay What Does It Mean to Orient Oneself in Thinking? (Kant AK 8: 133–146), Kant presents an analogy between geographical orientation (of one's body in a physical space) and orientation in thinking. His exposition begins with a description of geographical orientation, which includes a subjective ground of differentiation structured roughly according to the symmetry of the human body in terrestrial space.

> In the proper meaning of the word, to orient oneself means to use a given direction (when we divide the horizon into four of them) in order to find the others – literally, to find the sunrise. Now if I see the sun in the sky and know it is now midday, then I know how to find south, west, north, and east. For this, however, I also need the feeling of a difference in my own subject, namely, the difference between my right and left hands. I call this a feeling because these two sides outwardly display no designatable difference in intuition (Kant AK 8: 134–135, original emphasis).

Kant supports this claim by an example of navigation of one's body in the dark. Even if one is present with no objective data about space owing to lack of visual stimuli, orientation is still possible by means of a subjective ability to distinguish between left and right (Kant AK 8: 135). For Kant, a subject trying to navigate the world has the innate ability to notice the structure of its body in relation to the surrounding world. Analogously, reason also needs a navigational principle of its own once it leaves the turf of experience to venture into the speculative realm. Orientation in thinking thus refers to a "reason's feeling of its own need" (AK 8: 136) to ground its proper use on rational faith similar to subjective differentiation of left and right. Prudent reason chooses to restrain itself according to maxims because it realizes that over-enthusiasm would amount to squandering its freedom to think (Kant AK 8: 144–146). He writes in a footnote: "Thus to orient oneself in thinking in general means: when objective principles of reason are insufficient for holding something true, to determine the matter according to a subjective principle" (AK 8: 136).

What is remarkable about Kant's definition of orientation in thinking as a paradigm of rational self-constraint is that it uses an empirical case as its privileged model before the actual definition is given. We learn about what orientation in thinking really means in a footnote only after the imagery of finding the sunrise has been presented. Although the meaning of orientation in thinking is clear enough, Kant's way of presenting it seems driven by a temptation to talk about thinking in terms of spatial analogies. The reader is invited to imagine reason standing somewhere at dawn with a compass, even though thinking in general, and reasoning in particular, has little to do with the sort of space Kant refers to in his first exposition of orientation, quoted above (AK 8: 134–135).

But if there is no intrinsic relationship between thinking and geographical space, what was Kant's justification for talking about reason's behavior in these terms, which was probably a didactic choice? Could it be that there are other models to which thinking can refer in interpreting its orientation? To answer these questions, let us first examine the meaning of Kant's concept of orientation in more detail. We will gloss the point he makes about the distinction between left and right, and focus on the consequences of using this concept of orientation as a model for reason's need to constrain itself.

Kant bases his analogy on the space of sensomotoric orientation. In discussing the difference between left and right he has in mind a chiral structure or distinction between non-superimposable mirror images such as our hands (which appears to have been a small obsession of his). Earlier in his career he believed that incongruent counterparts were a means to critique Leibniz's concept of space as fully explicable by the relations between its 'occupants'. For Kant at this time, only an absolute space, as proposed by Newton, could make it possible to identify incongruent counterparts (e.g. Kant AK 2: 379). His dissatisfaction with the Leibnizian notion of space was a significant step toward the transcendental aesthetic and the entire critical system, where a Euclidean space would become one of the subjective conditions of objective knowledge, as constitutive of our form of outer sense. The development of Kant's thought points to tensions in the concept of orientation he uses in his essay on the topic:

1) If space is a form of sensibility in which one can orient oneself, it cannot be exhausted by relations between its contents because otherwise (following Kant's earlier argument) one could not identify left and right. This is the subjective 'feeling' which Kant refers to when he talks about chiral structures in the orientation essay and elsewhere.

2) Orientation utilizes the subject's ability to tell or 'feel' the difference between left and right in the structure of its sensibility, which in Kant's example, however, can only be an experiential act because hands are given as objects in space.

Kant's critical account of space as the form of outer sense delimits how we interpret the concept of orientation he discusses in the essay. The concept of orientation is understood as an empirical act of finding one's way in a space divided in four cardinal directions. The image of a compass is a bit misleading, however, since Kant's point is that orientation depends on an unexplained "feeling of a difference in my own subject" (Kant AK 8: 135) that persists even without perceiving any objects. The subject is equipped with a kind of intrinsic compass. But it is hard to see how there can be no objects at all, since the essay makes references to at least one body: the Kant-shaped object at the center of his geographical space. Add to this that the cited ability to differentiate between left and right has allegedly been imparted to us by "nature" (AK 8: 135), and it is clear that the concept of orientation refers to actual, embodied experience. Yet it is transcendentally constrained by the constitution of the human form of sensibility. So if reason can orient itself analogously to how we orient ourselves in darkness, this should – as far as Kant ought to think – first of all be understood as an analogy between a subjective principle of reason and the orientation of a bilaterally symmetric body represented in a Euclidean space. That is, the operative concept of orientation is tailored according to Kant's philosophical interpretation of an empirical case of orienting himself, which is in turn based on the concept of space that he thinks we are entitled to. We will refer to this as 'geographical orientation'.

Geographical orientation is the basis of Kant's construal of thought's need to assume things that it cannot know on objective grounds. But note here that Kant does not presuppose that thinking itself is spatial – in his architectonics of human cognition, space is a form of sensibility, and thus belongs to the apprehension of what is exterior to the subject. Hence, to orient oneself in thinking cannot mean in the space of thought but rather in the case of thinking (as opposed to experience). The problem is that Kant assumes the geographical model is suitable for explicating reason's need. At stake here are the languages, images and models we use to understand thinking. On the one hand, we think that Kant skillfully demonstrates the value of using a spatial concept such as orientation to think about thinking, but on the other hand, it is not clear to us why orientation in thinking must be understood with reference to the ge-

ographical model. If it is not necessarily so, it is possible that thought's self-orientation can be construed in other terms, through other topological analogies or empirical prostheses. Such alternatives may help us construct different norms for how thinking should behave with regard to itself, or at least open the floor for new lines of inquiry about the behavior of thought, unbinding reason from the common-sense model it finds in geographical orientation.

3. Topology of reason after Kant

Kant's analogy effectively discloses a model based on geographical orientation, where orientation functions as a heuristic to draw conclusions about what reason is justified in doing, indeed even what it needs to do. This means that the model has normative consequences. Though Kant may be right to explain how reason should behave using his orientational standard, orientation is understood as a geographical term which for Kant appears to have a subjective structure as a matter of objective fact – as demonstrated by his argument about incongruent counterparts that form the basis for navigation even in darkness. But why should reason 'feel' a subjective need that is explicable through this concept of orientation? In the next few paragraphs, we will argue that for the purposes of philosophy, there are benefits to considering alternative models which, on the one hand, concur that a topological explication of the standards of reason is meaningful but, on the other, are prepared to reject Kant's interpretation of what this means.

An answer to why the geographical interpretation of orientation appears privileged prima facie is a historical affinity between geographical space and reason. One of the more tantalizing ideas about the relationship between reason and space is developed by the time-space sociologist Bernd Schmeikal-Schuh, according to whom there is a structural homeomorphism between an original concept of orientation which reflects the structure of perceptual space and the operations of Boolean algebra (Schmeikal-Schuh 1993). Logic, it is hypothesized, may have been learned as a result of an increasing historical awareness about orientation. As Schmeikal-Schuh puts it plainly, "the laws of thought may have developed out of the original structure of orientation in space" (1993: 130). Here we have the outlines of a genetic account capable of explaining the basis of Kant's analogy. Even if his explanation cannot be verified as an actual historical process, Schmeikal-Schuh shows how reason can feel a need to orient itself not because it is spatial per se but because its fundamental

"grammar" is understood in terms which are translatable into a concept of orientation in space. In this respect, Schmeikal-Schuh's argument contributes two relevant points: 1) a claim about how thought acquired a particular logical structure; and 2) a justification of reason's felt need to orient itself in analogy to orientation in space. Although Schmeikal-Schuh does not understand logic in quite the same way as Kant, he nevertheless provides a cultural history to explain how thought can have found a model in geographical orientation.

Still, there is nothing to guarantee that thinking can be exhausted by this particular concept of orientation. The problem of such an interpretation, already at work in Kant, is an assumption that perceptual space, even if historically relevant, holds normative force for understanding orientation in the case of thinking. Even if reason feels a certain need to orient itself, why should orientation be understood on the basis of a particular logic? Even in the dark of night reason feels the need to constrain itself according to a concept of orientation which is drawn from an empirical case. A grammar of reason may well be inherent in the transcendental structure of space but its proper use is taught from an empirical case of geographical orientation. The problem is reminiscent of the one produced by Gilles Deleuze's criticism of a "dogmatic image of thought" (1994) which models its own nature on the basis of an empirical case. A thought which constrains itself in this manner is engaged in its own becoming-docile, a concrete process of learning what is proper conduct based on a privileged empirical case. Faced with darkness, it is as if the source of normativity in Kant's analogy comes from an assumed principle that thinking should acquiesce to that which is most familiar, namely an image of the human body in a Euclidean space.

Against the historicist rationale presented by Schmeikal-Schuh, an interlocutor might propose an essentialist alternative: that it is space as the form of outer sense, which is a precondition for experience, that is at work in the analogy – such that thought is justified in portraying itself in accordance with geographical space, not out of familiarity but because it refers to the necessary space presupposed by experience in general. The interlocutor might argue that not only does geography give us one model for orientation, it supplies a model which is derived from the structure of receptivity and is therefore basic. But the cost of such an argument is to bind reason's sensibility so that acquiescence to the model is a guarantor rather than a mere comfort: no acquiescence, no ability to understand orientation in the case of thinking. The consequence of the essentialist alternative would be to shackle freedom of thinking to the human sensory apparatus – a claim that Kant would not have supported, careful as

he was to distinguish between humans and the broader concept of rational beings (see for example AK 4: 412). Still, on the essentialist account there is just no more fundamental case than geographical orientation on which reason could model its felt need. But – and this is a crucial point – the interjection hinges on Kant being right about what sort of space is necessary for experience, or else it is hard to see why geographical orientation should be privileged.

In fact, it is no longer clear that it is, even on the terms of Kant's own argument. For example, in some of his earliest work Rudolf Carnap attempted to disambiguate between different meanings of space in philosophy, physics and mathematics. One of his conclusions was that Kant would have needed a topological concept of space in order to find a necessary ground for experience: a physical or projective space would simply not suffice (Carnap 1922). If Carnap is right, one could remain broadly Kantian but distinguish between the represented space of geographical orientation and the space necessary for experience. In the case of experience, we would be entitled to a loftier notion of space than that proscribed by Kant. In that case we might be able to accept the possibility of another normative account of reason's orientation based on a concept of space if we hold to such a topological notion that is irreducible to metrical or projective properties. In other words, it is the geographical orientation model that proves to be limiting insofar as geographical orientation foregrounds a concept of physical space with Euclidean roots. So even if it is historically warranted, Kant's image of reason's proper conduct derived from orientation in space appears to be, on different accounts, neither exhaustive of thought's potential nor (even to the Kantian acolyte) adequate to describe the necessary conditions of experience. In both cases, an image of reason's orientation does not need to be beholden to what Kant construed as geographical orientation. It is plausible to think there could be other models for orientation in thinking that are not restricted to an image of a bilaterally symmetric animal at the origin of a Euclidean projection.

4. Mathematical topology and ANNs

To recap, even though Kant does not say that thinking is spatial, he demonstrates how to think about thinking in spatial terms: that a topological modelling of thinking is possible and that it can be insightful. From Schmeikal-Schuh and Carnap, we can see now that the model of orientation of thinking does not have to be bounded (and should not be bounded) by the metric space

of geographical orientation. This opens up possibilities of integrating discoveries in the field of topology since Kant wrote his contribution. In other words, we are seeking a prosthetic model that would remain topological while allowing us to escape the boundaries of the models of orientation based on metric, projective space in general, including its narrowly Kantian version. Such a topological model of orientation would be especially welcomed if it already manifested some success in orienting other modes of conducting inferences beyond the case of human thinking – a sort of proof-of-concept for a topological model that aspires to be scaled up to the level of an alternative normative project. Interestingly, there is one such case at hand in the twenty-first century – the emergence of ways of apprehending the world explicitly based on topology in artificial neural networks (ANNs). The calculus behind ANNs relies on discoveries associated with the topological turn in computation, as explained by authors such as Pasquinelli (2019) and Cavia (2022), and more deeply with the topological turn in mathematics (Zalamea 2012), that resulted from liberating the notion of space from the confines of set theory. As Cavia and Reed reiterate in this volume (pp. 353–365), instead of space being treated as a set of points with additional structure, the structure becomes treated as inherent in space, which allows for an alternative approach to topology, as a general study of abstract structures. Or to put it differently: "Topological treatment of space is always a means of gleaning a structure latent in the space" (Cavia 2022: 112).

Most importantly, Zalamea associates this topological turn with a synthetic, constructive style of conducting mathematics, which pays special attention to "incessant pendular processes of differentiation and reintegration" (2012: 123). This style takes its guidance from Peirce's pragmatism, which "benefits from an attentive examination of the contaminations and osmoses between categories and frontiers of knowledge so as to articulate the diversity coherently" (Zalamea: 111). Such an orientation of thinking toward gestures of integration without homogenization proved to be fruitful in constructing new frontiers of mathematical discoveries, especially since Alexander Grothendieck's work on algebraic geometry and topoi theory (Zalamea: 133). Grothendieck established a relativistic perspective on the nature of truth in mathematics, which becomes indispensable from identifying its context (or locale, topologically speaking – see Cavia and Reed (2022) in this volume). This move allows proper maintenance of diversity/discontinuity within the field of mathematics, while at the same time facilitating categorial transitions between distant domains of this field (e.g. between geometry and logic), thereby leaving enough space for possible reintegration of what is diverse/discontin-

uous. Turning these discoveries into a normative model would mean that any integrative pursuit is linked in a pendular fashion to enunciation of a new platform of diversification – a maxim perhaps too abstract at the moment, yet already affording a glimpse of some of its political consequences. Our belief is that through close examination of the topological aspects of ANNs, we can arrive at a constructive project of reason (i.e. of its orientation) that would adopt these normative stakes endemic to mathematical topology.

To turn our attention to the case of ANNs, their explicit use of topology breaks down to the employment of dimensionality in their representation of data. ANNs represent parsed information from input data in a high-dimensional space, where the number of dimensions depends on the number of individual neurons in the network. Each dimension thus maps a unique subspace of the total datascape incommensurable to other subspaces within the datascape. Take an example of a simple dataset called MNIST, composed of images of handwritten numbers, each image of size 28x28 pixels. A neural network used to process this database would need 784 neurons on the input layer, meaning also 784 dimensions to represent the input data. That is a fairly high-dimensional space, but it would be just a beginning: to obtain meaningful results (e.g. to use ANN as an autoencoder capable of classifying handwritten digits with values from 0 to 9), one would need to add hidden layers and of course also its output layer, which would multiply the number of dimensions, ending with a number of dimensions with at least five or six digits, depending on the number of hidden layers and the number of neurons within each of them.

Still, such mapping of real-world data into its abstract, topological representation is an operation of reduction, which lies in rendering the continuous nature of reality (expressible as a topology with infinite dimensions) into its segmented, hence computable version. The resulting n-dimensional structure, known also as latent space, can thus be approached as a model of reality that uncovers the latent structure present in the data, and uses topology to figure out the relations between elements in the dataset. The training of such a model to successfully detect desired patterns, when presented with new inputs, also exhibits explicit topological aspects. As Cavia states:

> Manifold learning involves smoothing data into a continuous surface, as a planar representation on which locality can be expressed between points, an operation only possible via dimensionality reduction of real world data. The real is cast as a tangled complex of manifolds, and the ability of AI to

> recognize patterns becomes a matter of fitting a curve to a topology of points in this geometric interpretation. (Cavia 2022: 142)

Our conjecture is that procedures within the latent space of ANNs lead to a widening range of topological models suitable for interrogating how we describe human thinking, without resorting to a claim that ANNs "think" in some sense. Just as geographical orientation was a useful model for Kant in explaining the structure of orientation in thinking which turned out to have normative consequences (without assuming that by walking one somehow engages in an act of thinking), so the n-dimensional latent space of ANNs provides an alternative, topological rendering of inferential procedures – but with what consequence?

5. Beyond the critique: The project of topological reason

The consequence of topology's prevalence over set theory in setting out the foundations of contemporary mathematics is aptly described by Vladimir Voevodsky. While reiterating basic components of Zermelo-Fraenkel set theory with the axiom of choice, he notices that set theory is "based on the human ability to intuitively comprehend hierarchies" (Voevodsky 2014). Yet category theory – the most successful project of constructing topological foundations for mathematics – departs from this intuition at its fundamental layer, leading to an ability to appreciate non-hierarchical structurations and transits that are better suited to grasp the continuous nature of the real. No wonder, then, that Grothendieck – as one of the pivotal contributors to category theory – has been infusing mathematics with his anarchical political views (Plotnitsky 2012: 365). His method of "experimental mathematics" allows for the flourishing of "horizontal interactiveness" between logics emerging from the multiplicity of conceivable topologies (2012: 363, 367). Such a non-hierarchical account informs another approach to constructing and setting relations between abstract objects, one that spills over from the hierarchical intuition of set theory toward a project of reason that is not primarily reliant on the gestures of subsumption. On top of that, such a topologically informed model of reason lacks the disadvantage of Kant's empirically loaded model based on geographical orientation, since it removes familiarity with projective, metric space as the source of normative traction.

Subsumption, hierarchy, stiffness of categorial matrices – these are pathologies associated with instrumental, enumerative rationality criticized in contemporary writing on AI by authors such as Pasquinelli (2017a, 2017b). The template of this critique can be found in Heidegger's analysis of modern technology, which claims that its essence lies in enframing: a mode of epistemic rendering of the real that aims "to reveal the real, in the mode of ordering, as standing-reserve" (1977: 10) – as something that is "ordered to stand by, to be immediately at hand" (1977: 8). To put it differently, Heidegger here manages to link the optics on the real that modern technology presupposes with a mode of administration of the real. The administrative goal is to transform the real according to the template of the "standing reserve" that enframing by modern technology suggests. Similarly, it is claimed today that ANNs serve a function of classifying real-world objects (and subjects) into stereotypical categories (e.g. based on race or gender), thus following that tradition of enframing, otherwise explicable as instrumental, enumerative rationality. Hence it is widely assumed that these advanced practices of "marking territories and bodies, counting people and goods" (Pasquinelli 2019) solidify distinctions based on the aforementioned stereotypical categories, and perpetuate forms of social, political and economic control.

While it is surely true that no technology is politically neutral, our argument is that it may be too hasty to say that ANNs are somehow inherently tainted by the consequences of their emergence and use in a given socio-historical context. Such a critique numbs us to the deep topological turn propagated by AI. Although the political-economic and media-theoretical critique of AI correctly focuses on how ANNs may be used for purposes of social division and control (e.g. Crawford/Joler 2018; Srnicek 2017; Pasquinelli 2017b), it misses the point if it puts too much focus on the computational logic of ANNs, because this logic is not predestined for this purpose. Hence, as an alternative to the approach that binds the computational logic of ANNs to a given socio-historical context, we propose to think about it as a guide to a generic topological model (or image) that can be projected on the case of human thinking to yield various results. By opening up the space for diversification, this idea bears similarity to projects such as Yuk Hui's cosmotechnics, which also conditions technological rationality by a more general topological structure that goes by the name of cosmos (Hui 2016: 18–33). In this sense, cosmos is an example of a concept that unifies spatiality and normativity, referring to the ancient Greek meaning of the term that captures universe as a total "world-space" together with its normative dimension – the cosmic order

– and the aesthetic dimension of the perfectness of this order (Pascal 2014: 1218).

In a similar vein to the anthropological discussion of cosmology that works for Yuk Hui, we propose here that the computational logic of ANNs can function as a sort of philosophical lever, letting us gaze back at the normative standards that pervade our thinking. Such leverage can come from reflecting on the topological aspects of how ANNs operate and feeding these reflections back into the conversation about human reason, thus surpassing those critiques that lock the future of AI technologies firmly into the space of instrumental, enumerative rationality. Hence, to say that reason is topological is a constructive act, not an ontological uncovering of reason's essential affinities. It may therefore be better to speak of reason becoming topological, by analogy with the becoming topological of computation (as observed by Reed, Cavia or Pasquinelli) or culture (see Lury/Parisi/Terranova 2012). Such a constructive approach also carries the Kantian legacy further by reworking it: Kant's model of orientation in thinking was also constructive in how it posited thinking to be orientable according to a principle of rational faith, set out as an imperative. In the case of topological thinking, an alternative imperative can be found in reason's scope for excess – excess both of itself and of the continuous nature of reality it models (Plotnitsky 2012: 367–368). That contrasts with the docile vibe of Kant's model of orientation. Furthermore, the constructive project under discussion here aligns with an ambition of synthetic mathematics to maintain a pendular movement between integrative and differentiating acts of thought.

A hint toward such a constructive project that operationalizes the topological model of reasoning endemic to AI has been put forward by thinkers such as the aforementioned Yuk Hui or Ramon Amaro. The latter claims in this respect:

> what I'm thinking through, especially in terms of machine learning and artificial intelligence, is the potential for resistance within the spaces in between. That's why I was saying […] that this type of revolution, for me, ultimately comes down to revolution of us with ourselves, and actually how we consider our own self-actualisation in accord with these technologies. (Amaro/Hui/Dasgupta 2021: 55)

Beyond abolishing AI technologies as means of archaically cataloging different "genres of being human" (Wynter/McKittrick 2015: 31–32) lies the task of understanding how these genres get actualized – and even multiplied – in confrontation with such technologies. In this pursuit, unlocking the topological aspects of ANNs' computational logic and showing that they are more than their his-

torical context is both possible and politically salient. As Ramon Amaro sums up:

> there's a computational logic that cannot be comprehended by humans, and I see a great potential in that, in terms of race, in terms of gender dynamics, in terms of homophobia and so on and so forth. (Amaro/ Hui/ Dasgupta 2021: 55)

Again, an imaginary interlocutor may ask: How is it possible for a computational logic to engage at all with such socio-historically bounded phenomena such as race or gender dynamics? The answer lies again in an analogy with Yuk Hui's cosmotechnics: the topological model of reason delivered by ANNs functions as a generic space – a more generous image (compared with the geographical model of orientation) from which genres of thinking can be born, and where they can mutate as well as undergo massive processes of reconstruction. Replacing the underlying interpretive framework of thought can cause a cascade of spontaneous reverse-engineering operations that reshuffle the foundational assumptions of any system, thus restructuring the system from within. The effects of such restructuring could reverberate into every corner of the system, including those that interface with (and spill into) social, political and economic reality.

6. Conclusion

The picture of AI that this chapter presents is the one of a dual mirror. Evidently, AI can be appropriated by subsumptive logic of instrumental, enumerative rationality. But that is just one part of the story. Its second function is that of a mirror which generates what Reza Negarestani has labeled an "outside view of ourselves":

> whereby AGI or computers tell us what we are in virtue of what we are determinately not – i.e., contra negative theology or the uncritical and merely experiential impressions of ourselves. This objective picture or photographic negative may be far removed from our entrenched and subjectivist experience of ourselves as humans. (Negarestani 2018: 4)

This outside view involves an image of reason that assumes a degree of plasticity over what reason is, and it advertises a constructivist approach to delimiting its gestures, tools and competences. The proposition then is that the

topological model of thinking latent in the computational architecture of ANNs represents an applicable, insightful and normatively interesting model for the construction of reason; a model which would orient itself according to principles that transcend the most at-hand critiques of the use of AI in the current socio-historical context, still trapped in an "entrenched and subjectivist experience of ourselves as humans" (Negarestani 2018: 4). The emergent alternative topological model of reason we have discussed above runs parallel to these critiques, exploding the stability of anthropomorphic hierarchies, subsumptions and stiff categorial matrices (which is in the end also an ambition of the critiques of AI). Instead of giving all the credit to the grounds of thinking-as-we-know-it, the topological model of reason drawn from ANNs is historically extraneous, bringing an element of a productive determination from the outside that can contribute to reconstructing what it means for thinking to think.

Ultimately, this emerging image of reason should be seen as part of a longer history of modeling thinking and rationality on different operations in the world. Kant broke important ground in his orientation essay, but his was also a limited excavation because it developed its concept from geographical orientation, fundamentally binding reason to an anthropomorphic need. Although there is nothing intrinsically wrong with drawing from empirical models, the model's primacy contributed to a hegemonic idea about how reason ought to act, making it comfortably human in the process. The outside perspective we are starting to see – given, for example, by the computational logic of AI we have considered in this chapter – is interesting not because it constitutes an alien way of "thinking" but because it frees us to construct new grounding narratives about what reason ought to do beyond any familiar standards. The model gives us leverage on saying what it can conceivably mean to orient oneself in thinking by replacing an archaic concept of orientation that urged us to privilege familiarity. It is on this road that we can begin to dissociate rationality from its human context so as to take a step toward thinking about loftier standards of what thinking can mean, and to see whether the self-transformation of thinking can also transform its milieu.

Bibliography

Amaro, Ramon, Yuk Hui and Rana Dasgupta. 2021. Designing for Intelligence. In *Atlas of Anomalous AI,* Eds. Ben Vickers and K Allado-McDowell. London: Ignota, 53–73.

Bernecker, Sven. 2012. Kant on spatial orientation. *European Journal of Philosophy*, 20(4):519–533.

Carnap, Rudolf. 1922. *Der Raum: Ein Beitrag zur Wissenschaftslehre*. Kant-Studien Ergänzungshefte, 56. Berlin: Reuther & Reichard.

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

Cavia, AA and Patricia Reed. 2022. Pointless Topology: Figuring Space in Computation and Cognition. In this volume, 353–365.

Crawford, Kate and Vladan Joler. 2018. Anatomy of an AI System: The Amazon Echo as an Anatomical Map of Human Labor, Data and Planetary Resources. AI Now Institute and SHARE Lab. https://anatomyof.ai. Last access: 19 December 2022.

Deleuze, Gilles. 1994. *Difference and Repetition,* Trans. Paul Patton. New York: Columbia University Press.

Heidegger, Martin. 1977. *The Question Concerning Technology*. New York: Harper & Row.

Hui, Yuk 2016. *The Question Concerning Technology in China*. Falmouth: Urbanomic.

Kant, Immanuel. 1992a. Concerning the Ultimate Ground of the Differentiation of Directions in Space [AK 2: 375–382]. In *Theoretical Philosophy 1755–1770*, David Walford (Ed. and Trans.). Cambridge: Cambridge University Press.

Kant, Immanuel. 1992b. *Lectures on Logic*. Michael J. Young (Ed.). Cambridge: Cambridge University Press.

Kant, Immanuel. 1996a. What Does It Mean to Orient Oneself in Thinking? [AK 8:133–146]. In *Religion and Rational Theology,* Eds. and Trans. Allen W. Wood and George di Giovanni. Cambridge: Cambridge University Press.

Kant, Immanuel. 1996b. Groundworks to The Metaphysics of Morals [AK 4: 393–463]. In *Practical Philosophy*. Ed. and Trans. Mary, J. Gregor. Cambridge: Cambridge University Press.

Kant, Immanuel. 1998. *Critique of Pure Reason,* Trans. Paul Guyer and Allen W. Wood. Cambridge: Cambridge University Press.

Kogan, Gene. 2017. How Neural Networks Are Trained. Machine Learning for Artists. https://ml4a.github.io/ml4a/how_neural_networks_are_trained/. Last access: 19 December 2022.

Lury, Celia, Luciana Parisi and Tiziana Terranova. 2012. Introduction: the becoming topological of culture. *Theory, Culture & Society*, 29(4-5):3–35.

Malabou, Catherine. 2012. *Ontology of the Accident. An Essay on Destructive Plasticity*. Cambridge: Polity.

Negarestani, Reza. 2018. *Intelligence and Spirit*. Cambridge, Mass.: MIT Press.

Pascal, David. 2014. "Welt". In *Dictionary of Untranslatables: A Philosophical Lexicon,* Eds. Barbara Cassin, Emily Apter, Jacques Lezra, and Michael Wood. Princeton, N.J.: Princeton University Press.

Pasquinelli, Matteo. 2017a. Arcana Mathematica Imperii: The Evolution of Western Computational Norms. In *Former West. Art and the Contemporary after 1989,* Eds. Maria Hlavajova and Simon Sheikh. Cambridge, Mass.: MIT Press, 281–293.

Pasquinelli, Matteo. 2017b. The automaton of the anthropocene: on carbosilicon machines and cyberfossil capital. *South Atlantic Quarterly*, 116(2): 311–326.

Pasquinelli, Matteo. 2019. Three thousand years of algorithmic rituals: the emergence of AI from the computation of space. *E-flux journal* 101. https://www.e-flux.com/journal/101/273221/three-thousand-years-of-algorithmic-rituals-the-emergence-of-ai-from-the-computation-of-space/. Last access: 19 December 2022.

Plotnitsky, Arkady. 2012. Experimenting with ontologies: sets, spaces, and topoi with Badiou and Grothendieck. *Environment and Planning D: Society and Space*, 30(2):351–368.

Schmeikal-Schuh, Bernd. 1993. Logic from space. *Quality and Quantity*, 27(2): 117–137.

Srnicek, Nick. 2017. *Platform Capitalism*. Cambridge: Polity.

Voevodsky, Vladimir. 2014. The Origins and Motivations of Univalent Foundations. Institute of Advanced Studies. https://www.ias.edu/ideas/2014/voevodsky-origins. Last access: 19 December 2022.

Wynter, Sylvia and Katherine McKittrick. 2015. Unparalleled Catastrophe for Our Species? Or, To Give Humanness a Different Future: Conversations. In *Sylvia Wynter: On Being Human as Praxis,* Ed. Katherine McKittrick. Durham and London: Duke University Press, 9–89.

Zalamea, Fernando. 2012. *Synthetic Philosophy of Contemporary Mathematics*. Falmouth and New York: Urbanomic and Sequence Press.

Pointless Topology

Figuring Space in Computation and Cognition

AA Cavia & Patricia Reed

Abstract: *The topological turn in computing is elaborated both methodologically and philosophically via a treatment of inference proceeding from the notion of homotopy. In such a program, computation is unmoored from strictly discrete operations and opens up to geometric, or spatial domains, effectively suturing computational processes with procedures of situational embedding. Via this inferential scheme, hypothesis-generation as well as non-trivial transits between singular neighborhoods of thought can be systematically accounted for. The authors have adopted a dialogue format that echoes their presentation and structure of collaboration, while demonstrating a mode of encounter between discrete domains of knowledge.*

Introduction

The topological turn in computing over the last decade can be identified with the research program known as *univalent foundations* (Awodey 2014), an ambitious project that compels a re-evaluation of the core tenets of computational theory. Spearheaded by the work of Vladimir Voevodsky, univalence reorients computation around a form of structuralism rooted in geometry; this presents a treatment of types as continuous maps, forging invariances known as *homotopies* between spaces. The questions it raises reach into the depths of foundational mathematics, demanding a careful examination of what Lautman (2011) considers the key dialectic at the heart of mathematical practice – namely, the reciprocity of the discrete and the continuous, which we can couch in terms of algebra and geometry; the realm of symbols on the one hand and of spaces on the other.

Topology advances the view that all space comes with an attendant structure – that is to say, it is a means of gleaning structure in space. As a fork in mathematical intuition, it marks the move from naive to critical treatments of spatiality. In Euclidean geometry, a space precedes its axiomatization – points, lines and planes assume their own embedding in a surrounding domain which is extrinsic to the axioms provided; space is given in the first instance. Euler undermined the concept of the Euclidean plane with the development of graph theory, while it was Riemann's great contribution to show that there are many species of space, each with its own notion of locality, its own "shape", so to speak. In this new non-Euclidean world, structure is not absolute, but rather context sensitive to its embedding space. As Châtelet would note, this development marked "the liberation of geometry, 'freed' at last from the physical universe" (2000: 6), reframing the relation between algebraic structure and a geometric a priori.

Posing a dialectical relation between the two, as Lautman (2011) does, sidesteps the question of precedence, presenting the challenge of synthesis via what Lautman calls *mixtures*. The arc of post-war mathematics represents a major research program to fuse these two *ur*-branches of mathematical practice, which we can refer to at a high level as the project of algebraic geometry, most notably expressed in category theory and its various strands. This relativistic framework posits a plurality of spaces, each embodying its own logic, preceded by their articulation as mathematical structures such as *toposes*. The body of research associated with topos theory (Goldblatt 2014) presents a novel vocabulary for encoding space in terms of *schemes*, *sections*, *neighborhoods* and *sites,* objects whose algebraic and geometric properties are indistinguishable – providing a toolkit for treating space as a derived mathematical notion.

> From an epistemological perspective, there are notable ramifications to draw from this topological turn concerning possible systematic mobility through various knowledge 'sites', more conventionally called disciplines. Deploying mathematical or physics-based concepts at the level of sheer metaphor or terminological/analogical transfer brings the risk of triviality that works to undermine otherwise important efforts for transdisciplinary scheme-building. When faced with the methodological challenges posed by transdisciplinarity, where the discovery or invention of non-trivial movement through sites of knowledge is crucial, the construction of rigorous modalities of transfer becomes essential. We may think of this, provisionally, as topological learning, which is then also bound

to the question of topological pedagogies. Broadly stated, topology provides a conceptual scaffold for thinking about relations between singularity and unity in systematic ways. Methodologically, topological learning affords the preservation of specificity belonging to discrete neighborhoods of knowledge, while also inventing applied transits across distinct knowledge types.

In Olivia Caramello's theory of topos-theoretic 'bridges', the constructability of space is contingent on what she calls "dynamic unification" (Caramello 2016). Dynamic unification is not a generalization, but occurs where entities are related through a constructed third 'bridge object' that enables the transfer of information, not unlike Fernando Zalamea's evocation of a 'glueing procedure' when addressing mathematical sheaves (2012: 285). Put in Simondonian terms, the bridge object, or glueing procedure, is what enables transduction, the process driving the possibility of individuation. It is what allows for non-trivial comparisons and transfers between discrete entities without sacrificing what is specific to those entities. The guarantee of non-trivial transfers between discrete entities is enabled through Morita-equivalence, which mathematically speaking can preserve syntactic difference while mapping a common "semantical core" (Caramello 2010: 14).

For Caramello, this bridge object, otherwise known as an invariant, is a constructed representation of such a common semantic core between distinct entities. Applying such a topological model to the problems of disciplinary epistemologies in the kind of high-dimensional information contexts under consideration here serves as a fruitful heuristic for overcoming what Lewis Gordon calls "disciplinary decadence" (Gordon 2014). Gordon warns against the turning inward of disciplines that results in both the failure to recognize their limits and a fortification of their self-referential enclosures, secured through the sheer learning of codes and rehearsal of methods belonging to a field. Disciplinary sites calibrate not only what we think (content) but how we think (method), and when petrified they enact "an absolute conception of disciplinary life" (Gordon 2014). The petrification of a discipline is spatially isomorphic with the givenness of its neighborho od enclosure for thought – in other words, the notion of an embedding space 'free' from interference. While a lack of interference may be idealized as smooth and efficient, at the level of human cognition it is also the very condition that disables transfers and thereby prohibits transduction, resulting in a false conflation of metastability and fixity.

Sites and locales

This emphasis on the constructibility of space, which we can trace through topology and category theory, stems from an insistence that a modal structure, an inferential lattice of sorts, accompanies any spatial articulation – what we are referring to as a *site*. Discursive sites are bound not only to phenomenological constraints but to dynamic inferential loci, constraining the possible relations that can manifest in any given embedding space. The apotheosis of this view is the theory of *locales*, which asserts the precedence of an algebraic structure from which a space can be constructed. This is a form of *pointless topology*, outlining a propositional lattice, a kind of inferential frame which is point-free and is not derived from any geometric notion. As the constructive mathematician Andrej Bauer (2013) tries to explain:

> *In the usual conception of geometry, a space is a set of points equipped with extra structure, such as metric or topology. But we can switch to a different view in which the extra structure is primary and points are derived ideal objects. For example, a topological space is not viewed as a set of points with a topology anymore, but rather just the topology, given as an abstract lattice with suitable properties, known as a locale. In constructive mathematics such treatment of the notion of space is much preferred to the usual one.*

Doxastic spaces

Such approaches mark out the topological turn as an insistence on the situatedness of mathematical procedures such as proofs, embedding them in sites conditioned by locales. In this view, there can be no cleaving of the doxastic space of reasons from the mathematized space of geometry, they are both sutured with an inferential structure. The influence of the topological turn on contemporary AI is most markedly expressed by *the manifold hypothesis* (Fefferman/Mitter/Narayanan 2016) – the theory that real-world data can be represented as a complex of manifolds, or continuous surfaces, in an embedding space. Embeddings enable contemporary deep learning models to explore the 'latent space' of relations present in any given form of data, uncovering patterns which cannot be easily distinguished in the input space. This marks out deep neural nets as not simply flat networks of associations, but rather high-dimensional spaces induced by transformations in which the algebraic and the

geometric cannot adequately be decoupled. This enmeshing of the continuous and the discrete is another way of stating that an embedding space and its topological structure can only ever exist as a holistic composition. By contrast, earlier models such as *support vector machines* hinged instead on the programmer's ability to construct spaces by explicitly defining their structure up front via *kernel functions;* the models were not able to *learn* embeddings on their own. As such, the field of deep learning can be viewed as a paradigm shift in AI, from the construction of kernels to the induction of embeddings.

A central problem constraining the inferential limits of deep learning is in considering how an embedding space can come to be imbued with a modal structure, which is to say, how possibility can be couched in terms native to a geometric mode of representation. In a sense, how possibility as such can come to be embedded. This question is an expression of a historical theoretical problem in AI, namely *the frame problem*. First articulated by McCarthy & Hayes in the context of robotics in 1969, the problem concerns how a symbolic system can capture the result of actions in an environment, without having to explicitly delineate not just their effects but their non-effects on a variety of other entities. The question relates to the manner in which a situated computational agent negotiates not only its own boundedness but that of all other entities known to it – the means by which it localizes relations and constrains the effects of interaction in its own mental models of the world.

> The effects of interaction, understood as a mode of interference, and its feedback upon mental models of the world are driven by sensitivities to information. By comparison, disciplinary decadence can be described, geometrically, as an absolute fixing of a Euclidean site for certain knowledge types, and through a practice of self-bordering it amounts to conditions for informational desensitization. The degree of transformation of mental models corresponds to the receptivity to re-cognize "signals" or "alerts", as Ramon Amaro and Murad Khan have written in their outline of an expanded picture of interpellation beyond its pejorative guise as that which underwrites the self-transformational opportunity for updating mental models (Amaro/Khan 2020). From a topological perspective, the updating of mental models is akin to recognizing new conditions of situatedness: absent a static or a priori site from which to think, cognitive transformation is equal to the construction of other locales for embedding thought.

> Desensitivity to information is reinforced by an absolute fixing of thought to a given site of embedding. What is lacking is the comparative interference of a bridge object as a creativity-enabling catalyst, effectively serving as the creation of another site from which to interact with a world and its stuff; and it is through such comparative triangulation that abductive cognition or hypothetical reasoning is made possible. In an effort to demystify the type of creativity required for discovery, Lorenzo Magnani emphasizes the role of external representations as a "means to create communicable accounts" of the novel, making them amenable to generative interpellation beyond mere flights of personal imagination, and to the shareable updating of mental models (Magnani 2009: 2). Linking back to Caramello, we may suggest that 'bridge objects' function as just such a necessary external representation, and in some cases as an epistemic mediator for non-trivial transits to novel neighborhoods for embedding speculative cognition. It is via the systematicity of constructing such 'bridge objects' that abductive thought can be justifiably (that is, systematically) performed, via the morphing of spaces for the taking-place of reason as such. As Magnani emphasizes, these transits of abduction do not only operate within theoretical registers of thought, but may also take place via manipulative experiments (that is, environmental interactions with material stuff), introducing an opportunity for transfers between know-of and know-how in hypothesis construction, while broadening the scope of what a 'model' may be. The importance of manipulation in model-based reasoning, particularly vis-à-vis procedures of discovery, is the enablement of a "redistribution of the epistemic and cognitive effort" to contend with entities and information that "cannot be immediately represented or found internally" (Magnani 2004: 233).

The questions raised by interaction, conceptual revision and hypothesis formation pose numerous challenges to a theory of intelligence. The frame problem exposes a key limitation of deductive systems based on classical logic, namely that laws with open-ended sets of exceptions cannot be adequately captured; and the 'common sense' law of inertia, which encapsulates the fact that most actions do not alter most properties of most entities, is just such an open-ended law. For Dennett (1984), the frame problem represents a deep challenge to counterfactual reasoning. Along similar lines, Fodor (2000) considers the frame problem a matter of how we represent modality, which is to say the informational encapsulation of contingency, possibility and necessity. For Brandom

(2010: 79), this is transformed into the question of "doxastic updating", of how an agent updates their beliefs in order to accommodate real-time interaction, and it motivates a position he calls "pragmatic AI", which untethers itself from formal logic in an attempt to address the issue. The Bayesian riposte is to dispense with symbolic reasoning altogether, embracing an inductive scheme in which prior belief is incrementally updated on the basis of feedback – a continuous back-propagation of error fed into a generative model at every instance. This is precisely the strategy taken by deep learning, which evades the question of modality altogether, offering what Pearl (2018) critiques as a "model blind" approach to inference.

In the Bayesian view of intelligence, interference and learning are synonymous, precipitating a fluid updating of prior beliefs. However, rationalist critiques of Humean empiricism are valid here as challenges to the limits of such a scheme, particularly in relation to the epistemological claims made on behalf of deep learning models, which are theorized entirely along inductive lines. In essence, the problem concerns the distinction between prediction and explanation: Brandom (2010) and others reject the supervenience of normativity on empirical generalization, asserting that one cannot possess the affordances of modal reasoning, namely counterfactual robustness in one's inferences, without such capacities. For Brandom, the state of the actual world, its possibility space, and the way it ought to be, are all inextricably enmeshed in everyday acts of reasoning. For Sellars (2007), the account of perception which is central to empiricism is similarly parasitic on a modal structure, which in turn relies on a "space of reasons", whereby accounts of particulars are subsumed by general laws. Lastly, for computer scientist Judea Pearl (2018), model blind approaches to AI are hampered by their rejection of causality, conceived as an objective modal structure of reality. It is the link between the counterfactual and the causal which, in Pearl's view, characterizes the explanatory power of a model *qua* model, as that which singles out a theory as a robust explanation in the first place.

> As Badiou (2007) suggests, models are what allow us to think through participation, locating it at the juncture between the sensible (empirically available) and the intelligible (conceptual). Participation is a form of interaction that induces interference as receptivity to stimuli – or, put another way, sensitivity to information. Interference also serves as the critical operator for Turing in his accounts of organized and unorganized machines, in which the latter is linked to the status of an infant cortex. Machines

that are random in their construction are what Turing calls "unorganised machines", where interferences gradually set the conditions for machinic operations, a discrete state called a configuration (Turing 1969). The parsing of interference after a certain configuration is achieved produces an organized machine determined for some definitive purpose. Analogously, interference inhibition is isomorphic with doxastic preservation – or, more simply put, conceptual and/or practical habits. What is relevant to highlight from Turing's paper on machinic intelligence is his persistent assertion that so-called "proof" of the impossibility for machinic intelligence can be found in its failures or errors. This is a narrow property and/or expectation of intelligence: not only would every human also fail in this regard, but it undervalues how error is a critical form of interference for reorganizing thought and activity.

On two notions of frame

The challenge of modal reasoning appears irrevocably linked to questions of causality, error and doxastic updating. As such, the frame problem can be cast as a close cousin to the *problem of induction*, but it is not synonymous with the bind presented by Humean skepticism. In a sense, both are intrinsically spatially articulated problems raised by interaction. But the distinction rests on the fact that the former relates to our inferential models of the world as opposed to the causal structure of reality, while neither fully resolves the relationship between the two. Instead, one can say that they both compel a holistic account of reasoning in light of interaction; the prospect of unifying empirical generalization with inferential explanation.

Thus there are two notions of frame one can leverage to approach the frame problem, if one accepts it as a substantial theoretical challenge to computational reason. The first is a modal frame, such as those devised by Kripke (1963). These are foundational in model theoretic treatments of possible world semantics, and are the most common tool employed to resolve this bind. This sense of frame is a kind of meta-language which can be used to create correspondences with a target or object language. It is expressed as a pair <W,R> where W is a set and R is an operation (a binary relation on W). Elements of W are called nodes or worlds. Together the frame and a third component, an evaluation or forcing relation (), are called a model.

The second notion of frame is a generalization of a topological space known as a locale, as previously outlined by Bauer. In this sense a frame is a pointless topology representing a category of open subsets of a space. In this view, frames and locales are both lattices, algebraic forms which describe a topology but precede any geometric expression of a space.

Here we can ask, what kind of frame is the frame problem alluding to? It may be neither or perhaps both, as it is not a strictly mathematical use of the term. In a certain sense, it is a reference to a causal meta-representation one can conceive as a mental model possessed by an agent. In another sense, it is a fundamental challenge to any representational theory of mind. It poses the question of how we can reason about things as other than they actually are, and how we can judge those hypothetical situations to form abductive hypotheses – in effect, to ask which inferential moves are permitted and which are incompatible within our conception of the causal structure of the world, which is ultimately the question of navigation. As a result, the formal treatment of frames one chooses to approach the problem has implications for the semantic analysis of knowledge claims made on behalf of any AI that can be said to deal successfully with the issue it presents.

> Pointless topology as a frame for thinking abductive novelty translates into the space of search, such that queries shift from what an object of inquiry is, to where and how it is located/localized (Negarestani 2015). The conditions and contexts of embedding inquiries are open for manipulability, and by transforming the space of search new problems and knowledges can be discovered. It is notably on the activity of constructing problems (more so than identifying 'answers'), that Thomas Kuhn's proposal for paradigmatic scientific change hinges (1962: 37). This model is extrapolated to the social domain by Sylvia Wynter in her account of the necessity to mobilize 'outer views' from which to probe non-adaptation to existing normative spaces for the embedding of thought (Wynter 1984).
>
> There are two ways to consider the navigation of search-spaces outlined by Guerino Mazzola, who distinguishes between receptive and productive navigation – or, in terms we have been otherwise noting, adaptive and non-adaptive (abductive cognition). Receptive navigation is exemplified by an encyclopedia delineating a search-space where knowledge can be augmented, but only within the immutable configuration of its "orientation environment", namely the indexicality of alphabetical ordering (Mazzola 2002: 44). Such a receptive/reproductive mode is optimized for

navigating existing sites of knowledge in fragmented ways, but does not permit "less trivial search problems" (Mazzola 2002: 44). Productive navigation, in contrast, changes the very environment or neighborhood of search as a transformation in the conditions of embedding and encoding queries. The dynamism of search-space endemic to productive navigation undoes a more passive idea of navigation as merely the steering of a ship through existing, preconfigured spaces, turning it into the concept of a site of inseparable interaction between objects and conditions. The ongoing challenge inherent to productive navigation, as Mazzola notes, is how to make such novel embeddings of search intelligible to other agents; in other words, how to create such neighborhoods for thought as bridge objects, amenable to participation.

Locales of truth

If the frame problem is a problem about navigation, we can ask what model of cognition gives us the most convincing account of interactive revision. What affordances are required to reason in dynamic environments beyond the brute forcing of a search space? The most comprehensive riposte to the frame problem is broadly connectionist: by eschewing a representational model, one can adopt a form of semantic holism unfettered by symbolic reasoning, one that is closer to a more abstract topological notion, which is an embedding space. This approach addresses many of the doxastic issues which otherwise motivate an appeal to frames, and as such it allows deep learning to sidestep the frame problem entirely. However, this does not totally eliminate the need for inferential structure, as alluded to by Brandom, Sellars and Pearl; it cannot adequately account for normativity, causality or counterfactual reasoning, by appeal to connectionism alone. Instead it casts the frame problem in terms of navigating the modal structure of an embedding space. This directs us toward the second definition of frame, namely a pointless topology, as a more attractive option for approaching the frame problem, by adopting a paradigm which attempts to unify the plasticity of connectionism with the nomological capacity to reason via explanatory models.

It is through the invention of 'props' in either material (manipulable) or conceptual (theoretical) form, serving as mediators of interference, that a reconfigured search-space is rendered sharable, and thus open to partici-

> pation. These are artefacts inviting the abductive intelligibility of unfamiliar neighborhoods for the embedding of thought. For Magnani, abduction must be "an inference permitting the derivations of new hypotheses and beliefs"; as such it is not a new framework that merely explains an older framework, but provides a "very radical new perspective" (Magnani 2009: 32). It is in this way that the locale of a concept is coterminous with the metastable configuration of the vantage points it affords. The proposition here is that the transits between neighborhoods of thought, each belonging to specific conditions of embedding, are dependent upon the construction of bridge objects to pursue non-trivial modes of navigation across concept spaces. A "radically new" site for embedding thought must also privilege modes of access to it, meaning that the possibility of ramifying new search-spaces as a social activity hinges upon the shareable proof of non-triviality, so as to avoid the pitfall of merely constructing proverbial gated communities of thought.

The thread which can be used to suture locales (as frames) with artefactual intelligence is a topological treatment of inference proceeding from univalence. The implications of adopting locales as a guiding paradigm in approaching the frame problem are both methodological and philosophical. The motivation is to ally with the affordances of constructive mathematics, namely computable notions of structure, in order to expand the inferential toolkit available to computational reason, beyond the narrow confines of inductive learning and model blind approaches. This links locales more broadly to the topological turn, as a means of approaching computation from a vantage point which is intrinsically geometric – that is, anchored to sites. The bridging of univalence with the manifold hypothesis thus presents an opportunity for a unified account of computational reason rooted in topology, but it remains a highly speculative proposal. If we wish to gain traction on how rational agents navigate worlds, we should heed challenges such as the frame problem as serious philosophical questions arising from the concretization of models. Such problems provide us with a window onto the thorny and inhospitable landscape sprawling before any contemporary theory of intelligence, hampered as we are at every turn by the embedding of our own cognitive faculties and scientific practices.

Bibliography

Amaro, Ramon and Murad Khan. 2020. Towards Black Individuation and a Calculus of Variations. *e-flux Journal*. https://www.e-flux.com/journal/109/330246/towards-black-individuation-and-a-calculus-of-variations. Last access: 19 December 2022.

Awodey, Steve. 2014. Structuralism, invariance, and univalence. *Philosophia Mathematica*, 22(1):1-11.

Badiou, Alain. 2007. *The Concept of a Model: An Introduction to the Materialist Epistemology of Mathematics*. Trans. Zachary L. Fraser and Tzuchien Tho. Melbourne: re.press.

Bauer, A. 2013. Intuitionistic Mathematics and Realizability in the Physical World. In *A Computable Universe: Understanding and Exploring Nature as Computation*, 143–157. London: World Scientific Publishing.

Brandom, R.B. 2010. *Between Saying and Doing: Towards an Analytic Pragmatism*. Oxford: OUP.

Caramello, Olivia. 2010. The unification of mathematics via topos theory. https://doi.org/10.48550/arXiv.1006.3930.

Caramello, Olivia. 2016. The theory of topos-theoretic 'bridges' – a conceptual introduction. *Glass Bead Journal*. https://www.glass-bead.org/article/the-theory-of-topos-theoretic-bridges-a-conceptual-introduction/?lang=enview. Last access: 20 April 2022.

Châtelet, Gilles. 2000. *Figuring Space: Philosophy, Mathematics and Physics*. Berlin: Springer.

Dennett, Daniel C. 1984. Cognitive wheels: the frame problem of AI. *Minds, Machines and Evolution*:129–151.

Fefferman, Charles, Sanjoy Mitter and Hariharan Narayanan. 2016. Testing the manifold hypothesis. *Journal of the American Mathematical Society* 29(4):983–1049.

Fodor, Jerry A. 2000. *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology*. Cambridge, Mass.: MIT Press.

Goldblatt, Robert. 2014. *Topoi: The Categorial Analysis of Logic*. Amsterdam: Elsevier.

Gordon, Lewis R. 2014. Disciplinary decadence and the decolonisation of knowledge. *Africa Development* XXXIX:81–92.

Kripke, Saul A. 1963. Semantical analysis of modal logic. *Mathematical Logic Quarterly* 9(5-6):67–96.

Kuhn, Thomas S. 1962. *The Structure of Scientific Revolutions*. Chicago: Chicago University Press.

Lautman, Albert. 2011. *Mathematics, Ideas and the Physical Real*. Vancouver: A&C Black.

Magnani, Lorenzo. 2004. Model-based and manipulative abduction in science. *Foundation of Science* 9:219–247.

Magnani, Lorenzo. 2009. *Abductive Cognition: The Epistemological and Eco-Cognitive Dimensions of Hypothetical Reasoning*. Berlin: Springer Verlag.

Mazzola, Guerino. 2002. *The Topos of Music: Geometric Logic of Concepts, Theory, and Performance*. Basel: Birkhäuser.

McCarthy, John and Patrick J. Hayes. 1969. Some philosophical problems from the standpoint of artificial intelligence. *Machine Intelligence* 4:463–502.

Negarestani, Reza. 2015. Where is the Concept? (Localization, Ramification, Navigation). In *When Site Lost the Plot*, Ed. R. Mackay, 225–251. Falmouth: Urbanomic.

Pearl, Judea. 2018. Theoretical impediments to machine learning with seven sparks from the causal revolution. arXiv preprint arXiv:1801.04016.

Sellars, Wilfrid. 2007. *In the Space of Reasons: Selected Essays of Wilfrid Sellars*. Cambridge, Mass.: Harvard University Press.

Turing, Alan. 1969. Intelligent Machinery. In *Machine Intelligence 5*, Eds. Bernard Meltzer and Donald Michie, 3–23. Edinburgh: Edinburgh University Press. Originally published 1948.

Wynter, Sylvia. 1984. A ceremony must be found: after humanism. *Boundary* 2:19–70.

Zalamea, Fernando. 2012. *Synthetic Philosophy of Contemporary Mathematics*. Trans. Zachary Luke Fraser. Falmouth: Urbanomic.

Autor:innenverzeichnis

Miriam Akkermann ist seit 2019 Juniorprofessorin für Empirische Musikwissenschaft an der TU Dresden. Zu ihren Forschungsschwerpunkten zählen u.a. die Musik des 20. und 21. Jahrhundert mit Fokus auf die Einbindung von Technologien und Medien sowie die Analyse von Kompositionen, die Technologien oder Medien einbinden (Schwerpunkt 20. und 21. Jahrhundert) und die Entwicklung von Methodensets, die qualitative und quantitative Betrachtungen in einem Mixed Methods-Ansatz verbinden.

AA Cavia ist Informatiker sowie Theoretiker und lebt in Berlin. In seiner Arbeit beschäftigt er sich mit maschinellem Lernen, Algorithmen, Verschlüsselung und anderen Softwareartefakten.

Jakob Claus ist wissenschaftlicher Mitarbeiter der Professur für Theorie und Geschichte zeitgenössischer Medien an der Carl von Ossietzky Universität Oldenburg. In Berlin, London und Lüneburg hat er Kultur- und Medienwissenschaft studiert und promoviert derzeit zu kolonialen Konstellationen medialer Wissensproduktion in der Ethnologie um 1900.

Yaoli Du ist Doktorandin in der Kognitiven Anthropologie am Institut für Philosophie der Universität Leipzig. Ihr aktuelles Projekt befasst sich mit Semantic Web-Technologien und intelligenten künstlichen Agenten. Zu ihren Forschungsinteressen gehören ebenso die Auswirkungen der Digitalisierung als Transformation von Technologie und Gesellschaft auf die menschliche Kognition.

Jan Tobias Fuhrmann ist Georg-Christoph-Lichtenberg-Stipendiat des Landes Niedersachsen im MWK-Promotionsprogramm *Gestalten der Zukunft.*

Transformation der Gegenwart durch Szenarien der Digitalisierung an der Carl von Ossietzky Universität Oldenburg. Zu seinen Forschungsschwerpunkten zählen soziologische Theorie, Soziologie der Zeit, Digitalisierung und Algorithmisierung der Gesellschaft.

Catriona Gray promoviert an der University of Bath, UK. Ihre Forschung untersucht den Einsatz und die Regulierung von KI durch öffentliche Einrichtungen, insbesondere im Kontext von Migration und Asyl.

Jonathan Harth hat Soziologie, Philosophie und Psychologie studiert und arbeitet als wissenschaftlicher Mitarbeiter am Lehrstuhl für Soziologie an der Universität Witten/Herdecke. Zu seinen Forschungsschwerpunkten zählen die Soziologie der Digitalisierung (insbesondere Virtual Reality und Sozialität unter Bedingungen maschineller Intelligenz) sowie Religionssoziologie (westlicher Buddhismus).

Christian Heck arbeitet derzeit an der Kunsthochschule für Medien Köln als künstlerisch- wissenschaftlicher Mitarbeiter für Ästhetik und neue Technologien. Er unterrichtet dort im Bereich Experimentelle Informatik, wo er auch bei Prof. Dr. Georg Trogemann promoviert. Er ist aktives Mitglied im Forum InformatikerInnen für Frieden und gesellschaftliche Verantwortung (FifF), Glitch-Künstler, Cypherpunk und Codichter.

Michael Klipphahn-Karge ist Kunstwissenschaftler. Er studierte Bildende Kunst und Kunstgeschichte in Dresden, Berlin und Ústí nad Labem und ist derzeit Kollegiat des Schaufler Lab@TU Dresden, außerdem arbeitet er als Redakteur des begutachteten Online-Journals *w/k – Zwischen Wissenschaft und Kunst*.

Tom Lebrun ist Jurist und promoviert in Digitaler Kultur an der Laval University zum Thema von KI-generierten Texten. Er ist Stipendiat des Fonds de Recherche du Québec Société et Culture (FRQSC).

Lukáš Likavčan ist Global Perspective on Society Postdoctoral Fellow an der NYU Shanghai und dort ebenso assoziiertes Mitglied am Center for AI & Culture. Seine Forschungsschwerpunkte sind Technik- und Umweltphilosophie. Er ist Autor des Buches *Introduction to Comparative Planetology* (2019).

Hannah Link ist wissenschaftliche Mitarbeiterin am Institut für Soziologie der Universität Mainz sowie am SFB 1482 *Humandifferenzierung*. Im Zentrum ihrer ethnografischen Forschung stehen humantheoretische Annahmen über ›den Menschen‹ und die informationelle und maschinelle Implementierung dieser Annahmen in die Gestalt von Robotern.

Maximilian Locher hat Wirtschaftswissenschaften, Ethik und Organisationssoziologie studiert. Er arbeitet als wissenschaftlicher Mitarbeiter im Forschungsprojekt *KILPaD* an der Universität Witten/Herdecke, wobei sich seine Forschung auf die Digitalisierung von Produktionsunternehmen konzentrierte. Aktuell ist er Sekretär im Team Transfromation der IG Metall Baden-Württemberg. Sein Forschungsschwerpunkt liegt in der Digitalisierung und in diesem Zusammenhang Arbeits-, Technik-, Medien- und Organisationssoziologie.

Fabian Offert ist Assistant Professor für Geschichte und Theorie der Digital Humanities an der University of California, Santa Barbara. Seine Forschungs- und Lehrtätigkeit konzentriert sich auf die visuellen digitalen Geisteswissenschaften, mit einem besonderen Interesse an der Epistemologie und Ästhetik des maschinellen Sehens und Lernens.

Carl Christian Olsson ist Geograph, Schriftsteller und Philosoph. Er verfolgt an der Newcastle University ein Promotionsprojekt zu den Konsequenten naturalistischer Selbstverständnisse in der geografischen Ideengeschichte. Olsson war Teil der zweiten Kohorte von *The Terraforming* am Strelka Institute in Moskau und forscht derzeit am New Centre for Research & Practice.

Patricia Reed ist Autorin, Künstlerin und Designerin. Sie lebt in Berlin und setzt sich in ihrer Arbeit mit den Manifestationen planetarischen Denkens in materiellen Lebenswelten auseinander.

Jonathan Roberge ist Professor am Institut National de la Recherche Scientifique (INRS) in Montreal, Canada. Als Inhaber des Canada Research Chair in Digital Culture (seit 2012) gründete er das Laboratory on New Digital Environments and Cultural Intermediation (NENIC Lab). Er ist Mitherausgeber der Sammelbände *Algorithmic Culture* (Routledge, 2016) und *The Cultural Life of Machine Learning* (Palgrave, 2020).

Jan Georg Schneider ist seit 2010 Universitätsprofessor für Deutsche Sprachwissenschaft an der Universität Koblenz-Landau (Campus Landau) und seit September 2021 Vorsitzender der Deutschen Gesellschaft für Semiotik. Seine Forschungsschwerpunkte liegen in den Bereichen Allgemeine Sprach- und Zeichentheorie, Sprachnormenforschung und Medienlinguistik.

Yannick Schütte ist Kultur- und Medienwissenschaftler. Er studierte an der Leuphana Universität Lüneburg, der Universität Potsdam und der Université Bordeaux Montaigne. Nach zweijähriger Tätigkeit am Exzellenzcluster *Matters of Activity* der Humboldt-Universität zu Berlin arbeitet er derzeit für den Hatje Cantz Verlag in Berlin.

Nadine Schumann wurde in Leipzig zum Thema *Zur Methodologie der Zweiten-Person-Perspektive* im Bereich Philosophie der Psychologie promoviert. Sie arbeitete als Assistentin am Max-Planck- Institut für evolutionäre Anthropologie und ist Mitglied am Leipziger Forschungszentrum für Frühkindliche Entwicklung. Derzeit arbeitet sie wissenschaftliche Mitarbeiterin am Institut für Informatik und ist Gastwissenschaftlerin am Institut für angewandte Informatik (Infai) in Leipzig.

Katharina Zweig ist Informatikprofessorin an der TU Kaiserslautern, wo sie das Algorithm Accountability Lab leitet. Sie koordiniert dort auch den deutschlandweit einzigartigen Studiengang Sozioinformatik. Sie war 2018–2020 Mitglied der Enquetekommission Künstliche Intelligenz, erhielt 2019 den Communicator-Preis der DFG und ist Bestsellerautorin des Buches *Ein Algorithmus hat kein Taktgefühl* (2019).

Contributors

Miriam Akkermann is the Junior Professor of Empirical Musicology at TU Dresden, Germany. Her research interests include the music of the 20th and 21st centuries with an emphasis on technology and media as well as methods that combine qualitative and quantitative considerations in a mixed methods approach.

AA Cavia is a computer scientist and theorist based in Berlin, Germany. His practice engages with machine learning, algorithms, encodings, and other software artifacts.

Jakob Claus is a research assistant to the Professorship of Theory and History of Contemporary Media at the Carl von Ossietzky University Oldenburg, Germany. He studied cultural studies and media studies in Berlin and Lüneburg, Germany, and London, UK. His dissertation focuses on colonial constellations of media and knowledge production in ethnology around 1900.

Yaoli Du is a PhD student in cognitive anthropology at the Department of Philosophy at the University of Leipzig, Germany. Her current project is about semantic web technologies and intelligent artificial agents. Her research interests also include the impact of digitalization on human cognition.

Jan Tobias Fuhrmann is a Georg-Christoph-Lichtenberg-Fellow of the State of Lower Saxony, Germany, and a PhD student in the doctoral program *Shaping the future. Transformation of the present through scenarios of digitalization* at the Carl von Ossietzky University Oldenburg, Germany. His research focuses on sociological theory, the sociology of time, digitization, and the algorithmization of society.

Catriona Gray is a PhD candidate at the University of Bath, UK. Her research examines the adoption and governance of AI by public authorities, with a particular focus on migration and asylum settings.

Jonathan Harth studied sociology, philosophy, and psychology at Witten/Herdecke University, Germany, where he also received his PhD and works as a research assistant to the Chair of Sociology. His research interests are in the sociology of digitization, including virtual reality and sociability under conditions of machine intelligence. His research has also covered the sociology of religion and Western Buddhism.

Christian Heck is a researcher in the field of aesthetics and new technologies at the Academy of Media Arts Cologne, Germany. In addition to his doctoral research, supervised by Prof. Dr. Georg Trogemann, he also teaches experimental informatics and is an active member of the Forum InformatikerInnen für Frieden und gesellschaftliche Verantwortung (FifF), and also a Glitch artist, cypherpunk and code poet.

Michael Klipphahn-Karge is an art scholar. He studied fine arts and art history in Dresden and Berlin, Germany, and at Ústí nad Labem in the Czech Republic. He is currently a fellow of the Schaufler Lab@TU Dresden and works as an editor of the peer-reviewed online journal *w/k – Zwischen Wissenschaft und Kunst*.

Tom Lebrun is a lawyer and a PhD candidate in digital culture at Laval University Québec, Canada, where he also teaches AI and law. His research on AI text generation is currently funded by the Fonds de Recherche du Québec Société et Culture (FRQSC).

Lukáš Likavčan is a Global Perspective on Society Postdoctoral Fellow and a research affiliate of the Center for AI & Culture at NYU Shanghai, China. His research areas cover the philosophy of technology and environmental philosophy. He is the author of *Introduction to Comparative Planetology* (2019).

Hannah Link is a research associate at the Institute of Sociology as well as the Collaborative Research Center (CRC) 1482 *Studies in Human Categorisation* at the University of Mainz, Germany. Her work focuses on theories of the human and she conducts ethnographic research on the technological implementation of assumptions concerning 'the human' in the field of robotics.

Maximilian Locher studied economics, ethics, and organizational sociology. He was a researcher in the federally funded *KILPaD*-project at Witten/Herdecke University, Germany, where he focused on the digitization of manufacturing organizations. He is currently the union secretary for transformation at the IG Metall Baden-Württemberg, Germany. His research focuses on the digitization of work, media, organization, and management.

Fabian Offert is Assistant Professor for the History and Theory of the Digital Humanities at the University of California, Santa Barbara, US. His research and teaching focus on the visual digital humanities, with a special interest in the epistemology and aesthetics of computer vision and machine learning.

Carl Christian Olsson is a geographer, writer, and philosopher finishing his PhD at Newcastle University, UK, where his research examines the consequences of naturalistic self-understandings in geographical thought. He contributed to the second cycle of *The Terraforming* at Strelka Institute in Moscow, Russia, and is a researcher at The New Centre for Research & Practice.

Patricia Reed is a writer, artist, and designer based in Berlin, Germany. Her work concerns the ramifications of planetary thought upon the material domain of inhabitation.

Jonathan Roberge is Full Professor at the Institut National de la Recherche Scientifique (INRS) in Montreal, Canada. He founded the Laboratory on New Digital Environments and Cultural Intermediation (NENIC Lab) as part of the Canada Research Chair in Digital Culture he has held since 2012. His most recent edited volumes include *The Cultural Life of Machine Learning* (Palgrave, 2020) and *Algorithmic Culture* (Routledge, 2016).

Jan Georg Schneider holds a professorship for German linguistics at the University of Koblenz-Landau (Campus Landau), Germany and is the first representative to the Executive Board of the German Association for Semiotics. His research interests are in linguistics and semiotics, linguistic norms, and media linguistics.

Nadine Schumann received her PhD from the Department of Philosophy at the University of Leipzig, Germany, for a dissertation titled *On the Methodology of the Two-Person Perspective*. She worked as a research assistant at the Max Planck In-

stitute for Evolutionary Anthropology and is a member of the Leipzig Research Center for Early Childhood Development. She is currently a research assistant at the Institute of Computer Science and a visiting researcher at the Institute for Applied Computer Science (Infai) at the University of Leipzig.

Yannick Schütte is a researcher in cultural studies and media theory. He graduated from Leuphana University Lüneburg and University of Potsdam, Germany, and Université Bordeaux Montaigne, France. After two years at the Cluster of Excellence *Matters of Activity* at Humboldt University Berlin, Germany, he currently works for the Berlin-based publishing house Hatje Cantz Verlag.

Katharina Zweig holds a professorship in computer science and is head of the Algorithm Accountability Lab at TU Kaiserslautern, Germany where she also coordinates the only degree program in socioinformatics in Germany. From 2018 until 2020, she was a member of the German Parliament's Study Commission on Artificial Intelligence. Katharina Zweig was awarded the German Research Foundation's Communicator Award in 2019 and is the author of the bestselling book *Ein Algorithmus hat kein Taktgefühl* (2019).

Medienwissenschaft

Marco Abel, Jaimey Fisher (Hg.)

Die Berliner Schule im globalen Kontext

Ein transnationales Arthouse-Kino

2022, 414 S., kart., 48 SW-Abbildungen
30,00 € (DE), 978-3-8376-5248-2
E-Book:
PDF: 29,99 € (DE), ISBN 978-3-8394-5248-6

Martin Donner, Heidrun Allert

Auf dem Weg zur Cyberpolis

Neue Formen von Gemeinschaft, Selbst und Bildung

2022, 496 S., kart., 10 SW-Abbildungen, 5 Farbabbildungen
39,00 € (DE), 978-3-8376-5878-1
E-Book: kostenlos erhältlich als Open-Access-Publikation
PDF: ISBN 978-3-8394-5878-5
ISBN 978-3-7328-5878-1

Geert Lovink

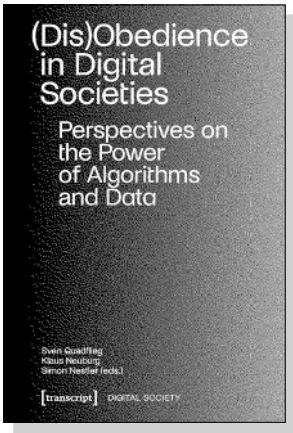
In der Plattformfalle

Plädoyer zur Rückeroberung des Internets

2022, 232 S., kart.
28,00 € (DE), 978-3-8376-6333-4
E-Book:
PDF: 24,99 € (DE), ISBN 978-3-8394-6333-8
EPUB: 24,99 € (DE), ISBN 978-3-7328-6333-4

Leseproben, weitere Informationen und Bestellmöglichkeiten finden Sie unter www.transcript-verlag.de

Medienwissenschaft



Sven Quadflieg, Klaus Neuburg, Simon Nestler (eds.)

(Dis)Obedience in Digital Societies

Perspectives on the Power of Algorithms and Data

2022, 380 p., pb., ill.
29,00 € (DE), 978-3-8376-5763-0
E-Book: available as free open access publication
PDF: ISBN 978-3-8394-5763-4
ISBN 978-3-7328-5763-0

Gesellschaft für Medienwissenschaft (Hg.)

Zeitschrift für Medienwissenschaft 27

Jg. 14, Heft 2/2022: Reparaturwissen DDR

2022, 180 S., kart.
24,99 € (DE), 978-3-8376-5890-3
E-Book: kostenlos erhältlich als Open-Access-Publikation
PDF: ISBN 978-3-8394-5890-7
ISBN 978-3-7328-5890-3

Olga Moskatova, Anna Polze, Ramón Reichert (eds.)

Digital Culture & Society (DCS)

Vol. 7, Issue 2/2021 –
Networked Images in Surveillance Capitalism

2022, 336 p., pb., col. ill.
29,99 € (DE), 978-3-8376-5388-5
E-Book:
PDF: 27,99 € (DE), ISBN 978-3-8394-5388-9

Leseproben, weitere Informationen und Bestellmöglichkeiten finden Sie unter www.transcript-verlag.de