KI zwischen Blackbox und Transparenz

Das Koppeln und Entkoppeln von Kontrollprojekten

Marco Schmitt, Christoph Heckwolf

1. Einleitung

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

Transparenz« befreit. Wenn auch für Konstrukteur:in oder Nutzer:in nicht im Einzelnen nachvollziehbar ist, wie die KI sich selbst strukturiert, dann befreit sie das auch von der Möglichkeit von Nachfragen oder einer auf den Prozess gerichteten Kritik. Hier findet sich dann häufiger eine Gegenkontrollstrategie, die darauf gerichtet ist hier Transparenz einzufordern und unter dem Etikett einer »Explainable AI«, einer erklärbaren künstlichen Intelligenz firmiert (Angelov et al. 2021). Das dies erforderlich wird, hängt mit gesellschaftlichen Transparenzansprüchen zusammen. Sollen Entscheidungen von erheblicher Relevanz an KI-Verfahren ausgelagert werden, muss sichergestellt werden, dass für Betroffene nachvollziehbar bleibt, warum so entschieden wurde. Andernfalls verlieren die Verfahren an Akzeptanz und Legitimität. Dies kann sich jedoch in verschiedenen kulturellen Umfeldern unterschiedlich stark zeigen, da es hier zu jeweils eigenständigen Umgangsformen, bei White (1992, 2008) »Stilen«, mit der Transparenzproblematik kommen kann, die sich dann auch unterschiedlich auf die Adaption und Entwicklung von KI-Verfahren auswirken (Züger/Asghari 2022). An dieser Stelle werden zum Beispiel immer wieder die Unterschiede zwischen China, den USA und Europa betont (Castro et al. 2019; Probst et al. 2018). Der Beitrag nutzt die Begriffe der Kontrolle und des Stils aus der Theorie von Harrison White, um die Transparenzproblematik im Bereich der neueren KI deutlich zu machen und den in der Forschung, wie in der Praxis entwickelten Kontrollstrategien nachzugehen (White 1992, 2008; Schmitt/Fuhse 2015). Dabei werden 19 Experten Interviews mit Forscher:innen und Entwickler:innen herangezogen, welche in den Jahren 2020 und 2021 von Mitarbeiter:innen des Lehrstuhls für Technik- und Organisationssoziologe der RWTH Aachen geführt wurden[1]. Des Weiteren werden Zeitungsartikel aus den USA und dem deutschsprachigen Raum aus den Jahren 2018 bis 2021 für weitere Analysen zur sich ausbildenden Diskursarena benutzt.

2. Das Transparenzproblem der neueren KI

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

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

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

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

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

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

3. Kontrolle, Kontrollversuche und Kontrollprojekte

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

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

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

4. Blackboxing als Kontrollversuch und als Kontrollprojekt

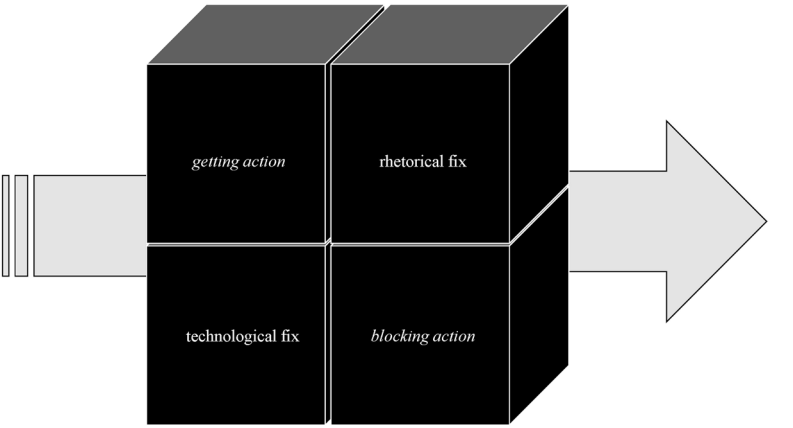
Ausgehend von dieser Beschreibung von Kontrollprojekten und Kontrolle ergibt sich eine doppelte Lesart des Verweises von KI-Forscher:innen auf den Blackbox-Charakter der Deep Learning-Verfahren: Zum einen ist der Verweis auf das Blackboxing als Kontrollprojekt lesbar, dass bestimmten Identitäten Freiheitsgrade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

Kontrollprojekte zwischen Technologie und Rhetorik sowie zwischen Blocking und Getting Action

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

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

Abbildung 1: Schematische Darstellung der 4 Foki von Kontrollversuche



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

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

Empirische Darstellung der Transparenzproblematik

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

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

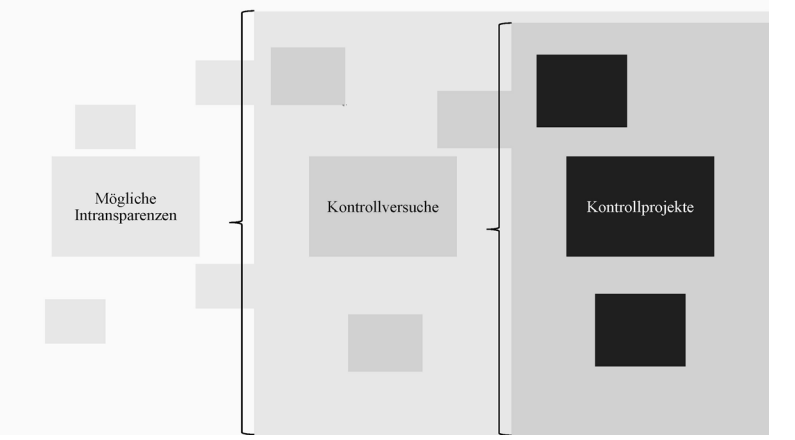
> »Es gibt auch noch so ein paar Netzwerkarchitektur spezifische Sachen, also dass man beispielsweise will, dass Gewichte, die innerhalb von diesen Netzwerken benutzt werden, eine bestimmte Eigenschaft haben, zum Beispiel, dass sie nicht beliebig groß werden oder auf einen bestimmten

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

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

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

Abbildung 2: Von Intransparenzen/Unsicherheiten zu Kontrollprojekten



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

Intransparenz der eingesetzten Lerndaten bzw. ihrer Fehlerquellen

Die hier bestehende Intransparenz ist auch im öffentlichen Diskurs um die KI-Systeme häufiger angesprochen worden[3]. Es ist hier nicht klar auf welcher Datengrundlage Ergebnisse eigentlich beruhen und ob diese Datengrundlage auch gut ist, also frei von Vorurteilen oder Rauschen. Aus der Kontroll-Perspektive weisen die Interviews hier auf einige Versuche und Projekte hin, die gängigerweise in Anschlag gebracht werden. Diese werden jedoch häufig nach außen nicht sichtbar, sodass eine Kontrolle dieser Daten von externer Seite häufig nicht gegeben ist.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Intransparenz des Trainingssettings und der vorgenommenen Änderungen am System

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Intransparenz auch der mathematischen Grundlagen

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Wenn man Blackboxing als Kontrollprojekt liest, stellt sich in den diskutierten Beispielen immer auch parallel dazu die Frage nach möglichen Kontrollprojekten, die hier in die Gegenrichtung arbeiten, also die Öffnung der Blackbox fordern und nicht nur auf Ergebnisse schauen wollen, sondern auch auf ein Verstehen des Zustandekommens pochen, da nur dann eine erfolgreiche Kontrolle gewährleistet ist. Da heute die Vertrauenswürdigkeit intransparenter Prozesse insgesamt als kritikwürdig angesehen wird, erheben sich Forderungen nach Möglichkeiten, die Intransparenz der KI-Systeme stärker in den Blick zu nehmen und Möglichkeiten zu erkunden, wie dies geschehen kann[5]. Interpretierbarkeit und Erklärbarkeit werden hier zu kommunikativen Vehikeln, die unterschiedliche Formen annehmen. Auch hier geht es darum einige Beispiele einzufügen, wie Kontrolle über Erklärung erreicht werden soll, was dabei mit der Intransparenzproblematik passiert und wer hier wie Kontrolle ausübt. Erklärung kann hierbei als »rhetorical fix« angesehen werden, der Kontrolle auf der Erzählungsebene ausübt. Gleichzeitig wird in der Diskussion jedoch häufig auch ein »technological fix« angestrebt, bei der die Verfahren quasi auto-

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

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

Deshalb sollen die Interviewdaten noch kurz einem Topic Modeling der öffentlichen Diskussion über KI in kulturell verschiedenen Weltregionen (Deutschland, USA, Großbrittanien und China) gegenübergestellt werden. Dazu wurden Artikel aus überregionalen Tages- und Wochenzeitungen aus den Jahren 2018 – 2021, welche jeweils mind. einmal »künstliche Intelligenz« bzw. »Artificial Intelligence« beinhalten, ausgewählt. Es wird dann eine Word Cloud aus den Topicmodeling-Ergebnissen generiert, da zu jedem Topic eine Wortliste erzeugt wird, deren Gewicht dann in Abhängigkeit zur Worthäufigkeit in den Texten abgebildet wird. Dabei ist zu beachten, dass die Größe der Wörter nicht alleine auf der Häufigkeit in den Topics basiert, sondern auch auf der »Stärke« des oder der Topics, die die Wörter beinhalten[6]. Nun ist eine Interpretation dieser Wortwolken sicher mit Vorsicht zu genießen, da hier vielfältige Auswahleffekte greifen. Doch in Anlehnung an Whites Konzept der Stile, das mit selbstähnlichen Verteilungen von Elementen operiert, die eingesetzt und erkannt werden, lassen sich doch grobe Ableitungen im Hinblick auf die wahrgenommenen Problemfelder und deren Relevanz ableiten.

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

Anwendungskontexte, wie Medizin, Finanztechnologie, Mobilität, Bilderkennung und Krieg erscheinen in den Topic Models, zu dem ist erkennbar, dass kritische Berichterstattung in deutschsprachigen Medien sehr viel prominenter zu sein scheint. Erklärung und Erklärbarkeit spielen hier aber noch keine große Rolle und auch die Vertrauensproblematik, die sich aus dem Nicht-Verstehen ergibt, ist in der öffentlichen Diskussion unterrepräsentiert.

Abbildung 3: Topic Model Visualisierung, China, 2018–2021

Abbildung 4: Topic Model Visualisierung, Deutschsprachiger Raum, 2018–2021

Abbildung 5: Topic Model Visualisierung Großbritannien & Irland, 2018–2021

Abbildung 6: Topic Model Visualisierung, Vereinigte Staaten, 2018–2021

Wie man an der Gegenüberstellung in einfachen Wordclouds leicht erkennen kann, verschieben sich die dominanten Foki der Diskussion um KI mit dem kulturellen Hintergrund der Zeitschriften. Während es in China um Marktchancen und Industrieentwicklung geht, konzentriert sich die Diskussion in Deutschland stark auf zukünftige Folgen, vor allem für den Arbeitsmarkt, in Großbritannien und Irland auf Forschung und Wandel und schließlich in den USA auf Wissen und die großen Techkonzerne. Was man

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

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

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

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

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

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

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

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

7. Ausblick

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

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

8. Literatur

Angelov, Plamen P./Soares, Eduardo A./Jiang, Richard/Arnold, Nicholas I./Atkinson, Peter M. (2021): »Explainable artificial intelligence: an analytical review«, in: WIREs Data Mining and Knowledge Discovery 11.

Callon, Michel/Latour, Bruno (1981): »Unscrewing the big Leviathan: how actors macro-structure reality and how sociologists help them to do so«, in: Karin Knorr-Cetina/Aaron Victor Cicourel (Hg.), Advances in social theory and methodology: Toward an integration of micro-and macro-sociologies, London: Routledge and Kegan Paul, S. 277–303.

Castro, Daniel/McLaughlin, Michael/Chivot, Eline (2019): »Who Is Winning the AI Race: China, the EU or the United States?«, in: Center for Data Innovation von August 2019, https://s3.amazonaws.com/www2.datainnovation.org/2019-china-eu-us-ai.pdf.

Dastin, Jeffrey (2018): »Amazon scraps secret AI recruiting tool that showed bias against women«, in: Reuters vom 11.10.2018, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

Diesner, Jana (2014): »ConText: Software for the Integrated Analysis of Text Data and Network Data«, in: Social and Semantic Networks in Communication Research.

Diesner, Jana et al. (2020): ConText: Network Construction from Texts, https://context.ischool.illinois.edu/.

Geitz, Eckhard/Vater, Christian/Zimmer-Merkle, Silke (2020): Black Boxes – Versiegelungskontexte und Öffnungsversuche: Interdisziplinäre Perspektiven, Berlin: De Gruyter.

Langer, Paul F./Weyerer, Jan C. (2020): »Diskriminierungen und Verzerrungen durch Künstliche Intelligenz. Entstehung und Wirkung im gesellschaftlichen Kontext«, in: Micheal Oswald/Isabelle Borucki (Hg.), Demokratietheorie im Zeitalter der Frühdigitalisierung, Wiesbaden: Springer, S. 219–240.

Latour, Bruno (1994): »On Technical Mediation. Philosophy, Sociology, Genealogy«, in: CommonKnowledge 3(2).

Latour, Bruno (1999): Pandora's hope: essays on the reality of science studies, Cambridge/Massachusetts: Harvard University Press.

Latour, Bruno/Woolgar, Steve (1986): Laboratory life: the construction of scientific facts, Princeton/N.J: Princeton University Press.

Lehmann, Katharina (2021): »Wenn der Code schwarz sieht«, in: Die Welt vom 19.11.2021, https://www.welt.de/wirtschaft/better-future/article235149630/Algorithmen-sind-nicht-objektiv.html.

Lloyd, Kirsten (2018): »Bias Amplification in Artificial Intelligence Systems«, in: Frank Stein/Alun Preece/Mihai Boicu, AAAI FSS-18: Artificial Intelligence in Government and Public Sector, Arlington: Cornell University.

Luhmann, Niklas (1991): Soziale Systeme: Grundriß einer allgemeinen Theorie. 4. Auflage, Frankfurt a.M.: Suhrkamp.

Mittelstadt, Brent/Russell, Chris/Wachter, Sandra (2019): »Explaining Explanations in AI«, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, New York: Association for Computing Machinery.

Perrigo, Billy (2023): »Exclusive: The $2 Per Hour Workers Who Made ChatGPT Safer«, in: Time vom 18.01.2023, https://time.com/6247678/openai-chatgpt-kenya-workers/

Probst, Laurent/Pedersen, Bertrand/Lefebvre, Virginie/Dakkak-Arnoux, Lauriane (2018): »USA-China-EU plans for AI: where do we stand?«, Brüssel: European Commission.

Rudin, Cynthia (2019): »Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead«, in: Nature Machine Intelligence 1, S. 206–215.

Schmitt, Marco/Fuhse, Jan (2015): Zur Aktualität von Harrison White, Wiesbaden: Springer VS.

Somani, Ayush/Horsch, Alexander/Prasad, Dilip K. (2023): Interpretability in Deep Learning, Cham: Springer International Publishing.

Van Aert, Robbie C. M./Wicherts, Jelte M./Van Assen, Marcel A. L. M. (2019): » Publication bias examined in meta-analyses from psychology and medicine: A meta-meta-analysis«, in: PLoS ONE 14 (4).

von Randow, Gero (2018): »A wie Algorithmus. Wer über die Digitalisierung spricht, kommt an drei Grundbegriffen der Informatik nicht vorbei. Was bedeuten sie politisch? «, in: Die Zeit vom 22.02.2018, https://www.zeit.de/2018/06/informatik-roboter-algorithmus-kuenstliche-intelligenz.

White, Harisson C. (1992): Identity and Control. A Structural Theory of Social Action, Princeton/New Jersey: Princeton University Press.

White, Harisson C. (2008): Identity and Control. How social formations emerge. Second edition, New Jersey: Princeton University Press.

White, Harrison C./Godart, Frédéric (2007): »Stories from Identity and Control«, in: Sociologica 3.

Züger, Theresa/Asghari, Hadi (2022): »AI for the public. How public interest theory shifts the discourse on AI«, in: AI & SOCIETY 38, S. 815–828.