Warum und wozu erklärbare KI?

Über die Verschiedenheit dreier paradigmatischer Zwecksetzungen

Suzana Alpsancar

Abstract: *Currently, explainable AI (XAI) is often touted as a remedy for the so-called black box problem of machine learning without any distinction. However, both the problem of seeing the issue in the blackness, i.e. opacity, of certain AI systems, and the dominant strategy to solve this problem by decreasing the opacity with XAI are misleading in their general applicability. This article calls for a nuanced and reflected investigation into the usefulness of XAI. To do this, it is not only important to name purposes in the first place, but also to make them concrete and take their paradigmatic differences seriously. To demonstrate this dissimilarity, the article adopts a perspective inspired by Th. Kuhn and highlights three paradigms of current XAI research, all of which, despite their fundamental differences, are currently being dealt with as if they represent a sort of ›normal scientific puzzle solving‹ in machine learning.*

Keywords: *black box; AI Ethics; explainable AI (XAI); human centered AI; value alignment*

1. Einleitung: Schwarze Kisten als Problemstellung

›Explainable AI‹ (XAI) ist ein in den letzten Jahren schnell wachsendes Forschungsfeld, welches sich darum bemüht, schwarze KI-Kisten zu lichten und eine Palette von Methoden vorgelegt hat, die verschiedene formale Aspekte der Funktionsweise von KI-Systemen beschreiben, die sonst im Verborgenen blieben (Hu 2020; Vilone/Longo 2021).[1] Das Forschungsfeld und die Debatte um erklärbare KI versteht sich selbst primär als Reaktion auf einen Bedarf, der aus der Forschung und Entwicklung und der gesellschaftlichen Anwendung von KI-Systemen hervorgegangen ist

1 Die Arbeit an diesem Beitrag wurde gefördert von der Deutschen Forschungsgemeinschaft (DFG): TRR 318/1 2021 – 438445824. Viele Gedanken sind im Austausch mit meinen wundervollen Kolleg:innen im TRR 318 entstanden, denen ich an dieser Stelle herzlich danke. – Für die Redaktion des Textes bedanke ich mich bei Amber Sophie Kieffer und Sebastian Mantsch für ihre Unterstützung.

(Capel/Brereton 2023). Dieser Bedarf wird als das Problem verstanden, für welches XAI die, bzw. eine wichtige Lösung ist. Damit fußen Forschung und Debatte auf einer bestimmten Problemkonzeption. Die gängige Fassung dieses Problems ist die Konzeption als ›Black-Box-Problem von Machine Learning‹. Diese Konzeption findet sich zum Beispiel bei Kamath und Liu, die in ihrem Buch *Explainable AI. An Introduction* (Kamath/Liu 2021) die Notwendigkeit des Forschungsfelds XAI mit dem Nachteil von Machine Learning (ML), opak zu sein, begründen:

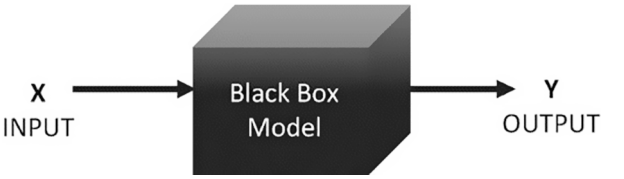
> »The field of explainable AI addresses one of the most significant shortcomings of machine learning and deep learning algorithms today: the interpretability of models.« (Kamath/Liu 2021: ix)

In dieser Problemkonzeption werden der jüngste Erfolg von KI sowie die Opazität ML-Verfahren allein zugewiesen. Die Opazität wird als Nachteil bewertet und als Preis deklariert, den man für den Gewinn von höherer Akkuratheit bezahlen müsse – ML-Systeme bringen »greater accuracy at the expense of complexity and explainability« (Kamath/Liu 2021: 2). Opazität wird so als Kehrseite des Erfolgs angenommen und beides als genuin und rein technische Angelegenheit gerahmt. XAI wird als eine zentrale bzw. die Lösungsstrategie eingeführt mit dem Problem der Undurchsichtigkeit umzugehen, abermals als rein technische Angelegenheit. Die verbreitete Metaphorik von ML-Systemen als Black-Box verstärkt den Eindruck, dass das Grundproblem, für das XAI die Lösung sein soll, allein auf Eigenschaften des technischen Systems zurückzuführen sei:

> »Many machine learning and deep learning models are essentially ›black-boxes‹ that do not reveal the internal mechanisms and nuances to their predictions.« (Kamath/Liu 2021: 2)

Dieser Blackbox-Charakter wird typischerweise als schwarze, geschlossene Kiste visualisiert (siehe Abbildung 1). Diese Problemkonzeption ist in zwei Hinsichten unzulänglich. Zunächst ist die Problemstellung ungenau und pauschal. Die Ursachen der Undurchsichtigkeit liegen nicht ausschließlich in technischen Eigenschaften (Burrell 2016).

Abbildung 1: Darstellung des Black-Box Problems



Quelle: Kamath/Liu 2021: 2

Burrell hat den Blick auf die Kontexte und die soziale, ökonomische und politische Einbettung der Technologien gelenkt und vorgeschlagen, drei verschiedene Typen von Opazität zu unterscheiden (Burrell 2016): Zunächst können Systeme für Dritte oder Außenstehende opak erscheinen, weil sie unter das Geschäftsgeheimnis fallen und damit wichtige Informationen über ihre Funktionsweise, ihren Aufbau, die verwendeten Datensätze u.ä. nicht einsichtig sind. Sodann ist eine Opazität relativ zum Kenntnisstand verschiedener Akteur:innen zu unterscheiden. Systeme können aufgrund einer fehlenden »digital literacy« für bestimmte Gruppen opaker als für andere sein. Von diesen beiden Hinsichten, die die soziotechnische Anwendung der Systeme in den Blick nimmt, unterscheidet Burrell einen dritten Typ, den sie epistemische Opazität nennt und der auf die Eigenschaften des Systems abhebt: sind die Eigenschaften des Systems generell nachzuvollziehen oder treten hier Grenzen und besondere Herausforderungen auf? Liptons Überlegungen (Lipton 2018) setzen bei diesem dritten Typ von Opazität an. In seinem kritischen Kommentar zum Mythos der Model-Interpretierbarkeit stellt er zunächst fest, dass es in der XAI-Debatte zwei grundverschiedene Bedeutungen von Erklärbarkeit/Opazität gibt: nämlich zum einen die *ex ante transparency* und zum anderen die *post hoc interpretability*. Der Unterschied ist ein zeitlicher: geht es um eine Einschätzung des Systems vor seinem Gebrauch, ex ante, und damit generell oder geht es darum rückblickend, von einem gegebenen Ergebnis des Systems ausgehend besser nachvollziehen zu können, wie es zu diesem Ergebnis kam? Lipton zufolge sind die Bemühungen im XAI-Bereich auf den zweiten Typ von Erklärbarkeit gerichtet. Die vielen verschiedenen Tools stellen Mittel dar, Systemergebnisse post hoc verstehbarer zu machen. Darüber hinaus führt er eine wichtige Differenzierung zur Rede von *ex ante transparency* ein, die verdeutlicht, dass es zu einfach ist von einer pauschalen Opazität von bestimmten Systemtypen, z.B. dem Machine Learning zu sprechen und spiegelbildlich von einer pauschalen Transparenz bestimmter Systemtypen, etwa der symbolischen KI. Lipton vertritt ein vergleichbares Anliegen wie Burrell, doch anders als sie hebt er nicht auf die Kontextualisierung der Systeme ab, sondern auf das Zusammenspiel der technischen Komponenten:

> »[…] what constitutes transparency? You might look to the algorithm itself: Will it converge? Does it produce a unique solution? Or you might look to its parameters: Do you understand what each represents? Alternatively, you could consider the model's complexity: Is it simple enough to be examined all at once by a human?« (Lipton 2018: 6)

Lipton unterscheidet drei Level der epistemischen Opazität, die je nach technischem Zusammenspiel mehr oder weniger opak sein können: (a) »algorithmic transparency«, »decomposability«, »simultability«. Die algorithmische Transparenz bezieht sich auf die Struktur der verwendeten Algorithmen. Diese werden häufig in transparente und opake Methoden eingeteilt. Zum Beispiel gelten Entscheidungsbäume, lineare Regressionen oder regelbasierte Systeme als ›white box models‹, während neuronale Netze per se als opak gelten. Diese Einteilung ist aber zu hinterfragen. Den Grund, den Lipton anführt, ist der der Komplexität, denn es macht praktisch einen Unterschied, wie komplex Modelle bzw. Systeme werden. Mit der Komplexität steigt die Undurchsichtigkeit:

> »A lack of transparency and interpretability is arguably less problematic for other ML methodology with a stronger ›white-box‹ character, most notably symbol-oriented approaches such as rules and decision trees. Yet, even for such methods, interpretability is far from being guaranteed, especially because accurate models often require a certain size and complexity. For example, even if a decision tree might be interpretable in principle, a tree with hundreds of nodes will hardly be understandable by anyone.« (Hüllermeier 2020: 206)

Liptons »algorithmic transparency« lässt sich sinnvoll auf Muster-Algorithmen (z.B. in Lehrbüchern) anwenden, die dann in transparente und opake sortiert werden können. Diese Zuschreibung ist für konkretisierte Algorithmen dann zu relativieren. Bei der Nutzung von Algorithmen stellt sich die Frage, wie nachvollziehbar das Zusammenspiel der wichtigen Komponenten des algorithmischen Systems sind. Lassen sich »input, parameter and calculation« (Lipton 2018: 14) für sich und in ihrem Zusammenwirken verstehen? Dann hätten wir eine Transparenz im Sinne der »decomposibility«. Davon unterscheidet Lipton weiter den Aspekt der ›simultability‹. Hiermit bezieht sich Lipton auf die Frage, ob das Systemverhalten im Ganzen und simultan zu verstehen ist:

> »[…], for a model to be fully understood, a human should be able to take the input data together with the parameters of the model and in reasonable time step through every calculation required to produce a prediction.« (Lipton 2018: 13)

Unter dem Gesichtspunkt der »simultability« hat es keinen Sinn, davon zu sprechen, Entscheidungsbäume seien per se transparent, neuronale Netze aber nicht, weil hier Größe, Komplexität, Zusammenspiel und Zeit in Betracht kommen.

Zur ungenauen und pauschalen Problemstellung kommt hinzu, dass die anvisierte Problemlösung (Erklärbarkeit) ebenfalls zu undifferenziert bleibt (Krishnan 2020; Freiesleben/König 2023; Alpsancar et al. 2024). Transparenz ist kein Selbstzweck. Erklärungen sind nicht per se hilfreich, nötig oder nützlich. Die Debatte um erklärbare KI kränkelt an einer pauschalen Inwertsetzung von Erklärbarkeit. Krishnan (Krishnan 2020) hat die Existenz und Relevanz des Black-Box-Problems in Frage gestellt und angeregt, die XAI-Debatte stärker an Fragen der Zwecksetzung von Verstehbarkeit auszurichten. Diesen Vorschlag aufgreifend lässt sich zunächst nach den gängigen Zwecken fragen, denen XAI dienen soll. Da es bisher wenig Erfahrungsberichte aus der Praxis gibt, d.h. konkrete Fallstudien zur Tauglichkeit von XAI, sondern das Feld weitestgehend im Zustand hoher Erwartungen, florierender Forschung und spekulativen Begründungen vibriert, blicke ich dazu in die Forschungsliteratur. Viele Forschungsartikel im Feld der XAI präsentieren hier, wie sie bestimmte Techniken weiterentwickelt haben, stellen neue Ansätze des technischen Erklärens vor oder geben einen Überblick über das Forschungsfeld (Adadi/Berrada 2018; Ras et al. 2018; Meske et al. 2022; Gilpin et al. 2018). Typischerweise findet sich in diesen Artikeln im Einleitungsteil (bzw. den einleitenden Sätzen) die Motivation für diese Forschung, d.h. hier wird aus dem Feld heraus artikuliert, warum und wozu es erklärbare KI gibt, und diese weiterentwickelt werden muss. Die Motivation beginnt i.d.R. mit der generalisierten Problemstellung, d.h. der Undurchsichtigkeit von ML-Systemen. Teilweise wird diese Undurchsichtigkeit als solche als Anlass genug gesehen, gegen sie vorzugehen, ohne das weiter erläutert würde, warum diese problematisch ist:

> »Explainable Artificial Intelligence (XAI) has experienced a significant growth over the last few years. This is due to the widespread application of machine learning, particularly deep learning, that has led to the development of highly accurate models that lack explainability and interpretability.« (Vilone/Longo 2021: 89)

Darüber hinaus werden häufig instrumentelle Gründe angegeben, d.h. die Dienlichkeit von XAI benannt, i.d.R. allerdings in abstrakter und generalisierter Form:

> »This opacity has created the need for XAI architectures that is motivated mainly by three reasons, […] (i) the demand to produce more transparent models; (ii) the need of techniques that enable humans to interact with them; (iii) the requirement of trustworthiness of their inferences. Additionally, […], models induced

> from data must be liable as liability will likely soon become a legal requirement.« (Vilone/Longo 2021: 89)

Ähnlich formulieren es Kamath und Liu, die vier Gründe auflisten: Erstens geht es um den Nutzen, den ML-Entwickler:innen aus XAI ziehen können: »We need interpretability to explain the model's working from both the diagnosis and debugging perspective«. Zweitens geht es darum, die Ergebnisse bzw. das Verhalten von KI-Systemen für Endnutzer:innen verständlich zu machen, damit diese die Systeme überhaupt bzw. besser nutzen können: »We need explanations for the end-user to explain the decisions made by the model and the rationale behind the decisions« (Kamath/Liu 2021: ix). Als drittes Motiv für XAI heben sie das Problem der ›biases/Verzerrungen‹ hervor, die es aufzudecken gelte:

> »Most datasets or models have been shown to have biases, and investigating these biases is imperative for model deployment. Explainability is one way of uncovering these biases in the model.« (Kamath/Liu 2021: ix)

Als vierten Grund stellen sie rechtliche Vorgaben heraus, die aus Sicht von Betreiber:innen und Hersteller:innen ein Compliance-Problem darstellen:

> »Many industries such as finance and healthcare have legal requirements on transparency, trust, explainability, and faithfulness of models, thus making interpretability of models a prerequisite.« (Kamath/Liu 2021: ix)

Rechtliche Regulierungen zu KI-Systemen finden sich zahlreiche (Nannini et al. 2023; Cath et al. 2018; Chakrabarti/Sanyal 2020; Roberts et al. 2021). In der EU wird insbesondere mit Bezug auf den *EU AI Act*, der Ende 2023 verabschiedet wurde und dessen Umsetzung noch ausgearbeitet werden muss (Laux et al. 2023), sowie mit Bezug auf die im Jahr 2018 in Kraft getretene *General Data Protection Right* darüber diskutiert, ob diese Regulationen ein sogenanntes ›right to explanation‹ fordern oder doch nur schwächere Auflagen wie Transparenzpflichten für bestimmte Anwendungen, Risiken oder ein ›right to information‹ (Goodman/Flaxman 2017; M.E. Kaminski 2019; Sovrano et al. 2022; Gyevnar et al. 2023; Panigutti et al. 2023). Es ist noch nicht ausgemacht, ob und für welche Fälle Erklärbarkeit von KI eine notwendige Bedingung für ihre Anwendung darstellt, geschweige denn, was genau unter Erklärbarkeit technisch, sozial und politisch zu verstehen sei (Doshi-Velez/Kim 2017; Gilpin et al. 2018; Ribera/Lapedriza García 2019; Rudin 2019).

Beim Problem der ›biases‹ (Verzerrungen) geht es darum, Diskriminierungsrisiken zu minimieren (vgl. Kolleck/Orwat 2020, für einen Überblick). Man spricht hier von diskriminierender KI (›racist algorithm‹) oder, wenn keine diskriminierenden Verzerrungen vorliegen, von ›ethical aglorithms‹ bzw. ›fair ML‹. Insofern Dis-

kriminierungen aufgrund von geschützten Merkmalen wie Alter, Rasse, Ethnie, Geschlecht, sexuelle Orientierung und Religion rechtlich verboten sind, sind der dritte und der vierte Grund Kamaths und Lius miteinander verbunden.[2] Ich schlage vor, beide unter dem allgemeineren Gesichtspunkt des Wunsches nach einer Vereinbarkeit von KI mit anerkannten Grundwerten und Normen zu fassen. Erklärbare KI wird hier als ein Mittel angesehen, welches (entscheidend) dazu beitragen kann, *KI-Systeme gesellschaftsfähig zu machen*. Hierbei geht es sowohl um rechtliche Normen und ethische Prinzipien als auch Fragen der sozialen Angemessenheit (Lipton 2018; Bellon et al. 2022).

Mit diesen Überlegungen lassen sich die von Kamath und Liu (Kamath/Liu 2021) sowie von Vilone und Longo (Vilone/Longo 2021) genannten Gründe, die sich in der einen oder anderen Formulierung immer wieder in der XAI-Debatte finden (Samek et al. 2017; Ribera/Lapedriza García 2019; Gilpin et al. 2018), zu drei *Hauptmotiven* zusammenfassen, die besagen, warum und wozu erklärbare KI entwickelt wird (Alpsancar et al. 2024):[3]

a. um KI-Systeme zu optimieren,
b. um eine effiziente Nutzung von KI-Systemen zu gewährleisten,
c. um KI-Systeme gesellschaftsfähig zu machen.

In der Zwecksetzung a. artikuliert sich die Perspektive der Entwickler:innen und Forschenden, denen es darum geht, Systeme besser einschätzen, verändern und optimieren zu können, z. B. in der Fehlersuche (›debugging‹). Genealogisch scheint mir dieser Zweck am Ursprung der derzeitigen XAI-Debatte in den Computer Sciences zu liegen, auch wenn historisch gesehen Erklärungen für Informationssysteme schon älteren Datums und aus Anwendungsbezügen erwachsen sind, insbesondere für Expertensysteme im medizinischen Bereich (de Bruijn et al. 2022; Meske et al. 2022). Dem zweiten Hauptmotiv liegt eine anwendungsbezogene Perspektive zugrunde. Es geht es um den gelingenden Gebrauch, der Voraussetzung für die erhofften Effizienzsteigerungen ist. Auf diese Weise kommen Endnutzer:innen und

2 Die Rechtslage divergiert freilich von Staat zu Staat. In der EU hat z. B. der Europäische Rat, zwischen 2000 und 2004, vier Gleichbehandlungsrichtlinien beschlossen, die in Deutschland durch das Allgemeine Gesetz zur Gleichbehandlung (AGG) umgesetzt werden. Eine gute Übersicht zu dieser Debatte bieten Kraus und Ganschow 2022.

3 Freilich lassen sich weitere wichtige Kategorien an Zwecksetzungen finden, etwa die Dienlichkeit von XAI für die Forschung in anderen Bereichen wie der Medizin oder Biologie, wo es darum geht, neue Einsichten und Erkenntnisse zu gewinnen (Markus et al. 2021; Guidotti et al. 2018; Miller 2019). Darüber hinaus könnte man diese Zweck-Kategorien binnendifferenzieren, etwa in ›XAI for contestability‹ oder ›subjects understanding‹ (Mittelstadt et al. 2019). Beides soll hier aber außen vor bleiben.

teilweise auch Anwendungskontexte in den Blick. Hierin artikuliert sich ein ökonomischer oder auch militärisch motivierter Wille. Es geht um Effizienz, Akzeptanz und ›Usability‹. Das dritte Hauptmotiv evoziert eine kollektive Perspektive. Es geht um die Frage nach den Regeln für den angemessenen Gebrauch und Einsatz von KI in diversen gesellschaftlichen Bereichen. Konkret betrifft dies die Sprechposition derjenigen, die KI regulieren wollen oder sollen bzw. diese Vorgaben oder Überlegungen umsetzen müssen oder wollen. Auch schließt hier eine gesellschaftliche Perspektive an, und zwar als Frage, welcher Einsatz und Gebrauch der Technologien eigentlich wünschenswert wäre. Die Frage, ob man diese Technologien überhaupt braucht oder haben möchte, kommt nicht vor: Die Technik ist da, nun muss sie eingehegt werden.

Mit den folgenden Überlegungen möchte ich zu einer kritischen Auseinandersetzung mit diesem motivationalen Horizont von XAI anregen. Hierzu werde ich an exemplarischen Fällen drei Problemlagen herausstellen, aus denen sich die drei genannten Hauptmotive nähren. Sie bilden meiner Ansicht nach den Grund von drei Paradigmen, die den Sinnhorizont der Forschung und Entwicklung von XAI stiften und damit formen, was im XAI-Bereich passiert: das Paradigma der epistemischen Güte von ML, das Paradigma der effizienten Handhabbarkeit von KI-Anwendungen sowie das Paradigma der Vereinbarkeit von KI mit anerkannten Grundwerten und Normen. Meine These ist, erstens haben wir es mit *drei verschiedenen Paradigmen* zu tun und zweitens wurde diese Verschiedenheit von der XAI-Forschung bislang zu wenig beachtet.

Lose angelehnt an Kuhn (Kuhn 1976), prägen Paradigmen in einer Forschungsgemeinschaft das, was als ein typisches und damit relevantes Forschungsproblem ist. Es gibt vor, worin zentrale Forschungsaufgaben liegen und wie man sich an deren Lösung machen sollte. Paradigmen haben in ihrer Vorbildfunktion einen normativen Charakter. Kuhn beschreibt sie als »allgemein anerkannte wissenschaftliche Leistungen, die für eine gewisse Zeit einer Gemeinschaft von Fachleuten maßgebende Probleme und Lösungen liefern« (Kuhn 1976: 10). So wie in einer Forschungsgemeinschaft ein Problem x durch das Lösungsverfahren z gelöst wurde, so geht man nun an neue Probleme in dem Forschungsgebiet heran. Das heißt, was als Forschungsproblem erkannt und wie es konzipiert wird, worin seine Herausforderung liegt, ist maßgeblich von den geltenden Paradigmen einer Forschungsgemeinschaft geprägt. Das Erkennen und Konzipieren von relevanten Forschungsproblemen gehen mit den Erwartungen einher, diese prinzipiell lösen zu können. Die Lösungsvisionen sind mehr als ein erstes Erkunden, sie sind zielgerichtetes Entwickeln und Testen, d.h. es existieren bestimmte strategische Vorstellungen darüber, wie die Forschungsprobleme zu meistern seien. Mit diesen formierenden Erwartungen und Zuschreibungen bündeln Paradigmen nicht nur das begriffliche und apparative Instrumentarium (Kuhn 1976: 41), sondern ebenso entsprechende Akteur:innen, Ressourcen und Infrastrukturen um sich

(Borup et al. 2006). Ich verstehe hier Paradigmen darüber hinaus als sinnstiftend und formgebend, in dem sie mit typischen Zwecken korrelieren, denen sich eine Forschungsgemeinschaft sinnvollerweise widmen kann. Bei Kuhn stand die Frage nach den Zwecken und Motiven der Forschung weniger im Vordergrund, da er seine Überlungen an der Geschichte der modernen Physik entwickelte, für die er verschiedene Phasen und damit Paradigmen der Forschung ausmachen konnte, die sich aber weitestgehend einem gemeinsamen obersten Zweck verschrieben hatten – die Natur zu erkennen. Im Fall von XAI haben wir es dagegen mit verschiedenen Zwecken zu tun. Meinem Eindruck nach wird dieser Verschiedenheit bloß diskursiv Rechnung getragen, während die Forschungspraxis weitgehend im Modus des normalwissenschaftlichen Rätsellösens der ML-Community verweilt. Mit Kuhn gesprochen, isoliert sich die XAI-Community in dieser Weise von solchen Problemen, »die sich nicht auf die Rätselform reduzieren lassen« (Kuhn 1976: 51), d.h. sowohl Fragestellungen als auch Lösungswege, Methoden, Werkzeuge, theoretische Annahmen, die außerhalb des eigenen Paradigmas liegen, werden abgelehnt, als nicht wichtig erachtet oder für zu schwierig gehalten.

2. Das Paradigma der epistemischen Güte von ML

Gut laufende Maschinen sind nicht erklärungsbedürftig. Aus einer Ingenieurs-Perspektive werden Sie es dann, wenn man Sie optimieren möchte oder wenn ein Fehler auftritt, den man beheben will. Anlass vieler Diskussionen ist dies das Problem des »overfitting«. Dieses bezeichnet die Überanpassung eines ML-Modells zu den Trainingsdaten im Vergleich zu anderen Datensätzen. Liegt eine Überanpassung vor, kann aus einer hohen Treffsicherheit auf den Trainingsdaten nicht auf eine hohe Treffsicherheit für andere Datensätze geschlossen werden. Damit ist die gute Performanz nicht verallgemeinerbar, weil die Maschine nicht tatsächliche Muster, sondern willkürliche gelernt hat. Dieser Effekt lässt sich ebenfalls als Clever-Hans-Effekt bezeichnen (Hernández-Orallo 2019; Lapuschkin et al. 2019), womit Beziehungen gemeint sind, die man fälschlicherweise als Kausalitäten einschätzt, die in Wirklichkeit bloße Korrelationen darstellen. Einen solchen Effekt überhaupt feststellen zu können setzt eine sogenannte ›ground truth‹ voraus, d.h. ein Wissen darüber, welche Muster den Tatsachen entsprechen und welche nicht (Freiesleben/König 2023).[4]

Als Beispiel für dieses Paradigma greife ich den häufig zitierten Artikel von Lapuschkin et al. auf, in dem die Autoren Strategien vorstellen, um ML-Modelle auf

4 Das Festlegen dieser Grundwahrheit ist nicht immer möglich und auch nicht immer gefragt, z.B. wenn es darum geht, neue Zusammenhänge und Muster zu erkennen. Sie praktisch festzulegen ist aufwendig, da die Datensätze entsprechend annotiert werden müssen.

Clever-Hans-Effekte zu prüfen (Lapuschkin et al. 2019). Der Name ›Clever-Hans-Effekt‹ geht auf die Aufdeckung eines Versuchsleiter-Erwartungs-Effekts in der Psychologie des frühen 20. Jhs. zurück. Der Mathematiklehrer Wilhelm von Osten hatte ein Zirkuspferd namens Hans darauf dressiert, richtige Antworten auf beispielsweise einfache Rechenaufgaben zu geben (etwa durch eine Anzahl von Hufschlägen), was er einem staunenden Publikum vorführte. Doch Hans konnte nicht rechnen. Eine Prüfung zeigte, dass das Pferd sein Antwortverhalten an der Körpersprache und anderen Signalen seines Besitzers orientierte. Die richtigen Ergebnisse basierten somit nicht auf einer korrekten Durchführung der Aufgabenstellung, sondern hatten kontingente Ursachen. Im Bereich der KI-Forschung spricht man von einem Kluger-Hans-Effekt:

> »[…] bzw. Clever-Hans-Effekt, wenn in einem Trainingsdatensatz, möglicherweise in versteckter Form, bestimmte Eingangsgrößen vorhanden sind, die mit der richtigen Ausgabe korrelieren, aber wenig mit der Ursache der jeweils adressierten Phänomene zu tun haben.« (Kraus/Ganschow 2022: 39)

Lapuschkin et al. haben einen solchen Effekt bei ML-Modellen nachgewiesen, die beim PASCAL VOC (Visual Object Classification) Wettbewerb gewonnen hatten (Lapuschkin et al. 2019). Es ist nicht trivial, Maschinen so zu konstruieren, dass sie ein Objekt klassifizieren können (Tisch, Stuhl, Tier, Person). Während Menschen dies beiläufig tun (Kaminski 2014), ist eine Objektklassifikation für Maschinen eine solch ausgezeichnete Leistung, dass es in der Informatik üblich ist, hierfür Wettbewerbe zu veranstalten. Dass man im Bereich der ›Computer Vision‹ überhaupt zu vertretbaren Ergebnissen gekommen ist, liegt hauptsächlich an der Vergrößerung der Trainingsdatensätze. In diesem Bereich lässt sich die Performanz-Steigerung der Maschinen aufgrund der gewachsenen Datensätze gut nachvollziehen: In den 1960er Jahren standen Forschenden typischerweise ein Dutzend Bilder zur Verfügung, mit denen sie ihre Maschinen trainieren konnten. In den 1990er Jahren waren es bereits Sätze mit Tausenden von Bildern. Mit Beginn der 2020er Jahre bewegt man sich üblicherweise im Millionenbereich (Crawford 2023: 1368).

Hinter den von 2005/06 bis 2012 organisierten PASCAL VOC Wettbewerben steht das von der EU geförderte Exzellenznetzwerk PASCAL[5] (Everingham et al. 2010; Everingham et al. 2015). Ich werde zunächst die Logik des PASCAL Projektes und der Wettbewerbe beschreiben, um die Dienlichkeit von XAI im Bereich der Bilderkennung zu verorten. Das Exzellenznetzwerk hatte der eigenen Forschungsgemeinschaft zunächst eine Infrastruktur für die Erprobung von ML-Techniken im Bereich der Bilderkennung bereitgestellt. Kern der Infrastruktur war ein frei zugänglicher Datensatz aus Bildern, die annotiert waren. Über eine dazugehörige

5 PASCAL steht für pattern analysis, statistical modelling und computational learning.

Software stellte die Forschergruppe außerdem Standards für die Evaluierung von Algorithmen bereit, die die Bilder aus dem Datensatz für das Lösen bestimmter Aufgaben im Bereich der Bilderkennung nutzen sollten (Everingham et al. 2010). Wichtig war, dass die Bilder annotiert und damit mit einer ›ground truth‹ versehen waren, d.h. es war bekannt und notiert, was auf den Bildern zu sehen ist (Everingham et al. 2015: 98f.). Seit 2006 wurde jährlich ein neuer Datensatz mit annotierter Grundwahrheit bereitgestellt.

Everingham et al. (Everingham et al. 2010) beschreiben, wie sie das Datenset für den Wettbewerb im Jahr 2007 erstellt und kuratiert haben. Welche Bilder aufgenommen und wie diese annotiert werden, beeinflusst auf entscheidende Weise, wie gut die gestellten Forschungsfragen beantwortet werden und die Antwortlösungen verglichen werden können. Die Güte der generierten Datensätze richtete sich nach der formulierten Absicht des Projektes (im Vergleich zu anderen verfügbaren Bild-Datensätzen). Ziel von PASCAL war es, ML-Modelle für ein möglichst breites Spektrum ›natürlicher Bilder‹ trainieren und testen zu können (Everingham et al. 2010: 305), was damals eine neue Herausforderung für die Forschungscommunity war. 2006 hat die Forschergruppe Bilder aus der Microsoft Research Cambridge Datenbank verwendet, die allerdings zu gestellt waren, um der Zielstellung ›natürliche Bilder‹ gerecht zu werden.[6] Seit 2007 haben die Initiatoren deswegen ihren Datensatz über Flickr generiert, wodurch eine Ausgangsbasis an Bildern gegeben war, die nicht alle gleichermaßen unter einer bestimmten Absicht erstellt und hinterlegt sind.

Die Annotation der gewonnenen Bilder geht von folgenden Grundsätzen aus: Jedes Bild wird einer der vorab festgelegten Objektklassen zugeordnet (für 2007 waren diese: »aeroplane, bird, bicycle, boat, bottle, bus, car, cat, chair, cow, dining table, dog, horse, motorbike, person, potted plant, sheep, sofa, train, tv/monitor«; Everingham et al. 2010: 308). Hinzu kommt ein Begrenzungsrahmen (»bounding box«), die den Umfang des Objektes sichtbar eingrenzt (Everingham et al. 2010: 308). Seit 2006 wurden weitere Annotationen eingeführt, die für das Training genutzt werden durften (aber für die Evaluierung nicht notwendig waren): »viewpoint«, »truncation« (wenn der Begrenzungsrahmen nicht das ganze Objekt erfasst, weil auf dem Bild z.B. nur ein Teil einer Person sichtbar ist etc.). Ab 2008 wurde auch ggf. »occluded« angegeben, wenn das Objekt (in Teilen) verdeckt sichtbar ist. Zudem wurden bestimmte Bilder als »difficult« markiert, wenn ihre Eigenschaften es im Vergleich zu anderen schwieriger für Algorithmen machen, Objekte zu erkennen (Sind alle Tiere auf dem Bild Kühe oder ist eines ein Schaf?). Diese Annotation sollte konsistent, akkurat und vollständig sein, weswegen die Annotator:innen geschult und der

6 Der Microsoft Research Cambridge-Datensatz war mit der Absicht gebaut worden, bestimmte Objektklassen möglichst gut abzubilden. Entsprechend zeigen die Bilder i.d.R bestimmte Objekttypen in zentrierter, ausgeleuchteter, zentraler Position. Sie haben damit wenig Varianz für die Zwecke des PASCAL Projektes.

Prozess des Annotierens überwacht und geprüft wurde. Annotieren ist ein sehr aufwendiger und je nachdem, von wem man diese Arbeit durchführen lässt, auch kostspieliger Vorgang.[7] Die Forschungsgruppe wollte die Annotationsarbeit zunächst vollständig auslagern, über die Amazon-Plattform »Mechanical Turk«, über die Arbeitskräfte günstig für kleinteilige Arbeit der digitalen Ökonomie zugänglich gemacht werden.[8] Für die hohen Ansprüche an die Annotation reichte es aber nicht aus (Konsistenz, Akkuratheit, Vollständigkeit), so dass nur Label für die Frage eingekauft wurden, ob eine Objektklasse zu sehen ist oder nicht (presence/absence), wodurch dennoch erheblich Zeit und Kosten gespart wurden (Everingham et al. 2015: 101).

Zu dem Projekt gehörte, neben der Bereitstellung der Forschungsinfrastruktur, ein jährlicher Wettbewerb, der PASCAL VOC Wettbewerb. Diese Ergebnisse wurden auf einem jährlichen Workshop zur Diskussion gestellt. Um am Wettbewerb teilnehmen zu können, mussten die Forscher:innen Algorithmen entwickeln, die zu Beginn (2005/06–2008) mindestens eine der zwei folgenden »principle challenges« lösen sollten: (1) Klassifizierung: »For each of twenty object classes, predict the presence/absence of at least one object of that class in a test image«;[9] (2) Detektion: »For each of the twenty classes, predict the bounding boxes of each object of that class in a test image (if any), with associated real-valued confidence« (Everingham et al. 2010: 304). Diese Kernfragen wurden um zwei zusätzliche Fragen ergänzt, die man aber nicht angehen musste, um am Wettbewerb teilzunehmen: (a) Segmentierungs-Probe: »For each test image, predict the object class of each pixel«, und (b) Personen-Gliederungs-Probe: »For each ›person‹ object in a test image (if any), detect the person, predicting the bounding box of the person, the presence/absence of parts (heads/hands/feet), and the bounding boxes of those parts« (Everingham et al. 2010: 305).[10] Der Wettbewerb bestand aus zwei Phasen: einer Trainings- und einer Testphase. Für das Training konnten die annotierten Datensets verwendet werden, für das Testen wurden un-annotierte Bilder herausgegeben (also Bilder, die eigentlich annotiert waren, aber deren Annotation absichtlich für die Wettbewerber entfernt

7 »For example, annotation of the VOC2008 dataset required around 700 person hours.« (Everingham et al. 2010: 336)

8 Typische Arbeiten sind: transkribieren, Objekte klassifizieren, Rückmeldungen zu Webseiten geben, Onlineinhalte umschreiben (Ipeirotis 2010). Die über die Plattformlogik forcierten Arbeitsbedingungen stehen häufig in der Kritik (vgl. exemplarisch Ellmer 2015).

9 Diese Klassifizierungsaufgabe läuft darauf hinaus, die Pixel richtig zu zuordnen: »to which class does each pixel belong?« (Everingham et al. 2015: 99).

10 Ab 2009 bis 2012 wurde die Segmentierungsaufgabe zur dritten Kernfrage gemacht. Der Test, wie gut Algorithmen bestimmte Körperteile und den Aufbau von menschlichen Körpern erkennen können (bounding boxes), blieb Zusatzaufgabe. Ab 2010 kam die Klassifizierung von Handlungen als weitere Zusatzaufgabe hinzu (Everingham et al. 2015: 100).

wurde), so dass Training und Test auf (logisch) verschiedenen Datensätzen erfolgten. Die Annotationen des Test-Datensets wurden absichtlich bis zum Abschluss des Wettbewerbs nicht publik gemacht, um das Phänomen des ›overfitting‹ vorzubeugen.[11]

Die Aufgaben dieser Wettbewerbe lassen sich als leitende Forschungsfragen der Forschungscommunity zu dieser Zeit verstehen. Die Bereitstellung der »ground truth« über die Annotation stellt sicher, dass man für diese Fragen die richtigen Antworten kennt. Entsprechend ließ sich nicht nur die performative Güte jedes eingereichten Algorithmus bewerten, sondern diese waren untereinander vergleichbar. Hiermit wurde einschätzbar, welche algorithmischen Methoden für welche Aufgaben zielführend sind. Die Ergebnisse aus der Serie von Wettbewerben und Workshops stellen zusammen eine wichtige ›Benchmark‹ für den Bereich der Computer Vision dar (Lapuschkin et al. 2019; Samek/Müller 2019; Hernández-Orallo 2019).

Das Problem des Clever-Hans-Effekts betrifft die Güte von ML-Modellen und damit zunächst die Frage ihrer epistemischen Evaluation. Wie zuverlässig sind die Aussagen, die man mit dem ML-Modell gewinnen kann? Zur Evaluierung von ML-Systemen gibt es verschiedene Gütekriterien bzw. statistische Evaluationsmetriken. Wichtig ist, dass sich die Güte der Systeme auf das Zusammenspiel des trainierten ML-Modells und den gegebenen Daten bezieht. Anders als bei konventionellen Algorithmen ist die Güte dieser algorithmischen Systeme deswegen entscheidend von den Daten abhängig. Hierbei sind Trainingsdaten, Validierungsdaten und der Testdatensatz zu unterscheiden.

Um die Rolle von XAI für die Evaluation der epistemischen Güte von ML-Systemen besser einschätzen zu können, vergleiche ich diese mit einem simplen Maß für die Fehlerquote, der *accuracy* (ACC).[12] Bei einer Klassifikationsaufgabe gibt das Maß der Akkuratheit (ACC) an, wie viele Bilder das System richtig klassifiziert hat.

$$ACC = \frac{Anzahl der korrekt klassifizierten Bilder}{Gesamtzahl der zu klassifizierenden Bilder}$$

Eine Akkuratheit von 100% sagt aus, dass das System jedes Bild im vorliegenden Datensatz richtig klassifiziert hat. Dies wäre die bestmögliche *Performanz*. Bei einer simplen Klassifikationsaufgabe mit nur zwei möglichen Antworten (Ist auf dem

11 Die Initiatoren gehen davon aus, durch den geringen Zeitraum für das Testen zu unterbinden, dass die Teilnehmenden ihre Testdaten händisch annotieren.

12 »Accuracy plays a role especially for data whose factual correctness can be conclusively determined and whose meaning is not ambivalent.« (Mohammed et al. 2024: 2)

Bild ein Pferd zu sehen – ja oder nein?), stellt eine Akkuratheit von 50% eine schlechte Performanz dar, da dies einer bloß zufälligen Zuordnung gleichkommt. Um die Akkuratheit eines Systems mit dieser Formel berechnen zu können, muss man wissen, welche Antworten für jedes gegebene Bild die richtige ist (dies war im PASCAL VOC Wettbewerb der Fall, auch wenn die Klassifikation hier komplexer war). Außerdem muss man die jeweiligen Ergebnisse des zu evaluierenden Systems kennen. Für die Berechnung der Akkuratheit ist demnach keine Einsicht in die Black-Box nötig. Die Art und Weise, wie das Ergebnis gewonnen wurde spielt keine Rolle. Auch ein Zirkuspferd wie Hans könnte demnach eine hohe *accuracy* aufweisen ohne Rechnen zu können. Im Gegensatz zu Gütekriterien für die Fehleranfälligkeit nutzt man erklärbare KI, um einen Einblick in die Frage zu erhalten, wie ein System zu seinem (richtigen) Ergebnis gekommen ist. Dies ist deswegen relevant, weil eine gute Performanz aus falschen Ursachen auf Datensatz_1 zu einer schlechten Performanz auf einem Datensatz_2 führen kann. Es geht um die Generalisierbarkeit der Nutzung der gewonnenen ML-Modelle über die Trainingsdaten (und ggf. Validierungsdaten) hinaus.

Lapuschkin et al. untersuchten zwei der erfolgreich am Wettbewerb teilgenommen Modelle für die Kategorie Objektklassifikationen (Lapuschkin et al. 2019): das Fisher Vectors (FV) Modell und ein vortrainiertes Deep Neuronal Network (DNN), welches die Autorengruppe für ihre Zwecke auf dem PASCAL VOC Datensatz von 2007 ›feingetuned‹ haben. Beide Modelle bewiesen »excellent state-of-the-art test set accuracy on categories, such as ›person‹, ›train‹, ›car‹, or ›horse‹ of this benchmark« auf (Lapuschkin et al. 2019: 4). Um eine Einsicht darin zu erhalten, wie die Modelle zu diesen guten Leistungen gekommen sind, setzten die Autoren die Methode der »Layer-wise relevance propagation« ein, in der sogenannte Heatmaps visualisieren, welche Zonen eines Bildes (Pixel) besonders einflussreich für die Klassifizierung des Modells sind (›Entscheidung‹). Die Heatmap für das DNN-Modell hebt die Zone hervor, auf der tatsächlich jeweils das Pferd auf den Bildern zu sehen ist. Dies spricht dafür, dass dieses Modell die Pferde aufgrund der richtigen ›Ursachen‹ korrekt klassifiziert. Die Heatmap des FV-Modells hingegen hebt eine Zone unten links auf den Bildern hervor, auf dem nicht die Pferde, sondern ein »source tag« zu sehen ist.[13] Aufgrund dieses Befundes sind die Forschenden die Pferdebilder ›händisch‹, d.h. mit menschlichen Augen durchgegangen und konnten bestätigen, dass sich diese Herkunftsangabe auf allen Pferdebildern in dem Datensatz befindet. Hiernach stellten sie den Verdacht auf, »the FV model has ›overfitted‹ the PASCAL VOC dataset by relying mainly on the easily identifiable source tag, which incidentally correlates with the true features, a clear case of ›Clever Hans‹ behavior« (Lapuschkin

13 Es ist ein schöner historischer Zufall, dass der Clever-Hans-Effekt hier bei der Klassifikation von Pferdebildern auftrat.

et al. 2019: 4). Um ihren Verdacht zu erhärten, haben die Autoren die Herkunftsangabe künstlich aus den Pferde-Bildern entfernt. Während das DNN-Modell gleich performte wie zuvor, büßte das FV-Modell signifikant an Performanz ein. Mehr noch, fügt man die Herkunftsangabe künstlich zu Auto-Bildern hinzu, klassifiziert das FV-Modell diese dann als Pferde: »a clearly invalid decision« (Lapuschkin et al. 2019: 4). Es liegt nahe, dass das trainierte FV-Modell mit hoher Wahrscheinlichkeit alle Bilder als Pferde klassifiziert, die das gleiche »resource tag« aufweisen. Mit diesem Beispiel geben Lapuschkin et al. zu bedenken, sich nicht ausschließlich auf gute Performanzergbnisse zu verlassen, sondern weitere epistemische Gütekriterien zur Evaluation von ML-Modellen heranzuziehen. Es muss die Frage gestellt werden, wie valide eine hohe Performanz eines Systems auf einem Datensatz$_x$ mit Blick auf andere Datensätze ist. Neben den Pferdebildern zirkulieren weitere eingängige Beispiele für Clever-Hans-Effekte in der Forschungsdiskussion zu XAI, etwa Huskies, die von Wölfen aufgrund von Schnee im Hintergrund der Bilder klassifiziert wurden (Ribeiro et al. 2016), oder auch »boats by the presence of water and trains by the presence of rails in the image« (Samek/Müller 2019; Cremers et al. 2019). Die Klassifikationsleistung aufgrund kontingenter Merkmale wird als Fehleinschätzung eingestuft, weil das korrekte Ergebnis aus falschen Gründen erbracht wird und damit die Klassifikationsleistung nicht auf andere Datensätze übertragbar ist. Wichtig ist hierbei, dass die richtige Klassifizierung in diesen Fällen dadurch klar wird, dass sich die korrekte Zuordnung an unserer (menschlichen) Klassifikation der Bilder orientieren kann. Für uns ist es i.d.R. völlig unproblematisch Wölfe von Huskies usw. zu unterscheiden. Komplizierter wird die Frage der epistemischen Güte von Klassifikationsleistungen, wenn nicht bekannt ist, welche Zuordnung die richtige ist bzw. welche Klassen es überhaupt gibt (Caruana et al. 2015).

Die Übertragbarkeit auf andere Datensätze wirft die Frage auf, wie verlässlich ML-Systeme in der realen Welt eingesetzt werden können (Kraus/Ganschow 2022: 39). Genauer sind zwei Fragen nach der Repräsentativität der Daten zu stellen, zum einen, wie angemessen repräsentieren Trainings- und Validierungsdatensätze die eigentlichen Daten der angedachten Anwendungsfelder, und zum anderen, wie gut repräsentieren die Daten die Wirklichkeit, für die sie stehen sollen. In diesen einfachen Fällen der Bildklassifikation lässt sich mit XAI-Methoden die Einsicht gewinnen, ob die ML-Modelle aus den ›richtigen Ursachen‹ gut performen oder nicht. Sie fügen sich in dieser Perspektive in den Werkzeugkasten ein, der dazu dient, die epistemische Güte von KI-Systemen zu bewerten. Als ein solcher Toolkasten ist XAI von Expert:innen für Expert:innen gemacht. Diese Bemühungen verlaufen im Rahmen dessen, was sich mit Kuhn (Kuhn 1976) als Normalwissenschaft bezeichnen lässt: Es gibt innerhalb der Community eine konsensuelle Problemstellung (Einsicht in die schwarzen Kisten zu erlangen, um deren epistemische Güte besser einschätzen zu können). Mir scheint, der größte Teil der XAI-Forschung arbeitet in diesem Paradigma und löst Rätsel dieses epistemischen Typs. Soll XAI anderen Zwecken dienen,

sollte man den modus operandi wechseln, etwa indem man sich ernsthaft realen Anwendungskontexten und damit diversen Nutzer:innen zuwendet.

3. Das Paradigma der effizienten Nutzung von KI-Systemen

Der zweite Zweck, der die XAI-Forschung (diskursiv) entscheidend motiviert, ist die Herausforderung, KI-Systeme in der Anwendung effizient handhaben zu können. Mit dem Beispiel des vom US-Verteidigungsministerium geförderten und über die »Defense Advanced Research Projects Agency« (DARPA) aufgesetzten Forschungsprojekts *XAI* möchte ich demonstrieren, dass diese Zwecksetzung einen Paradigmenwechsel in der Erforschung und Entwicklung von XAI erfordert.

Das DARPA-Projekt stellt eine wichtige Referenz in der Forschungsgemeinschaft dar und hat die Bezeichnung »XAI« maßgeblich verbreitet. Es kann als eine der ersten großflächigen Initiativen einer Regierungsorganisation angesehen werden, das Feld zu rahmen und voranzutreiben (Nannini et al. 2023: 1202; Hoffman et al. 2023). In der Retrospektive der leitenden Köpfe stellt es einen Meilenstein in der XAI-Forschung dar: »The program certainly acted as a catalyst to stimulate XAI research (both inside and outside the program)« (Gunning et al. 2021: 8f.). Das DARPA-Projekt geht über die Motivation von erklärbarer KI durch epistemische Fragestellungen hinaus, da explizit die Perspektive der ›End-User‹ adressiert wurde:

> »The stated goal of explainable artificial intelligence (XAI) was to create a suite of new or modified machine learning techniques that produce explainable models that, when combined with effective explanation techniques, enable end users to understand, appropriately trust, and effectively manage the emerging generation of AI systems.« (Gunning et al. 2021: 1)

Erklärbare KI sollte nicht nur von Expert:innen für Expert:innen entwickelt werden, sondern auch für Primäruser:innen. Das DARPA-Projekt stellt somit einen Perspektivwechsel in der Zwecksetzung von XAI dar, welches die Aufmerksamkeit der Forschung umlenkt, jedenfalls rhetorisch: Es geht um den realen Einsatz von KI-Systemen in diversen Anwendungsfeldern. Man verlässt den geschützten Rahmen der informatischen Laborforschung und Wettbewerbe. Mit den Usern kommt die *Praxis* ins Blickfeld, oder anders gesagt: Man wendet sich rhetorisch zugleich den Nutzer:innen und der Praxis zu.

An dem Projekt, das über vier Jahre lief (2017–2021), waren zwölf Forschergruppen verschiedener US-Amerikanischer Hochschulen und Forschungseinrichtungen beteiligt. Elf dieser Forschungsgruppen hatten einen informatischen Hintergrund (insbesondere aus den Communities des Machine Learning und der Human-

Computer-Interaction (HCI)). Hinzu kam ein Team, das sich der »psychology of explanation« (DARPA 2016: 15) widmen sollte. Für das Gebiet der HCI ist die Hinwendung zu Enduser:innen keineswegs neu (Biran/Cotton 2017); tatsächlich hat sich dieser Bereich der Computer Science herausgebildet, um Computersysteme für Nicht-Expert:innen nutzbar zu machen, etwa im Zuge der Entwicklung und Verbreitung von Minicomputern, dann Personal Computern über die Gestaltung von leichten handhabbaren Schnittstellen wie dem Graphical User Interface, Tastatur und Maus (Dix 2017; Petrick 2020; Harrison et al. 2007). Interessant ist aber, dass sich eine Diskrepanz zwischen der ›informatischen‹ Projektkonzeption und der Perspektive des Teams von Robert R. Hoffman, das für die »psychology of explanation« zuständig war, zeigen lässt. Die ›informatischen‹ Projektkonzeption ziehe ich maßgeblich aus der Ausschreibung des Verteidigungsministeriums zu dem Projekt des *Broad Agency Announcement – Explainable Artificial Intelligence (XAI) – DARPA-BAA-16-53*, vom 10.08.2016 (DARPA 2016). Die DARPA hatte David Gunning als Programmmanager berufen und ihn später zum Direktor des Projektes gemacht, der die wissenschaftliche Ausrichtung verantwortet und Experte in der KI-Forschung ist. Die Ausschreibung wird durch Folien von Gunning flankiert (*Distribution Statement »A«*), mit denen er vermutlich bei interessierten Partnern aus der Industrie und Forschung das Programm des Projektes vorgestellt hat (Gunning 2016). Hinzu kommt eine Publikation zum Ende der ersten Phase (Gunning/Aha 2019), in der Zwischenergebnisse berichtet werden, sowie die Retrospektive nach Abschluss des Projektes (Gunning et al. 2021). Die Beteiligung aus dem HCI-Bereich wird hier in einer bestimmten Weise, nämlich additiv, integriert, während die Gruppe um Hoffman dezidiert eine Sonderrolle zukommt.

Ich interpretiere die Diskrepanz zwischen Hoffmans Team und der informatischen Projektkonzeption so, dass letztere zwar rhetorisch eine Hinwendung zu realen Anwendungskontexten und Enduser:innen formuliert, praktisch jedoch im Modus der informatischen Normalwissenschaft operierte.[14] Hoffmans Team hingegen war vermutlich in einer ambivalenten Rolle: einerseits galt es die zugewiesenen Aufgaben zu erfüllen, andererseits hätte dieses Team die Hinwendung zur Praxis sicherlich anders aufgezogen. Die Differenz der Perspektiven von Gunning und Hoffman dient mir als Demonstration dazu, dass es sinnvoll ist, die Hinwendung zu User:innen und zur Praxis als einen Paradigmenwechsel in der XAI-Forschung zu verstehen, mit dem die Problemstellung, für die XAI eine gute Lösung sein will, neu zu fassen ist.

14 Tatsächlich kommentiere ich nicht die Forschungspraxis oder -ergebnisse, sondern die Konzeption des XAI-Projektes. Das heißt, ich Vergleiche die informatische Selbstbeschreibung von Gunning und wechselnden Ko-Autor:innen mit der von Hoffmans Team.

3.1 Informatische Projektkonzeption

Ich stelle hier die Motivation für das Projekt, die gestellten Forschungsfragen und daran die gebundene Konzeption des Untersuchungsgegenstands heraus sowie die anvisierte Zusammenarbeit bzw. Arbeitsteilung der Forschungsgruppen.

3.1.1 Motivation

Warum setzte die DARPA ein Projekt zu erklärbarer KI auf und warum im Jahr 2016? Das Kernmotiv, welches in den genannten Quellen zu finden ist, ist zweiseitig: Die eine Seite besteht aus einer zweifachen Charakterisierung der ML-Technik, zunächst deren jüngster Erfolg: »Dramatic success in machine learning has led to an explosion of new AI capabilites« (DARPA 2016: 5). Dieser Erfolg wird allein auf die Technologie bezogen (ML); kein Wort über die entscheidenden Randbedingungen (Big Data, Rechenkapazitäten, Verfügbarkeit von Services und Tools, Einbettung und Handhabe der Systeme). Zudem wird, wie in dem eingangs zitierten Lehrbuch zu XAI, das Forschungsgebiet der erklärbaren KI allein aus einem Manko der jüngeren Technologie hergeleitet: »These systems offer tremendous benefits, but their effectiveness will be limited by the machine's inability to explain its decisions and actions to human users« (DARPA 2016: 5). Die informatische Perspektive fokussiert einseitig auf das technische System.

Die andere Seite des Kernmotivs besteht in der Zielstellung der *effektiven Nutzbarkeit* von KI durch Laien:

> »Defense Advanced Research Projects Agency (DARPA) formulated the explainable artificial intelligence (XAI) program in 2015 with the goal to enable end users to better understand, trust, and effectively manage artificially intelligent systems.« (Gunning et al. 2021: 1)

Die Rede davon, dass Nutzer:innen den neuen KI-Systemen »appropriately trust, and effectively manage« (Gunning/Aha 2019: 44) können sollen, findet sich immer wieder in den Quellen zum Projekt (Gunning et al. 2019) und kann als übergeordnete Zielstellung aufgefasst werden. Damit ist die Hinwendung zu Usern/der Praxis prominent platziert.

Warum aber gerade das Verteidigungsministerium hohe Summen in die XAI-Forschung investieren wollte, wird nicht weiter erläutert. Bekannterweise ist die DARPA und das US-Militär generell eine der kontinuierlichen Förderer der Computertechnologien in den USA (Norberg 1996; Edwards 1997; Mahoney 2011), aber welches Anliegen konkret durch die Forschung zur erklärbaren KI befriedigt werden soll, bleibt äußerst vage:

> »The issue [XAI zu entwickeln – SA] is especially important for the Department of Defense (DoD), which is facing challenges that demand the development for more intelligent, autonomous, and symbiotic systems. Explainable AI will be essential if users are to understand, appropriately trust, and effectively manage this incoming generation of artificial intelligent partners.« (DARPA 2016: 5)

Die Ausschreibung artikuliert einen allgemeinen Bedarf an technischem Fortschritt, der an Erklärbarkeit gekoppelt wird ohne über konkrete Nutzungsweisen oder Gründe zu sprechen. Problematisch ist hier nicht die mangelnde Zwecksetzung (Freiesleben/König 2023), sondern die *mangelnde Konkretisierung* der Zwecke hinsichtlich ›user‹ und ›context‹. Der unterstellte allgemeine Bedarf des technischen Fortschritts, der XAI quasi naturwüchsig umfasst, wird sodann gekoppelt an die Hinwendung zum ›User‹ – was wiederum abstrakt bleibt. Mit der Hinwendung zum User und zur Praxis wird somit einerseits ein neuer Gegenstand des Forschungsinteresses formal instituiert, der anderseits bemerkenswert vage bleibt. Inhaltlich findet sich eine zaghafte Konkretisierung von Usern/Praxis in zwei Richtungen: Zum einen werden zwei spezifische Anwendungsfälle genannt, die als ›Use-cases‹ in dem Forschungsprojekt fungieren sollten, zum anderen wird auf viele weitere Anwendungsfelder pauschal verwiesen.

Der erste Anwendungsfall ist ein »intelligence analyst who receives recommendations from a big data analytics system«. Um seinen Beruf gut auszuüben, so die Zuschreibung, muss sie diese Empfehlung nachvollziehen können (Gunning et al. 2021: 2). Dieser Fall stehe für ein bekanntes Problem in der Praxis:

> »The data analytics challenge was motivated by a common problem: intelligence analysts are presented with decisions and recommendations from big data analytics algorithms and must decide which to report as supporting evidence in their analyses and which to pursue further. These algorithms often produce false alarms that must be pruned and are subject to concept drift. Furthermore, these algorithms often make recommendations that the analyst must assess to determine whether the evidence supports or contradicts their hypotheses. Effective explanations will help confront these issues.« (Gunning/Aha 2019: 45f.)

Diese Fallkonkretisierung vermittelt eine erste Idee eines Anwendungsszenarios, sagt jedoch nicht präzise, wann in welcher Form Erklärungen des Systems wie genau für die Analystin hilfreich wären. Der Erklärungsbedarf bleibt undefiniert. Allerdings prägt die Beschreibung die Vorstellung darüber, in welcher Form Erklärungen vorkommen sollen: nämlich *als fehlende Information zu […]*, im Prozess der Entscheidungsfindung der Analystin. Die Aufgabe von XAI besteht somit darin, die notwendigen Informationen verfügbar zu halten und auf verständliche Weise zu

übermitteln. XAI wird hier als Informationsverarbeitungs- und Vermittlungsaufgabe gedacht, die sich kontextunabhängig formulieren lässt.

Das zweite Beispiel stammt aus dem Bereich der militärischen Anwendung von Drohnen in Kampfschauplätzen. Das Militär setzt Drohnen (halb-automatisierte Agenten) ein, um Sachen für Truppen zu transportieren, Informationen einzuholen, Ziele zu identifizieren oder auch abzuschießen. Interessanterweise wird mit Bezug auf diesen zweiten Anwendungsfall von der KI als einem Partner gesprochen, der die Fähigkeiten von Soldat:innen ergänzt:

> »The autonomy challenge was motivated by the need to effectively manage AI partners. For example, the Department of Defense seeks semiautonomous systems to augment warfighter capabilities. Operators will need to understand how these behave so they can determine how and when to best use them in future missions. Effective explanations will better enable such determinations.« (Gunning/Aha 2019: 46)

Was in dieser Umschreibung des zweiten Fallbeispiels ›effectively manage‹ genau bedeuten soll, bleibt offen. Hierbei geht es weniger um fehlende Informationen und deren Vermittlung, sondern darum, das Systemverhalten in pragmatischen Zusammenhängen einschätzen zu können, z.B. die Flugrouten einer teilautomatisierten Drohne.

Anstatt auf die Verschiedenheit beider Anwendungsfälle einzugehen und zu überlegen, welche XAI-Tools für welche Konstellationen und Zwecke geeignet sein könnten, findet sich in den genannten Quellen eine gegenteilige Strategie: Man setzt auf Generalisierung der Anwendungsfälle. Dies passiert, indem weitere Fälle beispielhaft genannt werden – »Consider, for example, a doctor needing to explain a diagnosis to a fellow doctor, a patient, or a medical review board« (Gunning et al. 2021: 8) – und indem ganze Anwendungs*felder* ins Spiel gebracht werden.

Auf diese Weise entsteht der Eindruck einer unspezifischen Inwertsetzung von XAI für Anwendungen von KI, als seien Erklärungen immerzu hilfreich, gewollt oder gar notwendig. Zudem zeigt es, dass die Verschiedenheit der Anwendungskonstellationen in der Konzeption des DARPA-Projektes keine oder nur eine sehr marginale Rolle spielt. Auch die in der Ausschreibung eingeführten Use-Cases spielten in der Durchführung des Projektes keine prägnante Rolle, sondern man kam zu dem Entschluss »that it would be more valuable to explore a variety of approaches across a breadth of domains« (Gunning et al. 2021: 4). Man sah es nicht als zentral für die Forschung an, sich auf bestimmte Beispiele in ihrer Spezifizität festzulegen und einzulassen. Entgegen der proklamierten Hinwendung zu ›den usern‹ denkt man nicht von der Praxis, sondern von der Technologie aus.

So undifferenziert wie der Bezug zu Praxiskontexten ist, so ist er es auch gegenüber den intendierten oder potenziellen Nutzer:innen. Man redet allgemein von den

›Usern‹: »Explainable AI will be essential if users are to understand, appropriately trust, and effectively manage these artificially intelligent partners« (Gunning/Aha 2019: 44). Diese generalisierte Sicht auf intendierte und potenzielle Nutzer:innen von XAI wird indirekt leicht eingeschränkt, indem man den Untersuchungsgegenstand User/Praxis auf solche Anwendungstypen einschränkt, in denen professionelle Entscheidungsträger:innen durch KI-Systeme unterstützt werden:

> »The target of XAI is an end user who depends on decisions, recommendations, or actions produced by an AI system, and therefor needs to understand the rationale for the system's decisions.« (DAPRA 2016: 6)

Es geht nicht um den privaten Bereich, sondern Arbeitswelten. Es geht nicht um Betroffene der KI-assistierten Entscheidungen (Klient:innen, Patient:innen, Datensubjekte), sondern um die Entscheidungsträger:innen. Es geht aber auch nicht um organisationale Strukturen oder institutionelle Einbettungen der Systeme (z.B. die Befehlskette im Militär mit seinen klar geregelten Hierarchien oder die Ebene von Management und Direktion, die in Unternehmen oder Kliniken es zuerst verantworten die fraglichen KI-Systeme überhaupt einzukaufen und anzuwenden). DARPA wählt allein die Mikroperspektive der ›human-computer-interaction‹. Es geht ebenfalls nicht darum, zu vergleichen, ob die Berufstätigen ihre Arbeit mit oder ohne KI-Unterstützung besser ausführen. Die Optimierung der Arbeitsprozesse durch KI-Systeme wird nicht in Frage gestellt. Es geht allein darum mit XAI sicherzustellen, dass die neuen Assistenzfunktionen sinnvoll eingesetzt werden können. Die Handhabbarkeit der Systeme durch ihre Nutzer:innen wird so zum Faktor der Optimierbarkeit von Arbeitsabläufen durch die neue Technologie. Auf diese Weise wird das Projekt in den Rahmen einer abstrakten technischen Fortschrittserzählung gesetzt, was alternative Konzeptionen, etwa die des Ausprobierens, gerade verdeckt. Es geht bei XAI nicht darum, in der Kollaboration verschiedener Interessensvertreter:innen zu testen, in welchen Fällen, welche KI, welche Entscheidungsträger:innen sinnvoll unterstützen kann und inwiefern für wen, wozu und in welcher Art Erklärungen dabei hilfreich oder auch notwendig sein mögen. Die Notwendigkeit der jüngeren KI-Systeme wird schlicht gesetzt und mit ihnen und ihrer Charakterisierung als undurchsichtig die generalisierte Dienlichkeit von XAI. Durch diese undifferenzierte Bezugnahme auf User/Praxis entsteht eine generalisierte Suggestion: Ohne XAI können Entscheidungsträger:innen die neuen, effektiven KI-Systeme nicht sinnvoll nutzen und wären damit vom technischen Fortschritt ausgeschlossen – was es, so der weitere Subtext, zu verhindern gilt.

Dafür, dass die Anwendungsfelder und Fälle in der informatischen Konzeption als hinreichend ähnlich erachtet werden, spricht ebenfalls der Verweis in der DARPA-Ausschreibung auf eine Studie von Kulesza et al. zum »explanatory debugging« (Kulesza et al. 2015), welche dort als Vorbild für die Userorientierte Forschung im

DARPA-Projekt angepriesen wird. Dies ist bemerkenswert, weil die zitierte Studie einen bestimmten Praxisbereich in den Blick nimmt – solche bereits im Einsatz befindlichen Empfehlungssysteme, die Texte klassifizieren müssen, z.B. Spam-Filter für E-Mail-Programme oder die Auswahl von News Feeds (Kulesza et al. 2015: 128). Diese Studie als Vorbild für den militärischen Nutzen von XAI zu nehmen ist deswegen bemerkenswert, weil Anwendungen wie Spam Filter typischerweise anders als militärische Anwendungen nicht als moralisch, politisch oder sozial brisante Kontexte gelten. In der Risikoklassifikation des AI ACTs der EU (European Commission 2022) gelten Spam Filter als das Beispiel für eine Anwendung der Kategorie ›low risks‹. Aus einer Perspektive, die Risikoklassifikationen bzw. die soziale, ethische, politische, legale Sensitivität von Kontexten ernst nimmt, wäre es zumindest fragwürdig, inwiefern das Spam Filter-Beispiel überhaupt als Vorbild für sensitivere Kontexte dienen kann. Eine solche Überlegung scheint in der DARPA-Konzeption keine Rolle zu spielen. Kulesza et al. wählten ihr Beispiel aus technischen und forschungspragmatischen Gründen:

> »We chose text classification because (1) many real-world systems require it (e.g., spam filtering, news recommendation, serving relevant ads, search result ranking, etc.) and (2) it can be evaluated with documents about common topics (e.g., popular sports), allowing a large population of participants for our evaluation.« (Kulesza et al. 2015: 128)

Man könnte überlegen, ob diese Auswahlkriterien, die strategisch sinnvollsten sind, um die effektive Handhabbarkeit von KI-Systemen im Bereich des Militärs oder auch der Medizin, in denen Entscheidungen stark in das Leben von Menschen eingreifen, zu untersuchen. Mir scheint, die undifferenzierte Sicht auf User:innen und Kontexte verfehlt es, überhaupt Fragen der Übertragbarkeit von Studien mit Demonstratoren oder Prototypen oder andere empirisch gewonnene Einsichten präzise adressieren zu können. Die Präsentation der Projektergebnisse als Liste von Stichpunkten (»key takeaways«) verstärkt den Eindruck, als würde man sich entweder für die Übertragbarkeit von Einsichten auf andere Kontexte schlicht nicht interessieren oder als ginge man davon aus, dass diese Einsichten allgemein gültig seien (Gunning et al. 2019; Gunning et al. 2021). Ich demonstriere, wie unplausibel und wenig aussagekräftig die Befunde dadurch werden, an den ersten beiden gelisteten Befunden:

(i) »Users prefer systems that provide decisions with explanations over systems that provide only decisions. (Supported by 11 experiments across performer teams.)« (Gunning et al. 2021: 8)

(ii) »In order for explanations to improve user task performance, the task must be difficult enough that the AI explanation helps (PARC, UT Dallas).« (Gunning et al. 2021: 8)

Es ist fragwürdig, ob (i) allgemein zutrifft. Hoffman et al. widersprechen der Generalisierung: »Explanations are not needed all the time […]« (Hoffman et al. 2023: 241). Befund (ii) ist entweder trivial (wenn nichts erklärungsbedürftig erscheint, braucht es keine Erklärung) oder zu unspezifisch: bezogen auf welche Aufgabentypen, in welchen Kontexten und für welche Berufstätigen trifft dies zu? Was waren die konkreten Bedingungen der entsprechenden Studie (im Labor oder in der Praxis?) und aus welchen Gründen hält man diese Befunde übertragbar und auf was genau? Die Projektkonzeption ließ auf diese Weise wenig Platz für einen echten Paradigmenwechsel, für die man erstens die Problemstellung hätte überdenken müssen, um dann zweitens nach angemessenen Lösungsstrategien zu suchen. Die Organisation der Forschung bot darüber hinaus wenig Raum für die von der informatischen-normalwissenschaftlichen Perspektive abweichende Alternative von Hoffmans Team.

3.1.2 Organisation der Forschung

Das DARPA-Forschungsprojekt sollte drei Forschungsfragen verfolgen: »(1) how to produce more explainable models, (2) how to design explanation interfaces, (3) how to understand the psychological requirements for effective explanations« (Gunning/Aha 2019: 45). Die ersten beiden Fragen waren Aufgabe der elf technischen Forschungsgruppen, aus der ML- und HCI-Community, mit einer Ausrichtung auf Fragen des Interface-Designs.[15] Die dritte Frage lag in der Hand des psychologisch-orientierten Teams von Hoffman. Das Projekt war in zwei Phasen gegliedert. In der ersten Phase (18 Monate) sollten technische Demonstratoren entwickelt werden; in der zweiten Phase (30 Monate) sollten diese getestet werden. Am Ende sollten Prototypen herauskommen, die in einem open source XAI-Toolkit der Öffentlichkeit zur Verfügung gestellt werden.[16] Organisatorisch betrachtet, sah man drei Bereiche vor (Technical Areas (TA)): einen technischen Bereich (TA1), der die Fragen (1) und (2) umfasst, den Bereich der psychologischen Erklärung (TA2), der für die dritte Forschungsfrage zuständig war. Hinzu kam die Evaluation (TA3), welche vom U.S. Naval Reserach Laboratory (ein gemeinsames Forschungslabor der US Navy und des US Marine Corps) unter der Leitung von Eric Vorm verantwortet und organisiert wurde.

Ursprünglich gingen die Beteiligten, laut ihrer veröffentlichten Retrospektive, davon aus, dass die von den elf Teams entwickelten Erklärungstechniken durch

15 Hier HCI im Sinne des kognitiven Paradigmas (Harrison et al. 2007): Es geht um effektive Vermittlung Kanäle, Signale, Informationsdichte.

16 Das Toolkit kann man sich unter diesem Link herunterladen: https://xaitk.org/.

die Gestaltung der Schnittstellen nutzerfreundlich gemacht werden können. Mit den Nutzer:innen wird folglich die Frage der Schnittstellengestaltung zentral. Damit zerfällt die Aufgabe, nutzergerechte Erklärungen zu entwickeln, in zwei technische Komponenten; erstens ein »explainable model« und zweitens das »explainable interface« (Gunning/Aha 2019; Gunning et al. 2021). Diese Arbeitsteilung der Forschung ist von Seiten der Technik gedacht. Eigenschaften des technischen Systems (Opazität, Komplexität) bestimmen den Erklärungsbedarf – Entwicklung von erklärbaren Modellen. Sind diese einmal vorhanden, gilt es sie nutzergerecht zu vermitteln. Beide Schritte stehen additiv zueinander. User sind für die Schnittstellengestaltung relevant, nicht aber für die Frage des Erklärungsbedarfs (siehe kritisch hierzu z.B. Rohlfing et al. 2020). Hinzu kommt, dass die Endnutzer:innen negativ definiert sind, als nicht-Experten. Während die Expert:innen die (neuen) Erklärungstechniken verstehen, braucht es für die Endnutzer:innen die Schnittstelle. Ihr nicht-Expert:innen-Sein ergibt den Vermittlungsbedarf. Damit wird das Zuschneiden auf die ›user‹ auf eine Frage des fehlenden Wissens und der Modalitätsformen verengt, in denen das fehlende Wissen präsentiert werden soll.

Es wird in den Quellen nicht deutlich, ob man sich eine Integration der Ergebnisse des psychologischen Teams in die Entwicklung der neuen XAI-Tools vorstellte. Klar ist, diese Ergebnisse wurden als wichtig für die Evaluation angesehen und damit der Frage, wie man die Effektivität von XAI messen kann. Das psychologische Team sollte hierfür die Grundlage bieten und in diesem Sinn den anderen Projekten assistieren:

> »The program structure anticipated the need for a grounded psychological understanding of explanation. One team was selected to summarize current psychological theories of explanation to assist the XAI developers and the evaluation team.« (Gunning et al. 2021: 3f.)

Die vorhergesehene Hauptaufgabe des psychologischen Teams war »understanding the psychology of explanation by summarizing, extending and applying psychological theories of explanation« (Gunning et al. 2021: 2). Sie sollten die psychologische Fachliteratur zum Erklären überblicken und systematisieren (siehe hierzu den Report: Mueller et al. 2019), und auf dieser Erkenntnisbasis dann ein psychologisch informiertes »computation model« des Erklärens entwickeln, welches sie sodann anhand der Evaluationsergebnisse der XAI-Entwickler:innen validieren sollten (Gunning et al. 2021: 4). In der Retrospektive hat es sich als zu hochgegriffen herausgestellt, ein formalisiertes psychologisches Modell des Erklärens zu erstellen, stattdessen habe das Team um Hoffman beschreibende Modelle erstellt (Gunning et al. 2021: 4). Bemerkenswerterweise lässt sich aus den Publikationen ein Missverständnis darüber erkennen, worin das ›psychologische Modell‹ besteht, welches Hoffmans Team erstellen sollte. Was Gunning et al. (Gunning et al. 2021) als Ergeb-

nis der Arbeit des psychologischen Teams ausgeben (siehe Abbildung 2), markieren Hoffman et al. (Hoffman et al. 2023) als informatische Vorannahme, mit dem das DARPA Projekt gestartet sei (siehe Abbildung 3). Ihr eigentliches psychologisches Modell weisen sie abgrenzend hierzu als Weiterentwicklung zu dieser anfänglichen Annahme aus (siehe Abbildung 4).

3.2 Psychologische Konzeption

Aus der Retrospektive des ›psychologischen‹ Teams um Hoffman lassen sich weitere Diskrepanzen gegenüber der informatischen Perspektive aufzeigen. Zum Beispiel stellen Hoffman et al. zwölf Prinzipien zusammen, welche aus dem DARPA-Projekt bzw. der von diesem stimulierten XAI-Forschung, hervorgegangen sind (Hoffman et al. 2023). Diese zwölf Prinzipien sind weitaus sensibler für die Diversität der Nutzer:innen und Fragen der Relevanz von Erklärungen, als es in der informatischen Perspektive zum Ausdruck kommt. Auch wenn sich die Darstellung der Einsichten aus dem DARPA-Projekt des psychologischen Teams in Teilen so liest, als wären sie neu gewonnen, spricht vieles dafür, dass das psychologische Team die Erforschung von XAI für Nutzer:innen von Beginn an anders konzipiert hätte. Allein der fachliche Hintergrund von Hoffman, der Experte in den Bereichen des ›Cognitive Systems Engineering‹, des ›Human-Centered Computing‹ sowie der ›Human Factors‹-Forschung ist, spricht dafür, dass dem ›psychologischen‹ Team schon vor Projektbeginn klar war, »one explanation does not fit all« (Sokol/Flach 2020). Während es bei Gunning et al. (2021: 8) so klingt, als sei diese Einsicht eine der wesentlichen Lernschritte des Projektes gewesen – »different user types require different types of explanation« –, plädieren Clancey und Hoffman (2021) z. B. dafür, Einsichten wie diese aus der Forschung zu *Intelligent Tutoring Systems* für die XAI-Forschung fruchtbar zu machen. Aus letzterem Bereich seien Erkenntnisse aus mehreren Jahrzehnten Forschung zu gewinnen.

Während die informatische Perspektive von Seiten der Technik gedacht ist, denkt die psychologische HCI-Perspektive von ›den usern‹ aus.[17] Letzteres impliziert wenigstens drei zentrale Forschungsfragen, die in der informatischen Perspektive nicht zur Geltung kommen:

17 Mit der fachlichen Expertise von Hoffman scheint sein Team in der Forschung von den ›users‹ oder ›humans‹ ausgegangen zu sein und nicht von typischen psychologischen Ausgangspunkten wie kognitiven Prozessen oder Persönlichkeitsmerkmalen. Ich danke Ingrid Scharlau für den Hinweis, dass gegenüber solchen psychologischen Ausgangspunkten die Rede von ›dem User‹ einen ziemlich groben Forschungsgegenstand konstruiert, den man in der klassischen Psychologie so nicht findet.

1. Wann sind Erklärungen für Nutzer:innen überhaupt relevant?
2. Wozu dienen Nutzer:innen Erklärungen?
3. Wie geht man mit der Diversität von Kontexten und Nutzer:innen um?

Die informatische Sicht leitet den Erklärungsbedarf aus den Eigenschaften des technischen Systems ab (Komplexität, Opazität) während dieser Perspektivwechsel dazu einlädt, die Bedarfe von Nutzer:innen in verschiedenen Praxiskontexten zu erkunden. Wer so fragt, braucht ein anderes Wissen über Nutzer:innen und Kontexte. Es reicht dann nicht aus User Studies allein für die Evaluation der gewonnenen Werkzeuge einzubeziehen, denn dieses Wissens sollte das Forschungsdesgin von Anfang an informieren:

> »The design of XAI systems must be fully informed by a psychological model based on empirical evidence of what happens when people try to explain complex systems to other people and what happens as people try to reason out how a complex system works […].« (Hoffman et al. 2022: 366)

Indem die DARPA-Konzeption User studies allein für die Evaluation angesetzt hat, weist sie dem Wissen über Nutzer:innen und Kontexten sowie dem Wissen *von* User:innen einen bestimmten Platz zu: es betrifft allein die Effektivität der entwickelten XAI-Tools (Gunning et al. 2021: 4).[18] In der Entwicklung dieser Tools spielt dieses Wissen keine Rolle. Auch diese Platzierung des User-Wissens spricht für die Auffassung, dass die Erklärungen unabhängig von den konkreten, kontextuellen Bedarfen identifizierbar sind. Sie ergeben sich aus Mängeln bzw. Eigenschaften der KI-Systeme allein. In dieser Auffassung kommt das Selbstverständnis des Paradigmas der epistemischen Güte von ML deutlich zum Tragen, in denen es allein darum geht, die KI-Modelle für andere Expert:innen einsichtig zu machen und aufgrund der Homogenität dieser Expertengruppe die Verschiedenheit anderer Nutzer:innen nicht als relevante Kategorie erachtet wird.

Die informatische Perspektive verpasst es so, überhaupt nach der Bedeutung von Erklärung aus Sicht von User:innen in verschiedenen Kontexten zu fragen. Sie entwickelt keine Vorstellung über die Kontextabhängigkeit von Erklärprozessen. Zum Beispiel scheint allein die Tatsache, dass Nutzer:innen in ihren Arbeitswelten häufig mit multiplen Aufgaben und Zielen konfrontiert sind, in denen sich die KI-Systeme mit ihren Empfehlungen (besser oder schlechter) einfügen, einen praktischen Unterschied zu machen, wie Erklärungen effektiv sein können. In der XAI-Forschung geht man stattdessen von einer Vereinfachung aus: »Much XAI research

18 User studies waren ein zentraler Bestandteil der Evaluationsphase, mit insgesamt 12.700 Teilnehmenden, wovon »1900 supervised« und »10 800 unsupervised participants« waren, die z.B. über Amazons Mechanical Turk gewonnen wurden (Gunning et al. 2021: 7).

has assumed a one-person, one-task problem situation« (Hoffman et al. 2023: 243). Solche Vereinfachungen können forschungspragmatisch sinnvoll sein, man sollte sich jedoch über die Differenz zu den eigentlich angedachten Nutzungskontexten im Klaren sein. Um über diese gesättigteren Vorstellungen zu gewinnen, gibt es verschiedene Strategien, wie den Einbezug von Domänenexpert:innen oder anderen Stakeholdern im Sinne eines partizipativen Forschungsdesigns (Simon 2017; Dignum 2019; van der Hoven/Manders-Huits 2020; Friedman/Nissenbaum 1996) oder einer Befragung und Beobachtung von diesen in realen Anwendungskontexten oder im Labor. Zwar scheint es jüngst ein größeres Bewusstsein in der XAI-Forschung für die Notwendigkeit dieser Art von Forschung zu geben (Ribera/Lapedriza García 2019; Langer et al. 2021; Cabitza et al. 2023; Capel/Brereton 2023; Kim et al. 2024) – wenn man mit XAI den Zweck verfolgen will, für User:innen in verschiedenen Kontexten dienliche Erklärungen anzubieten – doch nach wie vor scheint es hierzu wenig empirische Untersuchungen zu geben (so der Befund von Langer et al. 2021). Wang et al. betonen, dass sich mit diesem Perspektivwechsel weitere Forschungsfragen anschließen – zum Beispiel wie sich die Relevanz von Erklärungen für User:innen überhaupt validieren lässt und wie man mit dem Unterschied zwischen Normen guter Erklärungen (aus der Argumentationstheorie und Logik) und dem, wie Personen tatsächlich anderen Dinge erklären, umgehen sollte (Wang et al. 2019: 601). Sollte sich XAI eher an der Norm einer guten Erklärung orientieren oder eher am Vorbild realer Erklärungen?

Die Kernfrage, die sich mit der ernsthaften Hinwendung zu User:innen und der Praxis stellen muss, ist die nach der Relevanz von Erklärungen.[19] Denkt man von der Technik aus, besteht der Bedarf an Erklärung scheinbar allgemein betrachtet, nämlich so lange wie die Technik opak und komplex ist und diese Charakteristika der Technik als erklärungsbedürftig angesehen werden. Nur wenn man diese Zwecksetzung von XAI, die von Eigenschaften der Technik ausgeht, unreflektiert auf eine andere Zwecksetzung von XAI überträgt – nämlich KI-Systeme in der Praxis effektiv nutzen zu können – entsteht der Eindruck einer generalisierten Inwertsetzung von XAI, die von Seiten der User:in und der Praxis gedacht nicht haltbar ist. Es ist sonach geboten, beide Zwecksetzungen voneinander zu unterscheiden: wir haben es hier mit verschiedenen Paradigmen von XAI zu tun.

Um die Frage nach der Relevanz von Erklärungen zu erforschen, schlägt das Team um Hoffman vor, von ›Triggern‹ zu sprechen, die an verschiedenen Stellen im Gebrauch von KI-Systemen auftauchen können. Typische Trigger sind Überra-

19 Was die Frage angeht, ob es überhaupt eine Erklärung in einer bestimmten Situation braucht, könnte man weiter pragmatisch-kontextuelle Faktoren (Relevanz) von individuell-kognitiven Faktoren (Angemessenheit, Sinnhaftigkeit) unterscheiden. Für diesen Hinweis danke ich Katharina Rohlfing.

schungen oder der Bruch mit Erwartungen (Hoffman et al. 2023). Zu diesem Schluss kommen ebenfalls de Graaf und Malle:

> »In short, people try to explain any given behavior (in self or other) if either (a) they themselves wonder why the behavior occurred or (b) they expect that someone else wonders why the behavior occurred.« (de Graaf/Malle 2017: 22)

Dieser psychologische Befund lässt sich gut an einen Kerngedanken der phänomenologischen Technikphilosophie anschließen. Erklärungsbedürftig ist das nichtselbstverständliche; das, was mit Vertrautem bricht. Wichtig ist hier, dass unser eingeübter Umgang mit Technik ebenso zum Bereich des Selbstverständlichen gehört wie alles andere auch. Technik ist in diesem Sinne Teil der Lebenswelt (als Inbegriff des Selbstverständlichen) und damit aus phänomenologischer Perspektive gerade das, was wir nicht (bewusst) zu verstehen brauchen, um sie nutzen zu können (Blumenberg 1981; Kaminski 2010). So wie wir keine Fachphysiker:innen sein brauchen, um eine Straßenbahn sinnvoll zu nutzen, liegt es ebenso nicht auf der Hand, wie und warum z.B. eine Einsicht, ähnlich der von Heatmaps, für die Bilderkennung notwendig sein sollte, um das Verhalten von Drohnen einschätzen zu können. Dem modernen Großstädter genügt es »daß er auf das Verhalten des Straßenbahnwagens ›rechnen‹ kann, er orientiert sein Verhalten daran; aber wie man eine Trambahn so herstellt, daß sie sich bewegt, davon weiß er nichts« (Weber 1992: 86). Max Webers Beispiel der Straßenbahn gibt zu bedenken, dass das nötige Verständnis über das Verhalten von Technik von anderer Art ist als das entsprechende Expertenwissen zum Bau dieser Technik.

Vor dem Hintergrund solcher Überlegungen ließen sich weitere Forschungsfragen für die XAI-Community gewinnen. Ein Fragebündel geht in die Richtung zu konzipieren, worin das Unerwartete eigentlich besteht: Sind es Abweichungen von sozialen Normen? Das Unvertraute? Das Neue? Wie zeigt sich dieses jeweils in verschiedenen konkreten Arbeitskontexten? Hieran angeschlossen ließe sich ein weiteres Fragebündel in die Richtung ausbuchstabieren, wofür genau XAI eingesetzt werden soll. Es könnte beispielsweise einen Unterschied machen, ob es darum geht Nutzer:innen bei der Aneignung von Neuem (neuen technischen Hilfsmitteln, neuen Aufgaben, etc.) zur Seite zu stehen (quasi als eine Art bessere Bedienungsanleitung) oder ob es darum geht Fälle, die von den Üblichkeiten abweichen oder schwerer einzuschätzen sind mit Hilfe von KI und ggf. Erklärungen zu dieser besser nachvollziehen zu können (Lebovitz et al. 2022). Weiter macht es einen Unterschied ob XAI Angebote machen soll, um etwas besser zu verstehen, oder ob es letztlich (z.B. im medizinischen oder juristischen Bereich) darum geht, Entscheidungen rechtfertigen zu können (Krishnan 2020).

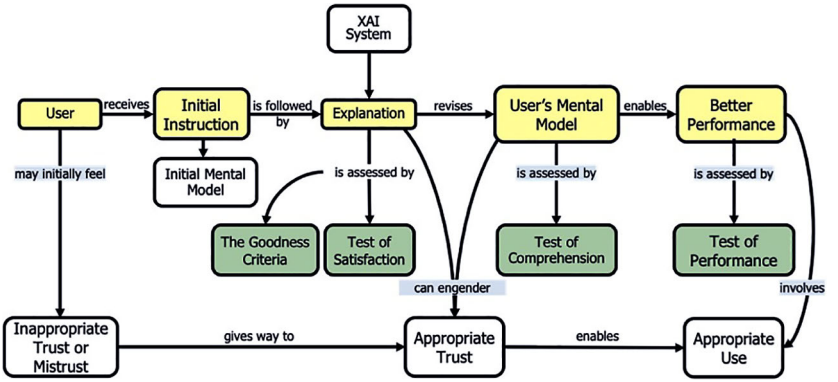
Die Frage nach der Relevanz von Erklärungen aus User-Sicht erfordert somit eine differenzierte Sicht auf die instrumentellen Zwecksetzungen von XAI und da-

mit ebenfalls auf die verschiedenen Kontexte und Nutzer:innen. Die praktische Notwendigkeit dieser Differenzierungen ergibt sich nicht in der Perspektive, die von der Technik aus den Bedarf an Erklärungen ableitet.

Diese verschiedenen Sichtweisen auf XAI implizieren außerdem verschiedene Vorstellungen darüber, wie die Technik und die Nutzer:innen zusammenspielen sollen. Diese Vorstellungen zum Verhältnis von ›user‹ und XAI hängen wiederum mit der Zwecksetzung von XAI zusammen, d.h. mit den Zielvisionen wie eine wünschenswerte Zusammenarbeit von XAI und Nutzer:innen aussehen würde. Diese Fragen hängen mit der Aufgabenstellung im DARPA-Projekt zusammen und fordern das psychologische Team, ein psychologisches Modell des Erklärens zu entwickeln. Die Darstellung der Ergebnisse dieser Modellerstellung demonstriert deutlich den Unterschied zwischen der informatischen und der psychologischen Perspektive im DARPA Projekt, denn das, was Gunning et al. (2021) als Ergebnis der Arbeit von Hoffmans Team herausstellen und als psychologisches Modell des Erklärens titulieren (Abbildung 2) weisen Hoffman et al. (2023) dezidiert als eben nicht-psychologisches, sondern informatisches Modell des Erklärens zurück – mit dem das DARPA-Projekt gestartet ist: Dieses Modell ist kein Ergebnis der Forschungsarbeit, sondern eine Explikation von Vorannahmen. Hoffman et al. üben harsche Kritik an diesem Modell, weil es den Prozess des Erklärens in der Logik einer Fütterung (›spoon feeding‹) modelliere (Abbildung 3): »The explanation is generated and then delivered, (ideally) to good effect.« (Hoffman et al. 2023: 239). Dieses Fütter-Modell impliziert eine Reihe von Unterstellungen. User werden hier als passive Empfänger von Erklärungen vorgestellt, die keinen aktiven Beitrag zum Erklärprozess liefern (siehe kritisch dazu Rohlfing et al. 2021): »the ›spoon feeding‹ paradigm is blind to the fact that users engage in a motivated, deliberative attempt to make sense of the AI system and any explanatory material that may be presented.« (Hoffman et al. 2023: 239)

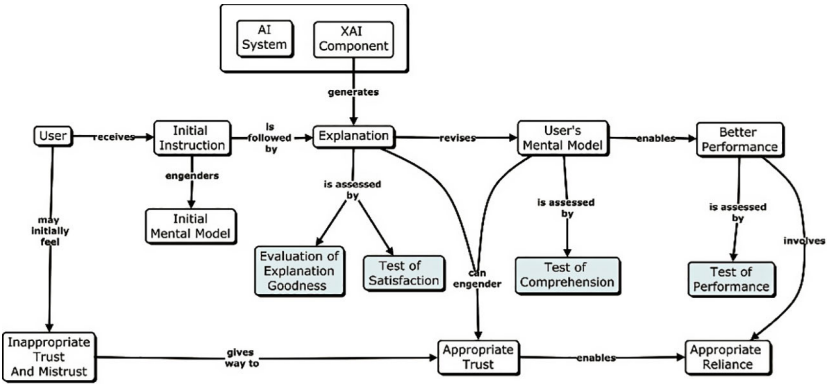
Die Logik des informatischen Modells basiert auf der Annahme einer simplen Informationsübertragung von einem Sender (XAI-System) zu einem Empfänger (user) durch bestimmte Kanäle (Schnittstellengestaltung). In dieser Vorstellung liegt der aktive Part beim XAI-System. Ob die Übertragung gelingt, dafür ist die Güte der Kanäle ausschlaggebend. Indem Hoffman et al. diese Vorstellung als »spoon feeding paradigm« ausweisen (Hoffman et al. 2023: 239), heben sie die unterstellte Passivität der User:in drastisch hervor, denn bei einer typischen Fütterung (von Haustieren, Babys oder pflegebedürftigen Personen) gehört es allein zur Rolle des Fütterungssubjekt den Mund zur passenden Zeit zu öffnen und wieder zu schließen. Der aktive Beitrag zur Verdauung des Futters liegt allein im Kauen und Schlucken – der eigentliche Verdauungsvorgang läuft dank funktionierender Organe wie automatisch ab. Die ganze Musik beim Füttern liegt auf der Seite der Fütternden. Die Gefütterten haben in der Regel wenig mitzubestimmen über das, womit sie gefüttert werden (Ist es gut für mich, brauche ich das?).

Abbildung 2: Darstellung der vermeintlichen Ergebnisse des psychologischen Teams nach Gunning et al.



Quelle: Gunning et al. 2021: 5

Abbildung 3: Rückweisung der informatischen Vorannahmen zum Verhältnis XAI-user (›spoon feeding model‹)



Quelle: Hoffman et al. 2023

Sie nehmen nur indirekt darauf Einfluss, wie viel sie bekommen. Versteht man XAI als Fütter-Aufgabe, treffen also die Fütternden (Systemdesigner:innen) allein alle Relevanzentscheidungen, die User:innen hingegen werden infantilisiert.[20]

20 Die Praxis des Fütterns lässt sich sehr wohl im familiären Bereich oder in der Pflege als eine Situation sozialer Aushandlungen verstehen, in der die Gefütterten keineswegs rein passiv sein müssen. Mir scheint aber die Wahl der Metapher von Hoffman darauf abheben zu wollen, dass User:innen hier als rein passive Gefäße modelliert sind. Diese Wahl spielt vermut-

Die Vorstellungen zum Verhältnis von User:in und XAI lassen sich darauf beziehen, welche Zielvisionen für die XAI-Entwicklung eigentlich angesetzt wird. Wie sehe eine optimale XAI in der Nutzung aus? Aus der informatischen Konzeption des DARPA-Projektes lassen sich hierzu nur indirekt Vermutungen aufstellen. Da diese Konzeption von der Technik (allein) her denkt, wäre die optimale XAI eine perfektionierte Technik. Worin die Perfektion der Technik liegen könnte, dafür scheinen mir unterschwellig zwei Ideale eine Rolle zu spielen: Das Ideal der Selbstevidenz und das Ideal der vollständigen Automatisierung.

Beim ersten Ideal der Selbstevidenz ist die KI nicht (mehr) fragwürdig bzw. sind die angebotenen Erklärungen so klar, dass keine Fragen übrigbleiben und man eine vergleichbare Situation hat wie bei einer vollkommen transparenten KI. Die Problemstellung (opake und komplexe technische Systeme) löst sich in dieser Vision quasi auf, ergo gibt es keinen Bedarf mehr an XAI bzw. hat XAI ihre genuine Aufgabe optimal erfüllt.[21] Die zweite Vision ist das Spiegelbild der Selbstevidenz für XAI. Eine vollkommen automatisierte XAI ist nicht angewiesen auf das Zusammenspiel mit ›usern‹ oder Kontexten, sie schafft es allein aus sich heraus flexibel jeden Erklärungsbedarf zu befriedigen. In beiden Zielvorstellungen spielen Nutzer:innen und die Kollaboration dieser mit der XAI keine tragende Rolle. Sie passen zu der Ignoranz gegenüber der Diversität von ›usern‹ und Anwendungsfeldern. In dieser Sicht erfüllt die perfekte Technik die Funktion, menschliche Arbeit zu substituieren.

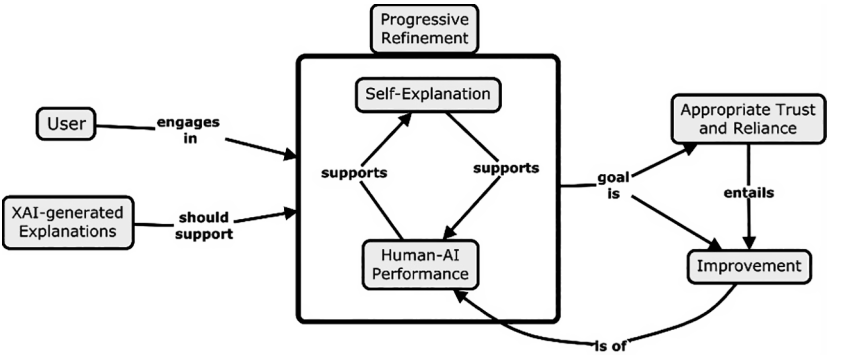
Das Team um Hoffman bietet in diesem Punkt ebenfalls eine alternative Sicht an: bei ihnen geht es nicht um Substitution, sondern Augmentation. Nicht die vollständige Automatisierung (oder die vollständige Auflösung des Erklärungsbedarfs im Sinne vollkommener Transparenz) ist das Ziel, sondern Kollaborationen zwischen Nutzer:innen und KI-Systemen. Dies betonen sie mit ihrem Prinzip »All Explanations involve Self-Explanation«. Es solle bei XAI nicht darum gehen, durch die automatisierten Erklärungen Mängel der Technologie zu beheben (ihre Opazität, ihre Komplexität), sondern darum Nutzer:innen in die Lage zu versetzen, sich das Fragwürdig-Gewordene mit Hilfe der XAI *selbst zu erklären*. Dementsprechend sollte eine oberste Design-Maxime wie folgt lauten: »Explain unto others in such a way as to help them explain to themselves« (Hoffman et al. 2023: 238). Die Ausgaben des XAI-Systems werden diesem Prinzip nach als Material und Hilfestellung für den Erklärungsprozess gesehen, der kollaborativ zwischen den Nutzer:innen und dem System stattfindet und in dem die User:innen einen aktiven Part spielen: Sie müssen in die Lage versetzt werden, sich das Verhalten des Systems mit Bezug auf das,

lich auf ein oft kritisiertes Modell des Lernens an, in dem Lernen ebenfalls mit rein passiver Nahrungsaufnahme verglichen wird.

21 Dieses verheißt jedenfalls das Versprechen der Transparenz. Dass diese allein praktisch nicht automatisch sinnvolle Anhaltspunkte liefert, wurde häufig diskutiert (vgl. Vogelmann 2019; Stamboliev 2023).

was für sie in einem spezifischen Kontext relevant ist, selbst erklären zu können. Der Output von XAI-Systemen (»representations«) darf also nicht mit der Erklärung selbst verwechselt werden, sondern assistiert den Erklärungsprozess (Hoffman et al. 2023). Das XAI-System erhält hierbei die Rolle eines Assistenten, denn letztlich geht es um das Vermögen der Nutzer:innen. Für diese Zielvision hat das Team ihr psychologisches Modell von XAI als Kollaboration erstellt (Abbildung 4).

Abbildung 4: Selbst-Darstellung der Ergebnisse des ›psychologischen‹ Teams: XAI als Kollaboration



Quelle: Hoffman et al. 2023

Bereits 2013 hatte sich Hoffman in Ko-Autorschaft in »The Seven Deadly Myths of ›Autonomous Systems‹« kritisch gegenüber eindimensionalen und überzogenen Konzeptionen der Autonomisierung von Technik geäußert (Bradshaw et al. 2013). Der springende Punkt hierbei ist, dass sie Autonomie nicht als Eigenschaft technischer Systeme missverstanden wissen wollen, sondern als Attribut des soziotechnischen Systems, d.h. dem Zusammenspiel von Mensch und Technik in spezifischen Kontexten mit ihren Bedingungen. Insofern Autonomie als Zielstellung für die technische Entwicklung angesehen wird, sollte man diese Zielstellung nicht einseitig auf die Eigenschaften des technischen Systems beziehen. Vielmehr sei das eigentliche Ziel effektive Teams von Menschen und Maschinen/Software in bestimmten Kontexten zu schaffen; es geht also darum die Autonomie dieses soziotechnischen Systems zu steigern. Der Kontextbezug ist dabei nicht außer Acht zu lassen. Autonomie ist auf Handlungssituationen bezogen: »Functions can't be automated effectively in isolation from an understanding of the task, the goals, and the context« (Bradshaw et al. 2013: 56). Im Design technischer Systeme sind, entweder explizit oder implizit, notwendigerweise Annahmen über die Kontexte der Anwendungen, ihre Nützlichkeit und beschränkte Bedingungen enthalten – Funktionen lassen sich

gar nicht anders konstruieren und denken (Bradshaw et al. 2013: 57; Vermaas 2010; Lenk 1982). Diesen Überlegungen folgend lässt sich vermuten, die informatische Konzeption des DARPA-Projektes, die offenkundig nur oberflächlich über Kontexte und User:innen nachgedacht hat, hat stillschweigend den Kontext der informatischen Forschung und den User-Typ des ML-Experten (eben in einer defizitären Variante charakterisiert durch mangelndes Wissen) auf das Forschungsthema XAI für Laien übertragen. Doch so wenig wie man Autonomie als eine diskrete, separierbare Komponente eines Systems missverstehen sollte, so sollte man auch Erklärbarkeit des Systemgeschehens konzipieren als »capability of the larger system enabled by the integration of human and machine abilities« (Bradshaw et al. 2013: 56). Als Vermögen (»capability«) ist die Verwirklichung von Autonomie bzw. Erklärbarkeit dann auf kontextuelle Bedingungen angewiesen. Ihre Relevanz und Angemessenheit, ergibt sich über die kontextuellen Zwecksetzungen des Handelns. Den Gedanken, nicht technische Eigenschaften, sondern das Team Mensch-Maschine, in den Fokus zu nehmen, stellen auch Miller et al. an den Anfang ihrer Einführung zum *Special Issue on Explainable AI*, welches von den Forschenden des DARPA-Projektes initiiert wurde: »AI can be seen as one integrative part of a cognitive work system and its broader social or organizational context.« (Miller et al. 2022)

3.3 Hinwendung zur Praxis als Paradigmenwechsel

Der Einbezug von Nutzer:innen und Kontexten ist wichtig und kann methodisch und theoretisch vielfältig angegangen werden. Die Arbeit des Teams von Hoffman ist nur ein Beispiel hierfür, das deswegen interessant ist, weil es im Kontext des DARPA-Projektes konzipiert wurde. Allein schon die hier demonstrierten Unterschiede zwischen den anfänglichen informatischen Vorstellungen von der Forschung an der ›psychology of explanation‹ und dem, was Hoffmans Team veröffentlicht hat, sprechen dafür, dass wir es bei der Hinwendung zu Nutzer:innen und zu Kontexten mit einem Paradigmenwechsel gegenüber der normalwissenschaftlichen Forschung der ML-Community im Paradigma der epistemischen Güte von ML zu tun haben. Ich plädiere dafür, diesen Paradigmenwechsel ernst zu nehmen. Wir haben es mit einer anderen Problemkonzeption zu tun als derjenigen, die das Paradigma der epistemischen Güte von ML sinnvoll anleitet.

Die Problemkonzeptionen sollten sich aus den Zwecksetzungen von XAI herleiten, über die sich die Community ernsthafter Gedanken machen sollte (Krishnan 2020; Freiesleben/König 2023). Mir geht es nicht darum eine ›human-centered‹ (Capel/Brereton 2023) gegenüber einer ›technology-centered‹ XAI auszuspielen, sondern darum, die Verschiedenheit dieser, mit dem jeweiligen Fokus auf sinnvoll einhergehende Zwecksetzungen zu beachten (Gunning et al. 2019). Will man sich ernsthaft mit der Güte von XAI für Laien in realen Anwendungsfällen beschäftigen, sollte man die Forschung entsprechend anders organisieren als es im Paradigma der epis-

temischen Güte von ML der Fall ist. Dies fängt bei der Gewichtung der beteiligten Disziplinen an, geht über die Erstellung der leitenden Forschungsfragen und Reflexion der leitenden Vorannahmen, bis hin zur Frage des Einbezugs von Nutzer:innen und der Kontextabhängigkeit von Erklärungen hinaus.

4. Das Paradigma der Vereinbarkeit mit Grundwerten und Normen

Mein drittes Paradigma, der Vereinbarkeit von KI mit Grundwerten und Normen, stellt erneut einen Perspektivwechsel dar. Es geht um ›ethical and social concerns‹, die insbesondere als Auseinandersetzung um ›bias‹, ›fairness‹ und ›accountability‹ von KI diskutiert werden. Genährt aus Sorgen, epistemischen und normativen Unsicherheiten im Umgang mit KI sowie dem vorauseilenden Gehorsam der Tech-Industrie (Lepri et al. 2018; Tworek 2019; Floridi 2021), welche beschwört, *AI for the social good* zu entwickeln, ist ein genuiner Diskursraum entstanden, der unter dem Namen *AI Ethics* firmiert. Dieser Diskurs konstituiert sich über das abstrakte gemeinsame Interesse, KI-Systeme gesellschaftsfähig zu machen und folgt größtenteils der Leitidee, *AI Ethics* als Selbstregulierung einzusetzen (Alpsancar 2023). Für Unternehmen geht es um ihr Selbstverständnis und Ansehen, Akzeptanzfragen und die Konformität mit rechtlichen Vorgaben. Die Politik will KI als Schlüsseltechnologie für wirtschaftlichen Erfolg ihrer Nationen fördern und sucht nach einem passenden regulativen Rahmen (Nannini et al. 2023). *AI Ethics* ist ein anwendungsorientiertes Forschungsgebiet vielfältiger Expertisen, welches ebenso in die Zuständigkeit der Computer Science und ihr angrenzende Disziplinen fällt, wie in die Produktentwicklung großer Tech-Konzerne. Zu diesem Diskursraum zählt ebenfalls eine in der breiteren Öffentlichkeit geführte Debatte um ›high profilic cases‹, über die in Blogs, Tech-Magazinen oder auch Buchpublikationen diskutiert wird (O'Neil 2016; Eubanks 2017; Benjamin 2019; Crawford 2021). Parallel zu der breiteren öffentlichen Debatte entstanden über die letzten Jahre zahlreiche ethische und politische Initiativen, die das Thema der Vereinbarkeit von KI mit gesellschaftlichen Grundwerten auf ihre Agenda setzten (Hagendorff 2020; Nannini et al. 2023). Im Jahr 2019 waren bereits mehr als 80 ethische Richtlinien oder KI-Kodizes öffentlich zugänglich (Jobin et al. 2019; Morley et al. 2020), die von Industrieverbänden wie der IEEE oder ACM, von Unternehmen wie IBM oder Google, oder von staatlichen Institutionen initiiert wurden, wie etwa die High Level Expert Group (2019) der Europäischen Kommission.

Dieser Diskurs adressiert zum einen bestimmte Einsatzgebiete, nämlich »socially significant and morally weighty contexts« (Walmsley 2021: 585; Floridi et al. 2018), in denen die KI-gestützten Entscheidungen einen erheblichen Einfluss auf das Leben derjenigen haben, die von diesen Entscheidungen betroffen sind, z.B. dem Finanzwesen (Zarsky 2016; Pfeiffer et al. 2023), dem Personalwesen (Starke

et al. 2022; Strohmeier 2020), dem Justizwesen (Corbett-Davies et al. 2017; Fortes 2020), dem Gesundheitswesen (Marabelli et al. 2018; López-Martínez et al. 2020) oder bei anderen staatlichen Hoheitsaufgaben (Allhutter et al. 2020). Zum anderen stehen Produkte und Services bekannter Tech-Unternehmen im Zentrum der Aufmerksamkeit, deren Einsatz diskriminierende Effekte hervorbrachte. Klassifikationsfehlern von Bilderkennungsprogrammen, die im geschützten Rahmen der laborähnlichen Wettbewerbe der ML-Community primär epistemisch von Bedeutung sind, kommen hier gesellschaftliches Gewicht zu, wenn farbige Personen durch die Google Photo App als ›gorilla‹ kategorisiert werden (Kasperkevic 2015), das Konzentrationslager in Dachau auf dem von Yahoo bereitgestelltem Foto-Dienst Flickr als ›gym‹ ausgewiesen wird (Hern 2018) oder in Facebooks Daily Mail Video Nutzer:innen, die Videos mit schwarzen Personen angeschaut hatten, ›weitere‹ Videos über Primaten empfohlen bekommen (AIAAIC 2021). Durch einen solchen »steady stream of examples and anecdotes of problematic biases, prejudices and other errors that have been automated and reinforced« (Walmsley 2021: 585), sei insgesamt eine große Skepsis gegenüber KI entstanden (Miller 2019: 1; Gilpin et al. 2018).

AI Ethics funktioniert größtenteils als Resonanzraum dieser Sorgen und Debatten. *Ethics* wird dabei hauptsächlich in attributiver Bedeutung verwendet: *ethische KI* ist gute KI. Synonym zu ethischer KI spricht man auch von einer KI, die ›sozialen Werten‹ gerecht wird. Die gesellschaftliche Verträglichkeit von KI ist für viele eine Aufgabe des *Engineerings*. Beispielhaft hierfür steht das Buch von Kearns und Roth, *The Ethical Algorithm* (Kearns/Roth 2019). Die beiden Informatiker verbinden ihre universitäre Forschung zum »socially aware algorithm design« mit einer Beratungstätigkeit für große Tech-Unternehmen wie Amazon, Apple und Facebook. Eine der größten Konferenzen dieser wachsenden Community ist die *ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)*.[22] Die Tech-Branche sponsert in diesem Themengebiet nicht nur einzelne Forscher (Roth und Kearns sind z.B. Amazon Scholars) und wissenschaftliche Tagungen wie die FAccT, sondern stellt auch sogenannte *Ethicists* ein, die in Unternehmen die gesellschaftliche Verträglichkeit der Produkte im Entwicklungsprozess sicherstellen sollen (Metcalf et al. 2019).

Der Diskursraum und seine zugehörigen Praktiken des Engineerings, des Ausrichtens, Sorgens und Debattierens trägt durchaus Züge von dem, was Petra Gehring für den Bereich der Bioethik als Realexperiment der Rechtspolitik

22 Die American Computing Machinery ist der größte nordamerikanische Fachverband der Computer Science und hat 2019 die Schirmherrschaft über die Konferenzreihe übernommen, die 2018 zunächst als FAT* Conference gestartet war. Fairness, Accountability und Transparency waren schon zuvor auf mehreren Workshops ein Thema. Auf größeren Konferenzen der Computer Science unterstreicht aber die Zusammenlegung zu einer eigenen Konferenz die gewachsene Bedeutung des Themengebietes.

beschrieben hat. Wie im Fall der Bioethik hallt der Ruf nach *AI Ethics* nicht nur durch die Massenmedien, Bildungseinrichtungen, Forschungsbetriebe und ganze Industrien, sondern auch in den »Vorhöfen und Foren legislativ tätiger oder Gesetzgebung zumindest erwägender Politik« (Gehring 2016: 144). Ebenso wie im Fall der Bioethik zeigt sich AI Ethics praktisch als ein »Mittelding aus Sachverständigeneinlassung, Meinung und Entscheidungsvorschlag, Beiträge zu Genehmigungsverfahren zu leisten, zur freiwilligen Selbstverwaltung von Körperschaften, Verbänden, Forschungseinrichtungen, zur parlamentarischen Arbeit und zum Regierungshandeln« (Gehring 2016: 146). Man kann AI Ethics als Mittel sehen, härtere Regulierungen zu umgehen (Floridi 2021), als Zwischenlösung auf dem Weg zu Regulierungen (Robles Carillo 2020) oder auch als Konkretisierungshilfe für die Rechtssprechung (Surden 2020). Der Rechtspolitik mag KI-Ethik ähnlich wie die Bioethik als Testfeld dienen, auf dem Normierungen abgewägt und deren Akzeptanz von Seiten verschiedener gesellschaftlicher Akteur:innen vorgeführt werden kann (Gehring 2016). Das Verhältnis zur Regulierung freilich ist komplex (Hilgendorf 2020), es sollte jedoch deutlich sein, dass man sich mit dem Paradigma des ›value alignment‹ auf einer politischen Spielwiese tummelt.

Der Diskurs der *AI Ethics* ist freilich breiter als XAI. Jedoch wird letzteres als ein entscheidendes Mittel angesehen, um die Gesellschaftsverträglichkeit von KI zu gewährleisten. Im Umkehrschluss erhält die XAI-Forschung eine ethische und soziale Zwecksetzung: XAI dient der Sicherung des ›value alignment‹ von KI. Diese dritte instrumentale Zwecksetzung beginnt interessanterweise bei der gleichen Problemkonzeption, wie es für die technisch-methodische XAI-Forschung typisch ist. Als Kern des Problems wird die schwärze von Kisten angesehen:

> »Such technologies are frequently described as a ›black box‹, capable of producing powerful results, but with little ability on the part of their creators to understand exactly how and why they make the decisions they do.« (Hern 2018)

Entsprechend schreibt man die ungewollten Nebeneffekte des Einsatzes von KI, die rassistischen und sexistischen Effekte, pauschal ›den Algorithmen‹ zu (Kasperkevic 2015). Die argumentative Verkettung verläuft hier wie folgt: die neuen ML-Systeme sind mächtig, aber opak. Um Fragen nach Fairness, Bias und Verantwortbarkeit adressieren zu können, brauche es mehr Erklärbarkeit und Transparenz dieser Systeme: »These explanations are important to ensure algorithmic fairness, identify potential bias/problems in the training data, and to ensure that the algorithms perform as expected« (Gilpin et al. 2018). Eine Einschränkung erhält diese allgemeine Valorisierung von XAI durch den Blick auf bestimmte Anwendungsbereiche, nämlich solche, in denen das menschliche Unvermögen die Maschine zu verstehen, zum gesellschaftlichen Problem wird, insbesondere dann, wenn Anwender:innen von KI-Systemen das Verhalten der Maschine gegenüber anderen rechtferti-

gen müssen. Konkreter wird die ethische-soziale Dienlichkeits-Zuschreibung selten. Ebenso vage bleibt die Zuschreibung dahingehend, unter welchen Bedingungen XAI überhaupt wie, für wen und für was hilfreich sein kann:

> »AI applications can have great societal impact, improving our societies and building a better world. Explainable AI can facilitate our greater adoption of AI applications by empowering us to address important issues like fairness, bias, verifiability, safety, and accountability.« (Kamath/Liu 2021: 10f.)

In den ethischen Richtlinien zu KI finden sich ähnliche pauschale Zuschreibungen dieser Dienlichkeit. Exemplarisch sei hier aus der *Recommendation on the Ethics of Artificial Intelligence* der UNESCO zitiert, die im November 2021 veröffentlicht und von allen 193 Mitgliedstaaten angenommenen wurde:

> »Transparency and explainability relate closely to adequate responsibility and accountability measures, as well as to the trustworthiness of AI systems.« (UNESCO 2021: 22)

Wie mehrere Meta-Reviews dieser ethischen KI-Richtlinien gezeigt haben, wird *explainability* (oder ein damit verwandtes Konzept wie *transparency* oder *interpretability*) als zentral für die KI-Entwicklung gesetzt (Hagendorff 2020; Morley et al. 2020; Jobin et al. 2019). Auch wenn Erklärbarkeit in diesen Richtlinien häufig selbst als Prinzip oder Grundwert deklariert wird, ist sie doch i.d.R als eine Art proto-ethischer Faktor angesehen, d.h. Erklärbarkeit wird über ihre Dienlichkeit für die Bewahrung und/oder Förderung der anderen ethischen Grundwerte valorisiert (Tsamados et al. 2022).

Es bleibt allerdings unklar, ob XAI allgemein als dienlich angesehen wird oder nur für bestimmte Fälle (welche?). Wird Erklärbarkeit als notwendige oder gar hinreichende oder doch nur als förderliche Bedingung für den Schutz und die Förderung der ethischen und sozialen Grundwerte verstanden? Zudem bleibt auszuhandeln, was unter den gesetzten Werten überhaupt jeweils zu verstehen ist, sowohl den ethischen und sozialen Werten auf der einen Seite als auch Erklärbarkeit/Transparenz auf der anderen Seite. Höherstufig wäre außerdem zu klären, an welchen Kriterien oder Maßstäben sich die Festlegung dieser Punkte überhaupt orientieren und wer hierfür zuständig sein soll. Meinem Eindruck nach geht derzeit zu wenig Energie in die Klärung dieser Punkte, während zugleich sehr viel in die (technische) Umsetzung des ›value alignment‹ investiert wird. Durch diese Diskrepanz klafft eine Lücke zwischen den Prinzipien der Richtlinien und der Praxis des ›Engineering‹ auf. Die rhetorisch-diskursive omnipräsente Inwertsetzung von XAI und die zahlreichen XAI-Techniken stehen einander unvermittelt gegenüber.

Auf die klaffende Lücke zwischen der Deklaration von »high-level principles« (Hickok 2021: 41) und der Umsetzung dieser Prinzipien in die Praxis wurde mehrfach hingewiesen. Häufig zieht man als Konsequenz aus dieser Diskrepanz den Schluss, ethische Richtlinien seien nichts weiter als zahnlose Tiger und AI Ethics seien gar insgesamt unnütz (Schwartz 2004; Popescu 2016; McNamara et al. 2018; Rességuier/Rodrigues 2020; Munn 2022). Die Richtlinien und weiteren Bemühungen in dem Bereich dienten höchstens der Tech-Industrie dazu, den Anschein einer Auseinandersetzung mit gesellschaftlichen Fragen zu wahren. Mit diesem *Ethics-Washing* versuche man härtere rechtliche Regulierungen vorzubeugen (Floridi 2021: 620; Wagner 2018; Yeung et al. 2020; Bietti 2020) und sich gegen tieferbohrende Fragen der Öffentlichkeit zu immunisieren (Mittelstadt 2019: 501).

Doch aus der berechtigten Sorge vor einer Verflachung der Ethik durch unternehmerisches Marketing (Bietti 2020) lässt sich noch ein alternativer Schluss ziehen: dass die eigentliche ethische Arbeit erst nach dem Setzen von Grundwerten beginnt. Man sollte nicht missverstehen, was ethische Kodizes überhaupt leisten können und was nicht. Im besten Fall dienen sie der Verständigung über und Kodifizierung von grundlegenden Prinzipien und Werten, denen ein Kollektiv seine Handlungen und Entscheidungen gegenüber verpflichtet. Diese Dokumente können jedoch nur ein kleiner Baustein einer Auseinandersetzung mit möglichen ethischen und sozialen Konsequenzen von technologischen Entwicklungen sein. Dass sie aus sich heraus nicht unbedingt eine hinreichende Motivation darstellen, sich de facto moralisch zu verhalten, ist klar und ein altbekanntes Problem, da typischerweise verschiedene Interessen in Konflikt miteinander geraten (Fehige/Wessels 1998). Folglich entstehen neue Arbeitspakete und weiterer Klärungsbedarf: Welche Pflichten sollte wer gegenüber wem haben? Welche Anreize lassen sich für welche Fälle schaffen? Welche Motive greifen? Mit welchen Interessen steht ein ethisches ›value alignment‹ im Konflikt? Welche fördert es? Welche übersieht es?

Zu dem Motivationsbedarf kommt eine Interpretationsaufgabe. Werte sind per se unterbestimmt und daher interpretationsbedürftig. Die eigentliche Herausforderung liegt somit darin, diese zusätzlichen Aufgaben institutionell zu integrieren und dies auf eine je nach Fall/Bereich angemessene und rechtfertigbare Weise. Für die Übersetzung von Grundwerten in technische Anforderungen im Designprozess gibt es in der Fachliteratur verschiedene Heuristiken aus dem Bereich des Value-Sensitive Designs sowie des Responsible Research and Innovation (van de Poel 2016; Simon 2017; Hallensleben et al. 2020). Zudem bietet eine ganze Reihe von »frameworks, tools, and checklists« (Hickok 2021: 41) eine Orientierungshilfe für die Prozesse dieser Umsetzung, die spezifischer darauf eingehen, welche Stakeholder und Akteur:innen eigentlich wann und für welche Fragen einbezogen werden sollten. Wichtig ist es den Status dieser theoretischen Heuristiken zu beachten. Diese sind Reflexionswerkzeuge, sie bieten Hilfe zur Selbsthilfe. Sie liefern Orientierungsmittel, mit denen man sich selbst in einem gegebenen

Fall eine Orientierung verschaffen kann (Hubig 2007). Der Form nach handelt es sich nicht um algorithmisches Wissen, welches sich mit geschickten technischen Mitteln automatisieren lässt. Ebenso wenig sollten die Kodizes mit Kochrezepten verwechselt werden, die dann gut funktionieren, wenn die Kochenden auf bewährte Routinen und Mittel zurückgreifen können. Die Herausforderung für eine ethische Orientierung von KI besteht gerade darin, nur in begrenztem Maße auf tradiertes Wissen und Erfahrungen zurückgreifen zu können. Die Richtlinien sollten eher mit Kompass und Karte verglichen werden, die auf hoher See (wo es sonst relativ wenig äußere Orientierung gibt) helfen können, seinen Weg zu finden. Sie geben dabei das Ziel der Reise nicht vor, sie präferieren keinen bestimmten Pfad, aber sie können die eigenen Entscheidungen unterstützen.

Mit diesen Überlegungen sollte klar geworden sein, dass sich die Dienlichkeit von XAI-Techniken für ethische und soziale Zwecke nur im Zusammenhang mit Antworten auf die oben genannten Fragen seriös konkretisieren lässt. In diesem Sinne schlage ich vor, die Dienlichkeit von XAI für ethische und soziale Zwecke als einen echten Paradigmenwechsel zu verstehen, der ganz andere Anforderungen stellt, als es die Routinen, Werkzeuge und Ansätze der Normalwissenschaft der XAI-Forschung erahnen lassen. Hinzu kommt, dass nicht nur die Angemessenheit erprobter Mittel und Ansätze zur Diskussion stehen sollte, sondern ebenfalls die Frage der Spielregeln: Wann braucht es überhaupt eine ethische Ausrichtung von KI-Systemen, wann nicht? Worin besteht ein guter Weg, ein KI-System an ethischen und sozialen Grundwerten zu orientieren? Wer sollte hierzu eine konkrete Vorstellung entwickeln? Wer sollte sie umsetzen? Woran kann man erkennen, ob diese Umsetzung gelungen ist? Nur im Licht der Antworten auf diese Fragen lässt sich erkennen, ob XAI in einem bestimmten Fall nützlich oder gar notwendig sein kann oder nicht. Folgend demonstriere ich an Beispielen wie verwickelt diese Fragen mit der Einschätzbarkeit über die Güte von XAI für ethische und soziale Zwecke sind.

4.1 Minimierung von Diskriminierungsrisiken

Bei der Frage nach der Vereinbarkeit von KI mit gesellschaftlichen Grundwerten und Normen geht es offenkundig um normative Fragen. Diese werden in ihrer Normativität überwiegend zu wenig beachtet. Ob etwas normativ gesehen richtig oder falsch ist, ist eine Frage der Anerkennung, nicht der Erkenntnis. Es gehört zum Selbstverständnis von Demokratien, dass normative Fragen nicht dogmatisch vorgegeben werden, sondern gesellschaftlich ausgehandelt werden können. Die meisten normativen Fragen sind i.d.R. nicht thematisch; Gesellschaften ziehen ihre normative Orientierung aus ihrer Kultur, ihrer Tradition, bestehenden Moral- und Rechtssystemen usw. Interessant ist nun, dass mit dem Einsatz von KI-Systemen eine Reihe spezifischer Aushandlungsfragen im Raum stehen. Dabei können Fälle

in ihrem normativen Gewicht und ihrer Komplexität sehr verschieden sein. Ein für solche Einschätzungsfragen simples Beispiel ist der ›rassistische Seifenspender‹, auf den der Softwareentwickler Chukwuemeka Afigbo in einem Twitter-Post vom 16.08.2017 aufmerksam gemacht hatte. Aufgrund der Einstellung des Lichtsensors spendete der Seifenspender farbigen Händen keine Seife, wohl aber weißen Papiertüchern. Afigbo kommentierte: »If you have ever had a problem grasping the importance of diversity in tech and its impact on society, watch this video« (Afigbo 2017). Eine Diskriminierung im sozialen Sinne liegt vor, wenn ungerechtfertigterweise Gleiche ungleich behandelt werden oder Ungleiche gleich (Beck et al. 2019). Entscheidend ist, dass diese Diskriminierung nicht gerechtfertigt ist. Viele Antidiskriminierungsgesetzte verbieten eine Diskriminierung aufgrund geschützter Merkmale wie Rasse, Alter, Geschlecht, Sexualität und Religion (Kolleck/Orwat 2020). Dass in diesem Fall eine solche ungerechtfertigte Ungleichbehandlung vorliegt, dürfte unstrittig sein. Wir können davon ausgehen, dass es nicht im Interesse der Hersteller und Betreiber dieses Seifenspenders lag, bestimmte Personengruppen vom Gebrauch des Seifenspenders auszuschließen. Entsprechend lässt sich dieser Fall als Beispiel für eine Fehleinstellung der Hardware ansehen, der unbeabsichtigte negative Effekte hervorbrachte. Diese hätten leicht verhindert werden können, hätte man den Seifenspender ausreichend getestet. Wie Krishnan argumentiert, steht dieser Fall für ein KI-induziertes Diskriminierungsproblem, für dessen Abhilfe keine XAI-Techniken gebraucht werden (Krishnan 2020).[23] Da sogenannte *biases* an sehr vielen Stellen aus verschiedenen Gründen und Ursachen im Lebenszyklus von KI-Systemen auftreten können, scheint es insgesamt plausibel, dass verschiedene Strategien helfen können, um Diskriminierungsrisiken zu minimieren und zu managen (Klier 2024).

Fallspezifische Antworten auf folgende Fragen können diese Einschätzung informieren: (1) Bestehen Diskriminierungsrisiken? Gegenüber wem? (2) Welche Interessen haben die beteiligten Stakeholder? (3) Bestehen Werte- und/oder Interessenskonflikte? (4) Wie kann wer von diesen Risiken und Interessen Kenntnis erlangen? (5) Welche Mittel der Risikobewältigung und des Umgangs mit Konflikten stehen zur Verfügung und können hier (von wem) als angemessen angesehen werden?

4.2 Entscheidungen nachvollziehbar machen?

Wie komplex diese Fragen werden können, zeigt der viel diskutierte Fall um den Einsatz des *Correctional Offender Management Profiling for Alternative Sanctions* (COM-

23 Wie plausibel diese Einschätzung ist, hängt davon ab, was man unter Erklärbarkeit versteht oder was XAI-Techniken leisten können. Für das Testen, wann der Seifenspender Seife gibt, braucht man die Kiste sicherlich nicht zu öffnen, aber freilich braucht es eine gewisse Kenntnis von dem System, um die Einstellungen zu berichtigen.

PAS) Systems, das in den USA im Justizwesen zum Einsatz kommt. Die jüngere Debatte hatte eine investigative Recherche von ProPublica ausgelöst, die die Berechnungen von COMPAS als rassistisch eingeschätzt haben (Angwin et al. 2016). Die Software von Northpointe Inc. (heute Equivant) liefert Vorhersagen zur Risikoeinschätzung und wird als Unterstützung benutzt, wenn Bewährungsstrafen festgelegt werden, die Höhe der Kaution bestimmt wird oder auch andere Urteile von Richter:innen gefällt werden. Hierzu ordnet COMPAS die Verurteilten in verschiedene Risikogruppen ein (hohes, mittleres, geringes) bzgl. ihrer Rückfallwahrscheinlichkeit und nutzt dazu sozioökonomische Daten, Daten zur familiären Situation sowie Persönlichkeitsdaten, z.B. Stresstoleranz (Kolleck/Orwat 2020). Die Daten wurden teils aus den Akten, teils aus Befragungen bezogen. Rasse/Hautfarbe werden nicht als Daten einbezogen, es besteht jedoch der Verdacht, dass diese über sogenannte Proxyvariablen (z.B. den Wohnsitz oder das Alter) dennoch in der Berechnung des Risiko-Scores einen gewichtigen Unterschied machen.

An die Veröffentlichung des Berichts von ProPublica schloss sich eine breite Debatte an. Flores et al. bemängelten statistische Fehler und ein mangelndes Verständnis für Daten auf Seiten der investigativen Journalist:innen und behaupteten, dass ›der Algorithmus‹ gegenüber Schwarzen und Weißen gleich kalibriert sei (Flores et al. 2016). Corbett-Davies et al. betonten, dass es einen trade-off im Systemdesign zwischen einer Optimierung auf das Kriterium der öffentlichen Sicherheit und Fairnesskriterien gibt, die auf eine faire Behandlung von Individuen unabhängig ihrer Zugehörigkeit zu bestimmten Gruppen (z.B. Rasse) setzt (Corbett-Davies et al. 2017). Hamilton ergänzte die Diskussion um den Aspekt, dass Frauen von COMPAS hinsichtlich der Rückfallwahrscheinlichkeit benachteiligt würden, da das System ihnen ein zu hohes Risiko zuweisen würde (Hamilton 2019). Insgesamt führte diese Debatte zu verschiedenen Einsichten. Zunächst, dass die verschiedenen Bewertungen von COMPAS auf verschiedenen Definitionen und damit Kriterien von Fairness zurückzuführen sind (Hedden 2021). Des Weiteren, dass sich algorithmische Systeme jeweils nur auf ein Fairness-Kriterium ausrichten lassen und nicht mehreren zugleich gerecht werden können (Chouldechova/Roth 2020; Corbett-Davies et al. 2023). Damit reicht es nicht einfach aus, zu zeigen, dass bestimmte KI-Systeme fair sind, sondern man muss spezifizieren, auf welches Fairness-Kriterium die Systeme eingestellt sind und sollte begründen, warum man dieses Kriterium für diesen Fall für angemessen hält. Folglich entstehen Anerkennungsfragen und es stellt sich die Frage, wer diese normative Festlegung eigentlich zu treffen hat bzw. welche Akteur:innen hieran beteiligt sein sollten.

Die Debatte hat außerdem gezeigt, wie wichtig die institutionelle Einbettung bei der Bewertung dieser Fragen ist. Rudin et al. argumentieren grundsätzlich gegen den Einsatz proprietärer Software im Justizwesen, denn das Geschäftsgeheimnis verhindere hier, dass Betroffene oder unabhängige Dritte bzw. die Gesellschaft überhaupt eine Bewertung der Software vornehmen können (Rudin et al. 2020). Da-

mit wechselt die Diskussion auf eine andere Ebene: Ist es für staatliche Institutionen bzw. ist in bestimmten Anwendungsbereichen der Gebrauch von proprietärer Software überhaupt legitim, wenn durch diese tief in das Leben von Individuen eingegriffen wird? Rudin et al. schlagen vor, transparente Systeme zu nutzen, die von unabhängigen Expertengruppen hinsichtlich geprüft werden können, z.B. hinsichtlich verschiedener Fairnesskriterien (Rudin et al. 2020). Zur Diskussion steht damit, ob und wem ein Anspruch darauf zukommen sollte, KI-Systeme in bestimmten Anwendungsbereichen prüfen und nachvollziehen zu können (Alfrink et al. 2023). Es hängt von der Einschätzung ab, was genau es heißen soll, dass Entscheidungen nachvollziehbar sind bzw., dass KI-Systeme überprüfbar sind und welche Mittel man an dieser Stelle für angemessen hält. Auch steht zur Frage, ob ein Einsatz proprietärer Software wie COMPAS mit dem Prinzip des ordnungsgemäßen Verfahrens (›due process‹) im Einklang steht.

Walmsley argumentiert, dass aus Sicht der Betroffenen ein entscheidender Unterschied zwischen der Klassifikation durch COMPAS und einer Klassifikation, die von Anwält:innen, Richter:innen oder Zeug:innen vorgenommen wird, besteht, denn sie können COMPAS nicht »cross-examine […] in the same way they could with a human witness« (Walmsley 2021: 592). Es geht also nicht nur darum, die Software in bestimmter Weise überhaupt nachvollziehbar zu machen (entweder, indem man nur transparente Software für bestimmte Fälle nutzt, oder durch XAI), sondern auch um die Frage, für wen Software in welcher Weise und in welchem Umfang nachvollziehbar sein muss (Burrell 2016). Die meisten Angeklagten werden kein Expertenwissen des Machine Learning aufweisen. Was kann es heißen, dass die Entscheidungen für sie dennoch nachvollziehbar sind? In wessen Hand legt man diese Übersetzungsarbeit? Gehört sie künftig zum Beruf der Anwältin? Oder sollte man Angeklagten eine Art Recht auf Beratung durch KI-Expert:innen zugestehen? Wenn ja, woher kommt diese Expertise und wer bezahlt sie? Die COMPAS-Debatte zeigt deutlich, dass man bei bestimmten Anwendungsbereichen zunächst eine ganze Reihe von Fragen der institutionellen Einbettung von KI klären sollte, bevor man sinnvoll einschätzen kann, welche XAI-Techniken für diese Fälle überhaupt zuverlässig sind.

4.3 Unternehmerische Techniken des Value Alignments

Es sind viele Fälle denkbar, bei denen es wie im Seifenspender-Beispiel ausreicht, die Systeme vor Gebrauch hinreichend zu testen, um Diskriminierungsrisiken zu minimieren. Andere Fälle, vielleicht all solche Anwendungen, die im Rahmen des EU AI Acts in die Gruppe der hohen Risiken gehören, ähneln eher dem Beispiel von COMPAS. Da hier die Anwendung von KI sowohl individuelle Grundrechte als auch das Gemeinwohl tangiert, spricht vieles dafür, höhere (regulative) Anforderungen

für ihren Einsatz zu stellen, etwa die Forderung nach Nachvollziehbarkeit der Entscheidungen.

Darüber hinaus gibt es wenigstens eine weitere Gruppe von Fällen, bei denen Stereotype reproduziert werden und Diskriminierungsrisiken zu verzeichnen sind, die jedoch nicht per se in die Gruppe der hohen Risiko-Anwendungen gehören. Hierzu gehören all jene Social Media und Onlineangebote, in denen KI-gestützt Inhalte moderiert oder generiert werden oder in andere Entscheidungen in Kommunikation, Information und Konsum eingreifen. Diese Fälle scheinen mir normativ höhere Unsicherheiten mit sich zu bringen, weil weniger klar ist, welche politische, soziale und moralische Brisanz man ihnen zuschreiben soll. Wäre es wünschenswert, dass der Gesetzgeber auch für diese Fälle bestimmte Auflagen vorsieht? Aus welchen guten Gründen? Ungeachtet dieser normativen Unsicherheiten haben Tech-Unternehmen längst diverse technische Strategien der Handhabe etabliert. Diese Fälle scheinen mir weitestgehend unter dem Radar der XAI-Community zu fliegen (Finke et al. 2022). Sie sind aber allein darum schon aufschlussreich, da man es hiermit, aus technischer Sicht, mit dem gleichen Problem wie bei COMPAS zu tun hat: »The problem of achieving agreement between our true preferences and the objective we put into the machine is called the *value alignment problem*: the values or objectives put into the machine must be aligned with those of the human« (Russell/Norvig 2021, Hervorh. im Original). Wenn man alle Fälle, vom Seifenspender über COMPAS bis zur Content Moderation als Problem des ›value alignment‹ konzeptualisiert, liegt es nahe, die in diesem Bereich etablierten technischen Lösungen auch für die anderen Fälle einzusetzen.

Eine simple technische Lösung für unerwünschte Nebeneffekte sind eingebaute Filter. Drei Jahre, nachdem der Post von Alciné zur Missklassifikation von Google Photos viral ging, berichtete das Tech Magazine *Wired* über einen von ihnen durchgeführten Test, der darauf schließen ließ, dass Google bestimmte Kategorien aus ihrem Klassifikationssystem schlicht gelöscht hat, um der rassistischen Zuordnung Herr zu werden. *Wired* hatte mit 40.000 Fotos die Klassifizierung getestet, darunter Fotos von Affen, die aber nicht als solche kategorisiert wurden. Google habe darauf bestätigt, die Kategorien ›gorilla‹, ›chimp‹, ›chimpanzee‹ und ›monkey‹ aus ihrem System gelöscht zu haben (Simonite 2018; Vincent 2018).

Filterfunktionen gibt es in verschiedenen Komplexitätsstufen. Eine weitere Strategie ist eine bestimmte Form der Mensch-Maschine-Arbeitsteilung, wofür das Einrichten des Large-Language-Models von OpenAI, auf dem ChatGPT-4 basiert, illustrativ ist. Informationen hierzu finden sich auf der Webseite von OpenAI sowie in dem *Technical Report* zu ChatGPT-4. OpenAI hat sich selbst das Ziel gestellt, seine Chatbots in Einklang mit drei Werten – den drei H's – zu entwickeln: Die Ausgaben der Bots sollten »honest«, »helpful« und »harmless« sein (Open AI 2023). Ehrlichkeit (»honesty«) ist begrifflich eine missverständliche Wahl, da sie sich auf den inneren Zustand eines Subjektes bezieht. Sachlich gemeint ist, dass die Aussagen

korrekt/wahr sein sollen (Heitzinger/Woltran 2024: 143). In seinem technischen Bericht gibt OpenAI zwei Schritte an, in denen sie das Large-Language-Modell erstellt haben: ein »pre-training« und ein »fine-tuning« des Modells. Aus beiden Lernwegen wurde dann das eigentliche Modell gebildet (Open AI 2023). Das pre-training fand automatisiert statt (überwachtes Lernen eines artifiziellen neuronalen Netzes), auf einer sehr großen Datenmenge, die zum Teil aus frei zugänglichen Internetquellen und andererseits aus lizenzierten Datenquellen stammen (OpenAI et al. 2024). Auf diesem Wege wurde das Modell in die Lage gebracht, zu bestimmen, welches Wort (token) sinnvollerweise auf ein vorheriges Wort erscheinen soll – die statistische Grundlage dieser Generierung von Texten. Der zweite Schritt dient dem Zweck des ›value-alignment‹. ChatGPT soll nur solche Text generieren, die mit ›unseren‹ Grundwerten vereinbar sind. Der Schritt des Fine-Tunings ist folglich eine Kontrollmaßnahme:

> »Then, we ›fine-tune‹ these models on a more narrow dataset that we carefully generate with human reviewers who follow guidelines that we provide them. Since we cannot predict all the possible inputs that future users may put into our system, we do not write detailed instructions for every input that ChatGPT will encounter. Instead, we outline a few categories in the guidelines that our reviewers use to review and rate possible model outputs for a range of example inputs. Then, while they are in use, the models generalize from this reviewer feedback in order to respond to a wide array of specific inputs provided by a given user.« (Open AI 2023)

Da die Daten, auf denen das Modell trainiert wurde, unerwünschte Inhalte beinhalten (›toxic content‹), lernt das Modell ebendiese (z.B. Beschimpfungen, diskriminierende Aussagen, Verleumdungen des Holocausts). Ohne die manuell-maschinelle Kontrollmaßnahme würde das Modell diese Inhalte ungefiltert ausgeben, wenn es die Token-Wahrscheinlichkeiten nahe legen würde. Die Kontrollmaßnahme, von Unternehmensseite auch als Sicherheitsmaßnahme verstanden (›do no harm‹), baut auf einer spezifischen Arbeitsteilung von Maschine und Mensch auf. Da die Maschine nicht allein beurteilen kann, ob generierter Text mit den drei H's vereinbar ist oder nicht, braucht es eine Beurteilung durch Menschen. Da die Menschen aber nicht den ganzen Datensatz prüfen können und praktisch nicht jeder denkbare Prompt vorab präventiv eingehegt werden kann, haben die ›human reviewers‹ die Prüfung nur auf einer vergleichsweise viel geringeren Datenmenge durchgeführt, damit diese Prüfung überhaupt für menschliche Arbeitskräfte (auch für große Gruppen) durchführbar ist. Die ›humans‹ waren dafür zuständig, vier vom Modell gebildete Antworten auf einen Prompt in eine Reihenfolge zu bringen (von der besten zur schlechtesten). Dieses Ranking wurde dann als Feedback in das Modell zurückgegeben, wodurch dieses seine Zuordnungen von Antworten zu Prompts

weiter optimierte, nach der Strategie des »reinforcement learning from human feedback (RLHF)« (Heitzinger/Woltran 2024: 143).

Für den Prüfvorgang waren den Arbeitskräften Instruktionen vorgegeben, worauf sie achten sollten, z. B. auf die Konformität mit geltendem Gesetz: »do not complete requests for illegal content« (Open AI 2023). Sodann gab es auch ›höherstufige‹ Anweisungen wie »avoid taking a position on controversial topics« (Open AI 2023). Außerdem wurden wöchentliche Treffen des OpenAI Management mit den Prüfer:innen anberaumt, um fragwürdige Fälle und offene Fragen zu klären: »This iterative feedback process is how we train the model to be better and better over time« (Open AI 2023).

Mittlerweile ist ein eigener Geschäftsbereich dadurch entstanden, KI so zu kalibrieren, dass sie sich nicht unerwünscht ›verhält‹. An diese etablierte Praxis lassen sich zahlreiche normative Fragen anschließen. OpenAI wirft in ihrem Blog selbst die Frage auf, wer eigentlich die Werte, Instruktionen und Standards bestimmen sollte, an denen diese Systeme orientiert werden (Open AI 2023). Es wäre auch zu diskutieren, ob durch solche normative Festlegungen privater Akteur:innen die kulturelle Hegemonie des Westens in bestimmten Bereichen des Internets weiter gefestigt wird (Goffi et al. 2021; Siapera 2022; Shahid/Vashistha 2023).

Dazu gesellen sich Fragen zu den Arbeitsbedingungen der menschlichen KI-Unterstützer:innen (Perrigo 2023; Hagendorff 2022); und dies umso mehr, weil die Arbeit der Einhegung dieser Software wohl auf Dauer stattfinden muss, wie die zahlreichen Beispiele des »Jailbreaks« zeigen, in denen die normativen Regeln des Systems umgegangen wurden (Xie et al. 2023). Hinzu kommen weitere Fragen zur Vereinbarkeit von Software dieser Art mit anderen hier nicht diskutierten gesellschaftlichen Grundwerten, wie Umweltschutz und Nachhaltigkeit (George et al. 2023; Khowaja et al. 2024). Hier ließe sich diskutieren, ob der hohe CO_2-Ausstoß gepaart mit dem Potential von Large-Language-Models, wie ChatGPT-4, in diversen elektronischen Geräten und Services als KI-Aufrüstung verbaut zu werden, nicht mit dem öffentlichen Interesse einhergeht, über die Umweltkosten dieses technischen Fortschritts aufgeklärt zu werden.

5. Über die Verschiedenheit der Zwecksetzungen

Die XAI-Forschung täte gute daran, ernsthaft über den Zweck nachzudenken, für den XAI entwickelt und optimiert wird (Krishnan 2020; Colaner 2022; Alpsancar et al. 2024). Freiesleben und König (2023) zufolge, kommen die meisten Beiträge in der XAI-Forschung ohne eine Zwecksetzung ihrer Tools aus, sie reflektieren nicht auf die Dienlichkeit ihrer Technologie, sondern kümmern sich abstrakt um deren Optimierung. Ich vermute, dieser Eindruck trifft zu einem großen Teil zu, er übersieht allerdings, dass sehr wohl Zwecke genannt und Dienlichkeiten gesetzt werden

– dies passiert allerdings erstens pauschal und generalisierend und zweitens eher rein diskursiv, d.h. ohne forschungspraktischen Bezug auf die Arbeit an den XAI-Technologien. Entsprechend rege ich nicht nur dazu an, über die Zwecke zu reflektieren und eine konkrete Zwecksetzung anzugeben, sondern ebenso über die kategoriale und forschungspraktische Verschiedenheit der Zwecksetzungen Klarheit zu gewinnen.

Im Paradigma der epistemischen Güte von ML ist der Sinn des Forschens innerhalb der ML-Community lokalisiert. Hier entwickeln einige Expert:innen für andere Expert:innen bessere Einsichten in bestimmte formale Zusammenhänge von ML-Systemen, um diese optimieren zu können und Fehler zu beheben. Die Relevanz der Erklärungen ergibt sich hier aus der Forschung der Community selbst: Welche Systeme sollen in welcher Hinsicht besser verstanden werden. Die Evaluation der XAI-Techniken und deren Maßstäbe ergeben sich aus dieser Zwecksetzung und sollten sich den üblichen methodischen Standards fügen.[24]

Im zweiten Paradigma verändert sich offenkundig die Zusammensetzung der Akteur:innen und Nutzer:innen, und Kontexte werden Teil des Forschungsgegenstandes, die gemäß der methodischen Standards aus der Psychologie, Sozial- und Kulturwissenschaft erforscht werden sollten. Der Methodenkasten erweitert sich disziplinär und die interdisziplinäre Zusammenarbeit wird zum Thema. Außerdem mag es von Bedeutung sein, in einigen Fällen partizipative Verfahren zu integrieren. Die Relevanz und Angemessenheit von Erklärungen lässt sich in diesem Paradigma nicht (allein) aus technischen Überlegungen gewinnen; es braucht folglich einen echten Perspektivwechsel. Mit meinem Kommentar zum XAI-Forschungsprojekt der DARPA wollte ich demonstrieren, dass es wichtig ist, sich über strukturelle Fragen bezüglich der Zuständigkeit von Forschungsfragen und Aufgaben Gedanken zu machen. Dies betrifft ebenfalls die Gewichtung der Beteiligung verschiedener Expertisen, also eine forschungspolitische Frage. Ein Klärungsbedarf besteht auch in konzeptioneller Hinsicht. Die von mir herausgestellte Differenz der ›informatischen‹ und der ›psychologischen‹ Perspektive im XAI-Projekt zeigt, dass praktische Unterschiede darin bestehen, wie man den Erklärungsprozess modelliert. Für eine erfolgreiche interdisziplinäre Zusammenarbeit scheint es hilfreich, sich über die Zielvision zu verständigen. Es ist etwas vollkommen anderes, XAI-Techniken mit

24 Nach Freiesleben und König (2023) ist die Community noch dabei diese auszubilden und befindet sich deswegen noch nicht in einem paradigmatischen Zustand der normalwissenschaftlichen Forschung im Sinne Kuhns. Dies mag der Fall sein, betrifft aber den Kerngedanken meines Arguments nicht – dass die normalwissenschaftliche Forschung eines Paradigmas nicht ohne weiteres auf die normalwissenschaftliche Forschung eines anderen Paradigmas übertragen werden kann und, dass wir es in der XAI-Welt mit wenigstens drei grundsätzlich zu unterscheidenden Paradigmen zu tun haben.

Blick auf das Endziel der Automatisierung zu entwickeln (unsere Erklärungen werden irgendwann so gut, dass alles selbst-evident ist, und dann brauchen wir keine User:innen für die Zwecke der XAI) oder, ob man die Techniken daraufhin optimiert, dass Teams von User:innen und KI-Systemen kollaborieren sollen. Diese Zielvorstellungen implizieren ein anderes Gewicht der Kontextualität des Gebrauchs von KI. Während er im Automatisierungs-Leitbild zu vernachlässigen ist, denn die ideale Automatisierung ist eben unabhängig von Kontexten (so die Ideologie), ist es in der Kollaborations-Vision gerade entscheidend, von welchen Kontexten und konkreten Konstellationen man ausgeht. Je nach Zielvision gilt es andere forschungspraktische Fragen zu berücksichtigen.

Der Diskurs der AI Ethics und das Problem des ›value alignment‹ konstituieren meiner Ansicht nach ein weiteres, drittes Paradigma des Forschens und Entwickelns von XAI, für das im Sinne Kuhns die erprobten Instrumente, Methoden und theoretischen Ansätze der anderen beiden Paradigmen nicht geeignet sein können – weil wir es hier mit einem anderen Typus von Problem zu tun haben, der andere Konzeptionen von Lösungsräumen und -strategien erfordert. Dieses Paradigma ist auch darum besonders, viel weniger im Bereich der Forschung verortet zu sein als ›in der Gesellschaft‹ – in Industrie, Politik, Recht und der darauf bezogenen Forschung und Entwicklung. Hier ist es nicht eine methodische Option, echte Kontexte und Nutzer:innen, z.B. über Feldstudien, Interviews oder partizipative Verfahren, in die (inter-)disziplinäre Forschung einzubeziehen, sondern die Bearbeitung des Problems findet primär in anderen gesellschaftlichen Bereichen als der Wissenschaft statt. Dementsprechend haben wir es in diesem Paradigma noch einmal mit einer grundverschiedenen Konstellation von Akteur:innen und Strukturen, sodann Machtverhältnissen zu tun.

Macht man sich die Verschiedenheit dieser drei Paradigmen bewusst, sollte es überraschen, dass de facto relativ ähnliche Lösungsstrategien des Engineerings für diese verschiedenen Probleme ausprobiert werden und ›laufen‹. Es wäre an der Zeit, das Engagement und die Expertise der Ingenieurswissenschaften, Data Science und Informatik mit anderen Expertisen so zu kombinieren, dass sich der alte Fehler eines ›technological fix‹ (de Bruijn et al. 2022) nicht im Bereich der Entwicklung und Optimierung von XAI wiederholt. Stattdessen könnte man das Feld der XAI diversifizieren, das (informelle) Wissen verschiedener sozialer Stakeholder einbeziehen, eine Debatte über gesellschaftliche Nützlichkeit von XAI und KI anregen und die mit ihrer Nutzung einhergehenden Risiken und Chancen bedacht abwägen, wobei man auch reflektieren sollte, warum, wie und für wen die (vermeintlichen) Zwecke von XAI und KI profitabel/nützlich sind.

Literatur

Adadi, A.; Berrada, M. (2018): Peeking inside the black-box. A survey on explainable artificial intelligence (XAI), in: *IEEE access*, 6, 52138–52160.

Afigbo, C. (2017): Post from @nke_ise, in: Twitter/X: 16.08.2017. [https://twitter.com/nke_ise/status/897756900753891328] (Zugriff: 28.12.2023).

AIAAIC (2021): Facebook labels black men ›primates‹. [https://www.aiaaic.org/aiaaic-repository/ai-algorithmic-and-automation-incidents/facebook-labels-black-men-primates] (Zugriff: 28.12.2023).

Alfrink, K.; Keller, I.; Kortuem, G.; Doorn, N. (2023): Contestable AI by design. Towards a Framework, in: *Minds and Machines*, 33(4), 613–639.

Allhutter, D.; Mager, A.; Cech, F.; Fischer, F.; Grill, G. (2020): Der AMS-Algorithmus. Eine Soziotechnische Analyse des Arbeitsmarktchancen-Assistenz-Systems (AMAS). Endbericht, Wien: Institut für Technikfolgen-Abschätzung (ITA).

Alpsancar, S. (2023): What is AI Ethics? Ethics as means of self-regulation and the need for critical reflection, in: International Conference on Computer Ethics, 1(1), 1–17. [https://soremo.library.iit.edu/index.php/CEPE2023/article/view/227].

Alpsancar, S.; Matzner, T.; Philippi, M. (2024): Unpacking the purposes of explainable AI, in: Arias-Olivia, M.; Pelegrin-Borondo, J.; Murata, K.; Palma, A.M.L.; Ollé Sensé, M. (Hg.), Smart Ethics in the Digital World. Proceedings of the ETHICOMP 2024. 21st International Conference on the Ethical and Social Impacts of ICT, Logroño: Universidad de La Rioja, 31–35. [https://dialnet.unirioja.es/descarga/articulo/9326091.pdf].

Angwin, J.; Larson, J.; Mattu, S.; Kirchner, L. (2016): Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks, in: ProPublica. [https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing] (Zugriff: 15.06.2024).

Beck, S.; Grunwald, A.; Jacob, K.; Matzner, T. (2019): Künstliche Intelligenz und Diskriminierung. Herausforderungen und Lösungsansätze (Whitepaper), München: Plattform Lernende Systeme.

Bellon, J.; Gransche, B.; Nähr-Wagener, S. (Hg.) (2022): Soziale Angemessenheit. Forschung zu Kulturtechniken des Verhaltens, Wiesbaden: Springer.

Benjamin, R. (2019): Race After Technology. Abolitionist Tools for the New Jim Code, Cambridge/Medford (MA): Polity.

Bietti, E. (2020): From Ethics Washing to Ethics Bashing. A View on Tech Ethics from within Moral Philosophy, in: FAT* '20. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, New York: Association for Computing Machinery, 210–219.

Biran, O.; Cotton, C. (2017): Explanation and justification in machine learning. A survey, in: *IJCAI-17 Workshop on Explainable AI (XAI)*, 8(1), 8–13.

Blumenberg, H. (1981): Technisierung und Lebenswelt unter Aspekten der Phänomenologie, in: Ders., Wirklichkeiten in denen wir leben. Aufsätze und eine Rede, Stuttgart: Reclam, 7–54.

Borup, M.; Brown, N.; Konrad, K.; Van Lente, H. (2006): The sociology of expectations in science and technology, in: *Technology analysis & strategic management*, 18(3/4), 285–298.

Bradshaw, J.M.; Hoffman, R.R.; Woods, D.D.; Johnson, M. (2013): The Seven Deadly Myths of Autonomous Systems, in: *IEEE Intelligent Systems*, 28(3), 54–61.

Burrell, J. (2016): How the machine thinks. Understanding opacity in machine learning algorithms, in: *Big Data & Society*, 3(1). [doi.org/10.1177/2053951715622512].

Cabitza, F.; Campagner, A.; Malgieri, G.; Natali, C.; Schneeberger, D.; Stoeger, K.; Holzinger, A. (2023): Quod erat demonstrandum? Towards a typology of the concept of explanation for the design of explainable AI, in: *Expert Systems with Applications*, 213, 1–16.

Capel, T.; Brereton, M. (2023): What is Human-Centered about Human-Centered AI? A Map of the Research Landscape, in: CHI '23. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, New York: Association for Computing Machinery, 1–23.

Caruana, R.; Lou, Y.; Gehrke, J.; Koch, P.; Sturm, M.; Elhadad, N. (2015): Intelligible Models for HealthCare. Predicting Pneumonia Risk and Hospital 30-Day Readmission, in: KDD '15. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York: Association for Computing Machinery, 1721–1730.

Cath, C.; Wachter, S.; Mittelstadt, B.; Taddeo, M.; Floridi, L. (2018): Artificial intelligence and the ›good society‹. The US, EU, and UK approach, in: *Science and engineering ethics*, 24, 505–528.

Chakrabarti, R.; Sanyal, K. (2020): Towards a ›Responsible AI‹. Can India take the lead?, in: *South Asia Economic Journal*, 21(1), 158–177.

Chouldechova, A.; Roth, A. (2020): A snapshot of the frontiers of fairness in machine learning, in: *Communication of the ACM*, 63(5), 82–89.

Clancey, W.J.; Hoffman, R.R. (2021): Methods and standards for research on explainable artificial intelligence. Lessons from intelligent tutoring systems, in: *Applied AI Letters*, 2(4), 1–8.

Colaner, N. (2022): Is explainable artificial intelligence intrinsically valuable?, in: *AI & Society*, 37(1), 231–238.

Corbett-Davies, S.; Pierson, E.; Feller, A.; Goel, S.; Huq, A. (2017): Algorithmic Decision Making and the Cost of Fairness, in: KDD '17. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York: Association for Computing Machinery, 797–806.

Corbett-Davies, S.; Gaebler, J.D.; Nilforoshan, H.; Shroff, R.; Goel, S. (2023): The Measure and Mismeasure of Fairness, in: *Journal of Machine Learning Research*, 24(312), 1–117.

Crawford, K. (2021): Atlas of AI. Power, Politics, and the Planetary Costs of Artificial Intelligence, New Haven/London: Yale University Press.

Crawford, K. (2023): Archeologies of Datasets, in: *The American Historical Review*, 128(3), 1368–1371.

Cremers, A.B.; Englander, A.; Gabriel, M.; Hecker, D.; Mock, M.; Poretschkin, M.; Rosenzweig, J.; Rostalski, F.; Sicking, J.; Volmer, J. et al. (2019): Trustworthy Use of Artificial Intelligence. Priorities from a Philosophical, Ethical, Legal and Technological Viewpoint as a Basis for Certification of Artificial Intelligence, Sankt Augustin: Fraunhofer Institute for Intelligent Analysis.

Defense Advanced Research Projects Agency (DARPA) (2016): Broad Agency Announcement. Explainable Artificial Intelligence (XAI), 1–52. [https://www.darpa.mil/program/explainable-artificial-intelligence] (Zugriff: 24.04.2024).

de Graaf, M.M.; Malle, B.F. (2017): How people explain action (and autonomous intelligent systems should too), in: AAAI Fall Symposium Series 2017, Washington [DC]: AAAI Press,19-26.

de Bruijn, H.; Warnier, M.; Janssen, M. (2022): The perils and pitfalls of explainable AI. Strategies for explaining algorithmic decision-making, in: *Government information quarterly*, 39(2), 1–8.

Dignum, V. (2019): Responsible Artificial Intelligence. How to Develop and Use AI in a Responsible Way, Cham: Springer.

Dix, A. (2017): Human–computer interaction, foundations and new paradigms, in: *Journal of Visual Languages Computing*, 42, 122–134.

Doshi-Velez, F.; Kim, B. (2017): Towards a rigorous science of interpretable machine learning, in: arXiv. [https://arxiv.org/abs/1702.08608] (Zugriff: 11.06.2024).

Edwards, P.N. (1997): The Closed World. Computers and the Politics of Discourse in Cold War America, Cambridge (MA): The MIT Press.

Ellmer, M. (2015): Digitale Arbeitsteilung. Amazon Mechanical Turks sozial konstruierte Designmuster und die Steuerung von Human-Computation-Arbeit, in: *Momentum Quarterly*, 4(3), 174–186.

Eubanks, V. (2017): Automating inequality. How high-tech tools profile, police, and punish the poor, New York: St. Martin's Press.

European Commission (2022): Regulatory framework proposal on artificial intelligence. [https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai] (Zugriff: 15.06.2024).

Everingham, M.; Eslami, S.M.A.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. (2015): The Pascal Visual Object Classes Challenge. A Retrospective, in: *International Journal of Computer Vision*, 111(1), 98–136.

Everingham, M.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. (2010): The Pascal Visual Object Classes (VOC) Challenge, in: *International Journal of Computer Vision*, 88(2), 303–338.

Fehige, C.; Wessels, U. (1998): Preferences. An introduction, in: Fehige, C.; Wessels, U. (Hg.), Preferences, Berlin/New York: de Gruyter, xx-xliii.

Finke, J.; Horwath, I.; Matzner, T.; Schulz, C. (2022): (De)Coding Social Practice in the Field of XAI. Towards a Co-constructive Framework of Explanations and Understanding Between Lay Users and Algorithmic Systems, in: Degen, H.; Ntoa, S. (Hg.), Artificial Intelligence in HCI, Cham: Springer, 149–160.

Flores, A.W.; Bechtel, K.; Lowenkamp, C.T. (2016): False Positives, False Negatives, and False Analyses. A Rejoinder to ›Machine Bias. There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks‹, in: *Federal Probation*, 80(2), 38–46.

Floridi, L. (2021): The End of an Era. From Self-Regulation to Hard Law for the Digital Industry, in: *Philosophy & Technology*, 34(4), 619–622.

Floridi, L.; Cowls, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; Dignum, V.; Luetge, C.; Madelin, R.; Pagallo, U.; Rossi, F. et al. (2018): AI4People – An Ethical Framework for a Good AI Society. Opportunities, Risks, Principles, and Recommendations, in: *Minds and Machines*, 28(4), 689–707.

Fortes, P.R.B. (2020): Paths to digital justice. Judicial robots, algorithmic decision-making, and due process, in: *Asian Journal of Law and Society*, 7(3), 453–469.

Freiesleben, T.; König, G. (2023): Dear XAI community, we need to talk! Fundamental misconceptions in current XAI research, in: Longo, L. (Hg.), Explainable Artificial Intelligence. First World Conference, Cham: Springer, 48–65.

Friedman, B.; Nissenbaum, H. (1996): Bias in computer systems, in: *ACM Transactions on Information Systems (TOIS)*, 14(3), 330–347.

Gehring, P. (2016): Ethik als Realexperiment von Rechtspolitik. Zum Dreiecksverhältnis von Bioethik, Recht und Politik, in: *Jahrbuch für Wissenschaft und Ethik*, 20(1), 143–162.

George, A.S.; George, A.H.; Martin, A.G. (2023): The Environmental Impact of AI. A Case Study of Water Consumption by Chat GPT, in: *Partners Universal International Innovation Journal*, 1(2), 97–104.

Gilpin, L.H.; Bau, D.; Yuan, B.Z.; Bajwa, A.; Specter, M.; Kagal, L. (2018): Explaining Explanations. An Approach to Evaluating Interpretability of Machine Learning, in: arXiv. [https://arxiv.org/abs/1806.00069] (Zugriff 14.06.2024).

Goffi, E.R.; Colin, L.; Belouali, S. (2021): Ethical Assessment of AI Cannot Ignore Cultural Pluralism. A Call for Broader Perspective on AI Ethic, in: *Arribat-International Journal of Human Rights Published by CNDH Morocco*, 1(2), 151–175.

Goodman, B.; Flaxman, S. (2017): European Union regulations on algorithmic decision-making and a »right to explanation«, in: *AI magazine*, 38(3), 50–57.

Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D. (2018): A survey of methods for explaining black box models, in: *ACM computing surveys (CSUR)*, 51(5), 1–42.

Gunning, D. (2016): Explainable Artificial Intelligence (XAI). DARPA/I2O. [https://sites.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf] (Zugriff: 14.05.2024).

Gunning, D.; Aha, D. (2019): DARPA's Explainable Artificial Intelligence (XAI) Program, in: *AI Magazine*, 40(2), 44–58.

Gunning, D.; Stefik, M.; Choi, J.; Miller, T.; Stumpf, S.; Yang, G.-Z. (2019): XAI – Explainable artificial intelligence, in: *Science Robotics*, 4(37). [doi.org/10.1126/scirobotics.aay7120].

Gunning, D.; Vorm, E.; Wang, J.Y.; Turek, M. (2021): DARPA's explainable AI (XAI) program. A retrospective, in: *Applied AI Letters*, 2, 1–11.

Gyevnar, B.; Ferguson, N.; Schafer, B. (2023): Bridging the transparency gap. What can explainable AI learn from the AI Act?, in: Gal, K.; Nowé, A.; Nalepa, G.J.; Fairstein, R.; Rădulescu, R. (Hg.), ECAI 2023. 26th European Conference on Artificial Intelligence, Amsterdam u.a.: IOS Press, 964–971.

Hagendorff, T. (2020): The Ethics of AI Ethics. An Evaluation of Guidelines, in: *Minds and Machines*, 30(1), 99–120.

Hagendorff, T. (2022): Blind Spots in AI Ethics, in: *AI and Ethics*, 2(4), 851–867.

Hallensleben, S.; Hustedt, C.; Fetic, L.; Fleischer, T.; Grünke, P.; Hagendorff, T.; Hauer, M.; Hauschke, A.; Heesen, J.; Herrmann, M. et al. (2020): From Principles to Practice. An interdisciplinary framework to operationalise AI ethics (Technischer Bericht), Gütersloh: Bertelsmann Stiftung.

Hamilton, M. (2019): The sexist algorithm, in: *Behavioral sciences & the law*, 37(2), 145–157.

Harrison, S.; Tatar, D.; Sengers, P. (2007): The three paradigms of HCI, in: Alt. Chi. Session at the SIGCHI Conference on human factors in computing systems San Jose, California, USA, New York: Association for Computing Machinery, 1–18.

Hedden, B. (2021): On statistical criteria of algorithmic fairness, in: *Philosophy and Public Affairs*, 49(2), 209–231.

Heitzinger, C.; Woltran, S. (2024): A Short Introduction to Artificial Intelligence. Methods, Success Stories, and Current Limitations, in: Werthner, H.; Ghezzi, C.; Kramer, J.; Nida-Rümelin, J.; Nuseibeh, B.; Prem, E.; Stanger, A. (Hg.), Introduction to Digital Humanism. A Textbook, Cham: Springer, 135–149.

Hern, A. (2018): Google's solution to accidental algorithmic racism. Ban gorillas, in: *The Guardian*, 12.01.2018. [https://www.theguardian.com/technology/2018/jan/12/google-racism-ban-gorilla-black-people?CMP=share_btn_url] (Zugriff: 28.12.2023).

Hernández-Orallo, J. (2019): Gazing into Clever Hans machines, in: *Nature Machine Intelligence*, 1(4), 172–173.

Hickok, M. (2021): Lessons learned from AI ethics principles for future actions, in: *AI and Ethics*, 1(1), 41–47.

High Level Expert Group (2019): A definition of AI. Main capabilities and disciplines. [https://digital-strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines] (Zugriff: 15.06.2024).

Hilgendorf, E. (2020): Robotik, Künstliche Intelligenz, Ethik und Recht. Neue Grundlagenfragen des Technikrechts, in: Hentschel, A.; Hornung, G.; Jandt, S. (Hg.), Mensch – Technik – Umwelt. Verantwortung für eine sozialverträgliche Zukunft, Baden-Baden: Nomos, 545–564.

Hoffman, R.R.; Miller, T.; Clancey, W.J. (2022): Psychology and AI at a Crossroads. How Might Complex Systems Explain Themselves?, in: *American Journal of Psychology*, 135(4), 365–378.

Hoffman, R.R.; Miller, T.; Klein, G.; Mueller, S.T.; Clancey, W.J. (2023): Increasing the Value of XAI for Users. A Psychological Perspective, in: *KI-Künstliche Intelligenz*, 37, 237–247.

Hu, M. (2020): Cambridge Analytica's black box, in: *Big Data & Society*, 7(2), 1–6.

Hubig, C. (2007): Die Kunst des Möglichen II. Grundlinien einer dialektischen Philosophie der Technik, Band 2. Ethik der Technik als provisorische Moral, Bielefeld: transcript.

Hüllermeier, E. (2020): Towards Analogy-Based Explanations in Machine Learning, in: Torra, V.; Narukawa; Y.; Nin, J.; Agell, N. (Hg.), Modeling Decisions for Artificial Intelligence, Cham: Springer, 205–217.

Ipeirotis, P.G. (2010): Analyzing the Amazon Mechanical Turk Marketplace, in: *XRDS: Crossroads, The ACM magazine for students*, 17(2), 16–21.

Jobin, A.; Ienca, M.; Vayena, E. (2019): Artificial Intelligence. The global landscape of AI ethics guidelines, in: *Nature Machine Intelligence*, 1, 389–399.

Kamath, U.; Liu, J. (2021): Explainable Artificial Intelligence. An Introduction to Interpretable Machine Learning, Cham: Springer.

Kaminski, A. (2010): Technik als Erwartung. Grundzüge einer allgemeinen Technikphilosophie, Bielefeld: transcript.

Kaminski, A. (2014): Sein und »als«. Notizen zu einer Denkfigur in Heideggers Werk, in: *Filozofija i društvo*, 25(4), 21–28.

Kaminski, M.E. (2019): The right to explanation, explained, in: *Berkeley Technology Law Journal*, 34(1), 189–218.

Kasperkevic, J. (2015): Google says sorry for racist auto-tag in photo app, in: *The Guardian*, 01.07.2015. [https://www.theguardian.com/technology/2015/jul/01/google-sorry-racist-auto-tag-photo-app?CMP=share_btn_url] (Zugriff: 05.06.2024).

Kearns, M.; Roth, A. (2019): The Ethical Algorithm. The Science of Socially Aware Algorithm Design, New York: Oxford University Press.

Khowaja, S.A.; Khuwaja, P.; Dev, K.; Wang, W.; Nkenyereye, L. (2024): ChatGPT Needs SPADE (Sustainability, PrivAcy, Digital divide, and Ethics) Evaluation. A Review, in: *Cognitive Computation*. [doi.org/10.1007/s12559-024-10285-1].

Kim, M.; Kim, S.; Kim, J.; Song, T.-J.; Kim, Y. (2024): Do stakeholder needs differ? Designing stakeholder-tailored Explainable Artificial Intelligence (XAI) interfaces, in: *International Journal of Human-Computer Studies*, 181, 1–12.

Klier, M. (2024): Grundlagen zu Bias & Fairness in KI-Systemen. [https://bias-and-fairness-in-ai-systems.de/grundlagen/] (Zugriff: 31.05.2024).

Kolleck, A.; Orwat, C. (2020): Mögliche Diskriminierung durch algorithmische Entscheidungssysteme und maschinelles Lernen. Ein Überblick, Berlin: Büro für Technikfolgen-Abschätzung beim Deutschen Bundestag (TAB).

Kraus, T.; Ganschow, L. (2022): Anwendungen und Lösungsansätze erklärbarer Künstlicher Intelligenz, in: Hartmann, E.A. (Hg.), Digitalisierung souverän gestalten II, Berlin/Heidelberg: Springer, 38–50.

Krishnan, M. (2020): Against interpretability. A critical examination of the interpretability problem in machine learning, in: *Philosophy & Technology*, 33(3), 487–502.

Kuhn, T.S. (1976[1962]): Die Struktur wissenschaftlicher Revolutionen. Bd. 2, Frankfurt a.M.: Suhrkamp.

Kulesza, T.; Burnett, M.; Wong, W.-K.; Stumpf, S. (2015): Principles of Explanatory Debugging to Personalize Interactive Machine Learning, in: IUI '15. Proceedings of the 20th International Conference on Intelligent User Interfaces, New York: Association for Computing Machinery, 126–137.

Langer, M.; Oster, D.; Speith, T.; Hermanns, H.; Kästner, L.; Schmidt, E.; Sesing, A.; Baum, K. (2021): What do we want from Explainable Artificial Intelligence (XAI)? A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research, in: *Artificial Intelligence*, 296. [10.1016/j.artint.2021.103473].

Lapuschkin, S.; Wäldchen, S.; Binder, A.; Montavon, G.; Samek, W.; Müller, K.-R. (2019): Unmasking Clever Hans predictors and assessing what machines really learn, in: *Nature Communications*, 10(1), 1–8.

Laux, J.; Wachter, S.; Mittelstadt, B. (2023): Three Pathways for Standardisation and Ethical Disclosure by Default under the European Union Artificial Intelligence Act, in: *Computer Law & Security Review*, 53, 1–11.

Lebovitz, S.; Lifshitz-Assaf, H.; Levina, N. (2022): To Engage or Not to Engage with AI for Critical Judgments. How Professionals Deal with Opacity When Using AI for Medical Diagnosis, in: *Organization Science*, 33(1), 126–148.

Lenk, H. (1982): Zur Sozialphilosophie der Technik, Frankfurt a.M.: Suhrkamp.

Lepri, B.; Oliver, N.; Letouzé, E.; Pentland, A.; Vinck, P. (2018): Fair, Transparent, and Accountable Algorithmic Decision-making Processes, in: *Philosophy & Technology*, 31(4), 611–627.

Lipton, Z.C. (2018): The Mythos of Model Interpretability. In machine learning, the concept of interpretability is both important and slippery, in: *Queue*, 16(3), 31–57.

López-Martínez, F.; Núñez-Valdez, E.R.; García-Díaz, V.; Bursac, Z. (2020): A case study for a big data and machine learning platform to improve medical decision support in population health management, in: *Algorithms*, 13(4), 1–19.

Mahoney, M.S. (2011): Histories of computing, Cambridge (MA): Harvard University Press.

Marabelli, M.; Newell, S.; Page, X. (2018): Algorithmic Decision-Making in the US Healthcare Industry. [https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3262379] (Zugriff: 15.06.2024).

Markus, A.F.; Kors, J.A.; Rijnbeek, P.R. (2021): The role of explainability in creating trustworthy artificial intelligence for health care. A comprehensive survey of the terminology, design choices, and evaluation strategies, in: *Journal of biomedical informatics*, 113, 1–11.

McNamara, A.; Smith, J.; Murphy-Hill, E. (2018): Does ACM's code of ethics change ethical decision making in software development?, in: ESEC/FSE 2018. Proceedings of the 2018 26th ACM joint meeting on European software engineering conference and symposium on the foundations of software engineering, New York: Association for Computing Machinery, 729–733.

Meske, C.; Abedin, B.; Klier, M.; Rabhi, F. (2022): Explainable and responsible artificial intelligence, in: *Electronic Markets*, 32(4), 2103–2106.

Metcalf, J.; Moss, E.; boyd, d. (2019): Owning Ethics. Corporate logics, Silicon Valley, and the Institutionalization of Ethics, in: *Social Research. An International Quarterly*, 86(2), 449–476.

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

Miller, T.; Hoffman, R.R.; Amir, O.; Holzinger, A. (2022): Special issue on Explainable Artificial Intelligence (XAI), in: *Artificial Intelligence*, 307. [10.1016/j.artint.2022.103705].

Mittelstadt, B.D. (2019): Principles alone cannot guarantee ethical AI, in: *Nature Machine Intelligence*, 1, 501–507.

Mittelstadt, B.; Russell, C.; Wachter, S. (2019): Explaining Explanations in AI, in: FAT* '19. Proceedings of the Conference on Fairness, Accountability, and Transparency, New York: Association for Computing Machinery, 279–288.

Mohammed, S.; Brandner, L.T.; Burtscher, F.; Hallensleben, S.; Harmouch, H.; Hauschke, A.; Heesen, J.; Hildebrandt, S.; Hirsbrunner, S.D.; Keselj, J. et al. (2024): A Data Quality Glossary. [https://zenodo.org/records/10474880] (Zugriff: 14.06.2024).

Morley, J.; Floridi, L.; Kinsey, L.; Elhalal, A. (2020): From What to How. An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices, in: *Science and Engineering Ethics*, 26, 2141–2168.

Mueller, S.T.; Hoffman, R.R.; Clancey, W.J.; Emrey, A.; Klein, G. (2019): Explanation in Human-AI Systems. A Literature Meta-Review, Synopsis of Key Ideas and Pu-

blications, and Bibliography for Explainable AI. [https://apps.dtic.mil/sti/citations/AD1073994] (Zugriff: 17.06.2024).

Munn, L. (2022): The uselessness of AI ethics, in: *AI and Ethics*, 3, 869–877.

Nannini, L.; Balayn, A.; Smith, A.L. (2023): Explainability in AI Policies. A Critical Review of Communications, Reports, Regulations, and Standards in the EU, US, and UK, in: FAccT '23. Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, New York: Association for Computing Machinery, 1198–1212.

Norberg, A. (1996): Changing computing. The computing community and DARPA, in: *IEEE Annals of the History of Computing*, 18(2), 40–53.

O'Neil, C. (2016): Weapons of math destruction. How big data increases inequality and threatens democracy, New York: Crown.

Open AI. (2023): How should AI systems behave, and who should decide? [https://openai.com/index/how-should-ai-systems-behave/] (Zugriff: 17.06.2024).

OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S. et al. (2024): GPT-4 Technical Report, in: arXiv. [https://arxiv.org/abs/2303.08774] (Zugriff: 17.06.2024).

Panigutti, C.; Hamon, R.; Hupont, I.; Fernandez Llorca, D.; Fano Yela, D.; Junklewitz, H.; Scalzo, S.; Mazzini, G.; Sanchez, I.; Soler Garrido, J. et al. (2023): The role of explainable AI in the context of the AI Act, in: FAccT '23. Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, New York: Association for Computing Machinery, 1139–1150.

Perrigo, B. (2023): Exclusive. OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic, in: *Time Magazine*, 18.01.2023. [https://time.com/6247678/openai-chatgpt-kenya-workers/] (Zugriff: 10.06.2024).

Petrick, E.R. (2020): Building the Black Box. Cyberneticians and Complex Systems, in: *Science, Technology, & Human Values*, 45(4), 575–595.

Pfeiffer, J.; Gutschow, J.; Haas, C.; Möslein, F.; Maspfuhl, O.; Borgers, F.; Alpsancar, S. (2023): Algorithmic Fairness in AI, in: Business & Information Systems Engineering, 65, 209–222.

Popescu, A.-I. (2016): In brief. Pros and Cons of corporate codes of conduct. *Journal of Public Administration, Finance and Law*, 9(9), 125–130.

Ras, G.; van Gerven, M.; Haselager, P. (2018): Explanation methods in deep learning. Users, values, concerns and challenge, in: Escalante, H.J.; Escalera, S.; Guyon, I.; Baró, X.; Güçlütürk, Y.; Güçlü, U.; van Gerven, M. (Hg.), Explainable and Interpretable Models in Computer Vision and Machine Learning, Cham: Springer, 19–36.

Rességuier, A.; Rodrigues, R. (2020): AI ethics should not remain toothless! A call to bring back the teeth of ethics, in: *Big Data & Society*, 7(2), 1–5.

Ribeiro, M.T.; Singh, S.; Guestrin, C. (2016): »Why Should I Trust You?«. Explaining the Predictions of Any Classifier, in: KDD '16. Proceedings of the 22nd ACM

SIGKDD International Conference on Knowledge Discovery and Data Mining, New York: Association for Computing Machinery, 1135–1144.

Ribera, M.; Lapedriza García, À. (2019): Can we do better explanations? A proposal of user-centered explainable AI, in: IUI Workshops 2. [https://api.semanticscholar.org/CorpusID:84832474] (Zugriff: 17.06.2024).

Roberts, H.; Cowls, J.; Morley, J.; Taddeo, M.; Wang, V.; Floridi, L. (2021): The Chinese approach to artificial intelligence. An analysis of policy, ethics, and regulation, in: *AI & society*, 36, 59–77.

Robles Carrillo, M. (2020): Artificial intelligence. From ethics to law, in: *Telecommunications Policy*, 44(6), 1–16.

Rohlfing, K.J.; Leonardi, G.; Nomikou, I.; Rączaszek-Leonardi, J.; Hüllermeier, E. (2020): Multimodal Turn-Taking. Motivations, Methodological Challenges, and Novel Approaches, in: *IEEE Transactions on Cognitive and Developmental Systems*, 13(2), 260–271.

Rohlfing, K.J.; Cimiano, P.; Scharlau, I.; Matzner, T.; Buhl, H.M.; Buschmeier, H.; Esposito, E.; Grimminger, A.; Hammer, B.; Häb-Umbach, R. et al. (2021): Explanation as a Social Practice. Toward a Conceptual Framework for the Social Design of AI Systems, in: *IEEE Transactions on Cognitive and Developmental Systems*, 13(3), 717–728.

Rudin, C. (2019): Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, in: *Nature Machine Intelligence*, 1(5), 206–215.

Rudin, C.; Wang, C.; Coker, B. (2020): The Age of Secrecy and Unfairness in Recidivism Prediction, in: *Harvard Data Science Review*, 2(1), 1–53.

Russell, S.; Norvig, P. (4. Auflage 2021): Artificial Intelligence. A modern approach, London: Pearson.

Samek, W.; Müller, K.-R. (2019): Towards Explainable Artificial Intelligence, in: Samek, W.; Montavon, G.; Vedaldi, A.; Hansen, L.K.; Müller, K.-R. (Hg.), Explainable AI. Interpreting, Explaining and Visualizing Deep Learning, Cham: Springer, 5–22.

Samek, W.; Wiegand, T.; Müller, K.-R. (2017): Explainable artificial intelligence. Understanding, visualizing and interpreting deep learning models, in: arXiv. [https://arxiv.org/abs/1708.08296] (Zugriff: 11.06.2024).

Schwartz, M.S. (2004): Effective corporate codes of ethics. Perceptions of code users, in: *Journal of business ethics*, 55, 321–341.

Shahid, F.; Vashistha, A. (2023): Decolonizing Content Moderation. Does Uniform Global Community Standard Resemble Utopian Equality or Western Power Hegemony?, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 1–18.

Siapera, E. (2022): AI content moderation, racism and (de) coloniality, in: *International journal of bullying prevention*, 4(1), 55–65.

Simon, J. (2017): Value Sensitive Design and Responsible Research and Innovation, in: Hansson, S.O. (Hg.), The Ethics of Technology. Methods and Approaches, London/New York: Rowman & Littlefield, 219–236.

Simonite, T. (2018): When It Comes to Gorillas, Google Photos Remains Blind, in: *Wired*, 11.01.2018. [https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/] (Zugriff: 28.12.2023).

Sokol, K.; Flach, P. (2020): One Explanation Does Not Fit All, in: *KI-Künstliche Intelligenz*, 34(2), 235–250.

Sovrano, F.; Sapienza, S.; Palmirani, M.; Vitali, F. (2022): Metrics, explainability and the European AI act proposal, in: *J*, 5(1), 126–138.

Stamboliev, E. (2023): Proposing a Postcritical AI Literacy. Why We Should Worry Less about Algorithmic Transparency and More about Citizen Empowerment, in: *Media Theory*, 7(1), 202–232.

Starke, C.; Baleis, J.; Keller, B.; Marcinkowski, F. (2022): Fairness perceptions of algorithmic decision-making. A systematic review of the empirical literature, in: *Big Data & Society*, 9(2), 1–16.

Strohmeier, S. (2020): Algorithmic decision making in HRM, in: *Encyclopedia of electronic HRM*, 1, 54–59.

Surden, H. (2020): Ethics of AI Law. Basic Questions, in: Dubber, M.D.; Pasquale, F.; Das, S. (Hg.), The Oxford Handbook of Ethics of AI, New York: Oxford University Press, 719–736.

Tsamados, A.; Aggarwal, N.; Cowls, J.; Morley, J.; Roberts, H.; Taddeo, M.; Floridi, L. (2022): The ethics of algorithms. Key problems and solutions, in: *Ai & Society*, 37(1), 215–230.

Tworek, H. (2019): Social Media Councils, in: Owen, T.; Docquir, P.F.; Donovan J.; Etlinger, S.; Fay, R.; Girard, M.; Gorwa, R.; Kimmelman, G.; Klonick, K.; McDonald, S.M. et al. (Hg.), Models for Platform Governance. A CIGI Essay Series, Waterloo: Centre for International Governance, 97–102.

UNESCO (2021): Recommendation on the Ethics of Artificial Intelligence. [https://unesdoc.unesco.org/ark:/48223/pf0000381137] (Zugriff: 7.06.2023).

van de Poel, I. (2016): An Ethical Framework for Evaluating Experimental Technology, in: *Science and Engineering Ethics*, 22(3), 667–686.

van der Hoven, J.; Manders-Huits, N. (2020): Value-sensitive design, in: Olsen, J.K.B.; Pedersen, S.A.; Hendricks, V.F. (Hg.), The Ethics of Information Technologies, London: Routledge, 329–332.

Vermaas, P.E. (2010): Focussing Philosophy of Engineering. Analyses of Technical Functions and Beyond, in: van de Poel, I.; Goldberg D. (Hg.), Philosophy and Engineering. An Emerging Agenda, Dordrecht u.a.: Springer Netherlands, 61–73.

Vilone, G.; Longo, L. (2021): Notions of explainability and evaluation approaches for explainable artificial intelligence, in: *Information Fusion*, 76, 89–106.

Vincent, J. (2018): Google ›fixed‹ its racist algorithm by removing gorillas from its image-labeling tech, in: *The Verge*, 12.01.2018. [https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognitionalgorithm-ai] (Zugriff: 28.12.2023).

Vogelmann, F. (2019): Transparency's Trap. Problems of an Unquestioned Norm, in: Berger, S.; Owetschkin, D. (Hg.), Contested Transparencies, Social Movements and the Public Sphere. Multi-Disciplinary Perspectives, Cham: Springer, 35–54.

Wagner, B. (2018): Ethics As An Escape From Regulation. From ›Ethics-Washing‹ To Ethics-Shopping?, in: Bayamlioglu, E.; Baraliuc, I.; Janssens, L.A.W.; Hildebrandt, M. (Hg.), BEING PROFILED:COGITAS ERGO SUM. 10 Years of Profiling the European Citizen, Amsterdam: Amsterdam University Press, 84–89.

Walmsley, J. (2021): Artificial intelligence and the value of transparency, in: *AI & Society*, 36(2), 585–595.

Wang, D.; Yang, Q.; Abdul, A.; Lim, B.Y. (2019): Designing Theory-Driven User-Centric Explainable AI, in: CHI '19. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, New York: Association for Computing Machinery, 1–15.

Weber, M. (1992): Wissenschaft als Beruf, in: Ders., Wissenschaft als Beruf 1917/1919. Politik als Beruf 1919 [Studienausgabe der Max Weber-Gesamtausgabe, Band I/17], Tübingen: Mohr Siebeck, 71–112.

Xie, Y.; Yi, J.; Shao, J.; Curl, J.; Lyu, L.; Chen, Q.; Xie, X.; Wu, F. (2023): Defending ChatGPT against jailbreak attack via self-reminders, in: *Nature Machine Intelligence*, 5(12), 1486–1496.

Yeung, K.; Howes, A.; Pogrebna, G. (2020): AI Governance by Human Rights–Centered Design, Deliberation, and Oversight. An End to Ethics Washing, in: Dubber, M.D.; Pasquale, F.; Das, S. (Hg.), The Oxford Handbook of Ethics of AI, New York: Oxford University Press, 76–106.

Zarsky, T. (2016): The trouble with algorithmic decisions. An analytic road map to examine efficiency and fairness in automated and opaque decision making, in: *Science, Technology, & Human Values*, 41(1), 118–132.