Verantwortung in Zeiten ›künstlicher Intelligenz‹

Eine Problemexposition am Beispiel medizinischer Diagnostik*

Susanne Hahn

Abstract: *The ascription of responsibility is a social practice with a high significance for human coexistence. The challenge posed to this practice by the use of artificial intelligence, for example in medical diagnostics, needs to be specified in more detail. An examination of the classic concept of responsibility makes it possible to identify ascription gaps. Assuming that an important function of the ascription of responsibility, namely the guidance of action to avoid harm, should be preserved, it is worth taking a look at the historical handling of ascription gaps using the sketch of the regulation of steam boilers in the 19th century. Some epistemological considerations on the predominantly prognostic function of algorithms in diagnostics as well as existing proposals for the certification of artificial intelligence serve to work out questions that are decisive for the assignment of responsibilities.*

Keywords: *responsibility; responsibility gap; role obligations; demand for transparency; social practice*

1. Das Ausgangsproblem: Herausforderung für die Verantwortungszuschreibung durch den Einsatz ›künstlicher Intelligenz‹

Die Zuschreibung von Verantwortung ist eine soziale Praxis mit einem hohen Stellenwert für das menschliche Zusammenleben. Die Zuordnung von Handlungsfolgen, insbesondere solchen unerwünschter Art, zu Akteuren als ihren Urhebern, hat

* Die Arbeit an diesem Beitrag wurde gefördert durch das Center for Advanced Internet Studies (CAIS) Bochum und die Stiftung MERCATOR. – Seit der Arbeit an diesem Thema im Rahmen eines Fellowships 2019 haben sich sowohl in der Sache selbst (Stichwort ChatGPT) als auch in der Reflexion und Regulierung von KI (Stichwort AI Act) viele weitere Neuerungen oder Beschleunigungen ergeben. An der Grundstruktur von Handeln, Technikeinsatz und Verantwortungszuschreibung hat sich dadurch nichts verändert. Der vorliegende Aufsatz versucht sich an einer Analyse dieser Grundstruktur.

sowohl präventive als auch wiedergutmachende Funktionen. Diese soziale Praxis, die auch rechtlich normiert ist, wird durch den Einsatz künstlicher Intelligenz herausgefordert, so beispielsweise in der medizinischen Diagnostik. Hier, wie in vielen anderen Bereichen, ist in den letzten Jahren eine stark anwachsende Entwicklung und Anwendung des maschinellen Lernens zu verzeichnen, insbesondere in der Bilderkennung, aber auch in der Verarbeitung anderer Merkmale zu Diagnosen bzw. Voraussagen.

So vielversprechend der Einsatz von Algorithmen auch sein mag, verhindert er jedoch nicht, dass Personen Schäden durch Fehldiagnosen erleiden. Dies ist der Ansatzpunkt für die Frage nach der Verantwortung: Wem ist Verantwortung zuzuschreiben, wenn ein Patient aufgrund einer Einschätzung durch einen Algorithmus beispielsweise nicht zu einer Präventivuntersuchung überwiesen worden war und er einen Herzinfarkt erlitten hat? Wer trägt Verantwortung, wenn eine Patientin nach einer Einschätzung durch eine maschinelle Bilderkennung als krebserkrankt gilt und zu weiteren Eingriffen wie Gewebeentnahmen geschickt wird, ohne dass sich ein Tumor bzw. eine Tumorvorstufe findet. Wer ist verantwortlich, wenn eine bösartige Veränderung nicht entdeckt wurde? Und – um neuere Entwicklungen zur Früherkennung von Alzheimer zu nennen – wer trägt Verantwortung, wenn jemand mit 55 Jahren unter Einsatz eines Algorithmus die Diagnose erhält, in der Zukunft mittelfristig wahrscheinlich an Alzheimer zu erkranken, er aber mit 89 an einer Lungenentzündung bei guter geistiger Gesundheit stirbt?

Durch die Einschaltung einer Technik – hier der Einsatz von Algorithmen zur Klassifikation bzw. Voraussage von medizinisch relevanten Zuständen – entstehen *Zurechnungslücken*: Es ist nicht mehr klar, welchem Akteur ein eintretender Schaden zuzurechnen ist, wodurch der Adressat für eine Kompensation und eine mögliche Sanktion ebenfalls nicht mehr identifizierbar ist. Durch diesen Umstand ist eine für das menschliche Zusammenleben wichtige soziale Praxis herausgefordert und die Frage, ob der Einsatz dieser Techniken entsprechend reguliert werden soll, auf der Tagesordnung.[1] Der vorliegende Aufsatz dient nicht dazu, eine affirmative oder negative Antwort auf diese Frage zu liefern. Es geht vielmehr ausschließlich darum, dieses Problem so zu analysieren, dass die aufgeworfene Frage spezifiziert werden kann und deutlich wird, dass zu ihrer adäquaten Behandlung zum einen weitere Fragen nach der technischen Handhabbarkeit von algorithmischen Voraussagemodellen zu beantworten sind und zum anderen Abwägungen hinsichtlich der Chancen und Risiken der Technik notwendig sind.

Diese Problemexposition soll vermitteln, dass zur Bearbeitung dieser Aufgaben die Expertise aus Mathematik, Statistik, Informatik, Wissenschaftsphilosophie, Moralphilosophie, Rechtswissenschaft und Medizin erforderlich ist. Das hierzu

1 Auf vielen Ebenen und in vielen Gremien wurden und werden bereits Richtlinien zum Umgang mit künstlicher Intelligenz entwickelt. Für einen Überblick vgl. Hagendorff 2020.

entwickelte Szenario umfasst fünf Schritte: Zunächst sind einige Klärungen zum Verständnis künstlicher Intelligenz zu leisten und zu erläutern, welche Formen in der medizinischen Diagnostik, die hier als Beispielfeld dienen soll, zum Einsatz gelangen könnten. Sodann ist das klassische Konzept von Verantwortung zu erläutern, um darzulegen, in welcher Weise diese Form der Verantwortungszuschreibung durch den Einsatz künstlicher Intelligenz herausgefordert wird. Unterstellend, dass eine wichtige Funktion der Verantwortungszuschreibung, nämlich die Handlungssteuerung zur Vermeidung von Schäden, erhalten bleiben soll, bietet sich ein Blick in den historischen Umgang mit Zurechnungslücken an. Eine Skizze der Regulierung von Dampfkesseln im 19. Jh. liefert strukturelle Anknüpfungspunkte. Einige erkenntnisphilosophische Überlegungen zur vorwiegend prognostischen Funktion von Algorithmen in der Diagnostik sowie vorliegende Vorschläge zur Zertifizierung von künstlicher Intelligenz dienen dazu, Fragen herauszuarbeiten, die für die Zuordnung von Verantwortlichkeiten entscheidend sind.

Angewandte Philosophie zeichnet sich dadurch aus, dass man verallgemeinernde Überlegungen auf spezifische Handlungsfelder bezieht oder auch umgekehrt eben solche Verallgemeinerungen aus diesen gewinnt. Diese Handlungsfelder sind durch einen eigenen Fundus an Wissen und Verfahren charakterisiert. Das gilt unabhängig davon, ob man sich mit Organtransplantation, Gentherapie, Umwelttechnik oder – im betrachteten Fall – mit Anwendungen maschinellen Lernens beschäftigt (vgl. Bayertz 1991: 28ff.). In allen Fällen stellt sich die Frage, wie tief und umfassend man sich mit dem jeweiligen Gegenstandsbereich auseinandersetzen muss, um begriffliche Sortierungen vornehmen zu können, die wiederum Voraussetzungen für normative Einschätzungen darstellen.

Die Antwort hängt davon ab, welche – beispielsweise – ethischen Prinzipien und Normen man potenziell durch eine Handlungsweise betroffen sieht. Wenn man z. B. vermutet, dass die Steuerbarkeit und Transparenz von Handlungen durch den Einsatz künstlicher Intelligenz beeinträchtigt werden und diese Eigenschaften für die moralphilosophische Betrachtung relevant sind, dann sollte man die zugrundeliegenden Verfahren so weit erfassen, dass diese Merkmale deutlich werden.

Damit ergibt sich der erste Schritt zur Realisierung des hier verfolgten Projekts, die Grundlage für die Formulierung handhabbarer Szenarien und Fragen zu liefern: Es gilt also zunächst, unterschiedliche Formen maschinellen Lernens so aufzubereiten, dass Fragen zu ihrem gerechtfertigten Einsatz bearbeitbar werden.[2] Die relative Explizitheit dieses Vorgehens dient zum einen dazu, Unterschiede in den

2 Die Nicht-Expertin auf diesem Gebiet ist auf Darstellungen angewiesen, die sich an interessierte Laien richten. Die hier vorgelegte Skizze orientiert sich zunächst – auch in den Abbildungen und Beispielen – an dem für Einsteiger sehr nützlichen Buch von Steven Finlay (Finlay 2017) und zieht daneben Bart Baesens (Baesens 2014) heran.

Spielarten künstlicher Intelligenz zu verdeutlichen und in ihrer Relevanz für die Verantwortungsthematik einzuordnen. Zum anderen sollen die Darstellungen Anhaltspunkte für das Verständnis von Schlagworten wie Transparenz, Erklärbarkeit, Nachvollziehbarkeit etc. liefern.

2. Formen maschinellen Lernens als derzeit realisierte Form ›künstlicher Intelligenz‹

Mit dem Begriff künstliche Intelligenz, abgekürzt KI (»artificial intelligence«, AI), ist für einige Menschen die Erwartung verbunden, dass Menschen, als Träger »natürlicher« Intelligenz (jetzt/in Kürze/in absehbarer Zeit/langfristig) durch Maschinen ersetzbar werden. Diesem *starken* Verständnis künstlicher Intelligenz im Sinne von Menschenähnlichkeit (hier allerdings oft nur in Bezug auf die erwünschten Eigenschaften) steht ein *schwaches* Verständnis gegenüber, das auf bestimmte, vor allem auf Kalkulationen bezogene, Fertigkeiten abstellt.[3] Im Zentrum dieses Verständnisses stehen Algorithmen maschinellen Lernens.[4] Wenn im Folgenden von künstlicher Intelligenz die Rede ist, dann ist damit dieser problembezogene Einsatz maschinellen Lernens gemeint. Diese auf statistischen Verfahren basierende Technologie ist durch enorm gestiegene Rechnerkapazitäten auf der einen Seite und zugleich stark wachsender Verfügbarkeit von Daten in den letzten Jahren zu einem dominanten Zweig in der Informatik und Datenwissenschaft geworden (vgl. Engemann 2018: 254; Lepri/Oliver/Letouzé/Pentland/Vinck 2018: 612). – Die Folgefrage lautet: Was ist maschinelles Lernen?

> »Machine learning is the use of mathematical procedures (algorithms) to analyze data. The aim is to discover useful patterns (relationships or correlations) between different items of data. Once the relationships have been identified, these can be used to make inferences about the behavior of new cases when they present themselves.« (Finlay 2017: 5)

Auch wenn die Erfolge in der Anwendung maschinellen Lernens häufig eine mystische Aura autonom agierender, aber nicht identifizierbarer Akteure hervorrufen, ist – wie in dem obigen Zitat ablesbar – festzuhalten, dass diese Verfahren auf der *An-*

3 Unterschiedliche Bedeutungen von künstlicher Intelligenz werden z.B. dargelegt bzw. erörtert in Bringsjord/Govindarajulu 2020; Mainzer 2016.

4 Zur Charakterisierung von Algorithmen sowie zur Unterscheidung regelbasierter, »klassisch« programmierter Algorithmen und Algorithmen maschinellen Lernens vgl. Fry 2018.

wendung mathematischer und statistischer Verfahren beruhen.[5] Die Zielsetzung besteht darin, in den vorliegenden Daten, d.h. in den bisherigen Verläufen, Merkmalskonstellationen zu finden, die von den zur Mustererkennung betrachteten Fällen auf neue Fälle übertragen werden können. In übertragener Redeweise handelt es sich um Lernen aus Erfahrung: »In den-und-den Fällen hat die-und-die Merkmalskonstellation zum Verlauf A geführt. Dieser Fall weist eine große Ähnlichkeit zu dieser Merkmalskonstellation auf. Also ist Verlauf A wahrscheinlich.« Eine typische Anwendung ist die Erstellung von Prognosen, ob eine Person einen Kredit zurückzahlen wird. Die ermittelten Beziehungen zwischen Merkmalen sind allerdings, wie in statistischen Verfahren üblich, Korrelationen; über Kausalverhältnisse ist damit noch nichts gesagt. Ein klassisches illustratives Beispiel liefert die Merkmalskonstellation zwischen der Anzeige eines Barometers, dem atmosphärischen Druck und den daraus abgeleiteten Prognosen für die Wetterlage. Die Korrelationen, die in »beiden Richtungen« zwischen den Barometeranzeigen und dem atmosphärischen Druck bestehen, sind nur in einer Richtung auch Kausalbeziehungen: Nicht die Veränderung der Barometeranzeige verursacht die Druckveränderung, sondern die Druckveränderung verursacht umgekehrt die Änderung der Anzeige des Barometers. Werden reine Korrelationen betrachtet, ist weder etwas über den tatsächlichen Einfluss einer Größe auf künftige Entwicklungen gesagt, noch über das Zustandekommen von Mustern.[6]

Der warnende Hinweis auf die lediglich ermittelten Korrelationen deutet bereits an, dass im Hintergrund des maschinellen Lernens verschiedene erkenntnis- und wissenschaftsphilosophische Probleme stehen. Neben dem genannten Problem von Korrelation vs. Kausalität ist dies allgemein das Problem der Induktion: Unter welchen Bedingungen ist man berechtigt, von einer endlichen Zahl beobachteter Fälle mit bestimmten Merkmalskonstellationen auf eine allgemeine Aussage über Typen dieser Fälle – und damit auch auf bislang unbeobachtete Fälle – zu schließen (vgl. Abschnitt 6.1)?

Eine weitere für normative Betrachtungen wichtige Eigenschaft des Einsatzes maschinellen Lernens sind die enthaltenen Verfahrensschritte. Üblicherweise sind dies die folgenden (vgl. Finlay 2017: 48ff.):

---

5 »[…] it [sc. AI] is only ›intelligent‹ in the narrowest sense of the word. It would probably be more useful to think of what we've been through as a revolution in computational statistics than a revolution in intelligence.« (Fry 2018: 14)

6 Die Nicht-Berücksichtigung von Kausalitätsbeziehungen hat auch damit zu tun, dass die mathematische Darstellung von Merkmalsbeziehungen eine Unterscheidung zwischen bloßer Korrelation und Kausalität erschwert. Diesem Defizit will Judea Pearl abhelfen, indem er versucht zu zeigen, wie man Kausalbeziehungen formal-mathematisch darstellen kann (Pearl 2018).

- Zusammenstellung eines Dateninputs
- Aufbereitung der Daten
- Entwicklung der Voraussagemodelle
- Formulierung von Entscheidungsregeln bzw. von Wenn-Dann-Verknüpfungen[7]
- resultierende Maßnahme.

Für die spätere Diskussion ist es wichtig zu bemerken, dass ein großer Teil der analytischen Arbeit in der inhaltlichen Durchdringung einer Fragestellung, welche Merkmalsbeziehungen es zwischen unterschiedlichen Zuständen oder Ereignissen gibt, bereits in Prozessen liegen, die *vor* der Bereitstellung des Dateninputs und der Aufbereitung der Daten liegen. Daneben wird mit dieser Identifizierung von Verfahrensschritten deutlich, dass notwendig Entscheidungsregeln formuliert werden müssen. Dieser Verfahrensschritt liegt wiederum nicht in unbeeinflussbaren technischen Prozessen, sondern ist Gegenstand einer menschlichen Entscheidung (vgl. weiter unten zur Bildung von Risikoklassen und Bestimmung eines Schwellwertes). Der hier zunächst interessierende Einsatz des eigentlichen maschinellen Lernverfahrens besteht lediglich im Schritt zur Entwicklung des Voraussagemodells.

> »A predictive model (or just model going forward) is the output generated by the machine learning process. The model captures the relationships (patterns) between which have been uncovered by the analytics process. Once a model has been created it can be used to generate new predictions. Organizations then use the model's predictions to decide what to do or how to treat people. *So machine learning is a process, and a predictive model is the end product of that process*.« (Finlay 2017: 38f.; Hervorhebung – SH)

Für die Fragen der Verantwortungszuschreibung ist der Umstand bedeutsam, dass es – anders als durch die Bezugnahme auf neuronale Netze in öffentlichen Debatten nahegelegt – *verschiedene Techniken maschinellen Lernens* gibt, um ein Voraussagemodell zu gewinnen. Gemeinsam ist allen Modellen, dass die generierte Voraussage durch eine Zahl repräsentiert wird, die die Wahrscheinlichkeit angibt, mit der das in Frage stehende Verhalten oder Ereignis eintritt. Die Verfahren maschinellen Lernens unterscheiden sich jedoch hinsichtlich ihrer Nachvollziehbarkeit.[8]

---

7 Zur Frage, ob die Rede von »Entscheidungen« durch Algorithmen berechtigt und adäquat ist, vgl. Hahn 2024.

8 Die »Opakheit« maschineller Mustererkennung ist nicht nur Gegenstand der Suche nach technischen Lösungen, sondern auch der kritischen Reflexion. – Bezogen auf Arten der Opakheit, die unterschieden werden, ist hier im Folgenden die Art gemeint, die sich aus dem maschinellen Lernverfahren selbst ergibt, nicht etwa aus eigentumsrechtlichen Gegebenheiten (vgl. Burrell 2016; Durán/Jongsma 2021).

Dieser Unterschied lässt sich an einem – fiktiven – Beispiel aus der medizinischen Diagnostik illustrieren, das der erwähnten Arbeit zur Darstellung maschinellen Lernens entnommen ist (vgl. Finlay 2017). Die ausführliche Darstellung dieses Beispiels dient dem oben bereits erwähnten Ziel, nachvollziehbar darzustellen, an welchen Stellen die etablierte Praxis der Verantwortungszuschreibung herausgefordert wird und Anschauungsmaterial für die Verwendung der »Transparenzbegrifflichkeit« zu liefern. Der Dateninput besteht im Beispiel aus zufällig ausgewählten Berichten von 500.000 Patienten, die noch keine Anzeichen einer Herzkrankheit aufweisen. Die Daten umfassen Angaben zu Alter, Geschlecht, Vorerkrankungen, Blutdruck, Body-Mass-Index, Alkoholkonsum, Rauchen, Gewicht sowie zum Einkommen.

Diese Daten werden mit dem weiteren Verlauf in den darauffolgenden fünf Jahren konfrontiert: Wer von diesen Personen entwickelt eine Herzkrankheit und wer nicht? In der Sprache der Voraussagemodelle hat man an dieser Stelle zwei Sorten von Daten: *Beobachtungsdaten* – das sind in diesem Fall die Daten aus den medizinischen Berichten und *Ergebnisdaten* – in diesem Fall die Dokumentation des Gesundheitsverlaufs über fünf Jahre, konzentriert darauf, ob diese Personen eine Herzkrankheit entwickeln oder nicht.

Zusammen ergeben diese Daten das sogenannte *Entwicklungssample*. Die Zusammenstellung des Entwicklungssamples enthält bereits die ersten beiden genannten Schritte bei der Anwendung von Verfahren maschinellen Lernens, nämlich die Bereitstellung des Dateninputs und die Aufbereitung der Daten. Auf dieser sollen aussagekräftige Parameter ermittelt werden, die es erlauben, bei neuen Fällen auf die Entwicklung oder Nicht-Entwicklung einer Herzkrankheit zu schließen. In diesem Beispiel soll diese Einschätzung als Entscheidungsbasis für die Einladung zu einer umfassenderen Herz-Kreislauf-Untersuchung dienen (im Fünfer-Schritt der Anwendung von Verfahren maschinellen Lernens ist dies der letzte Schritt, d.h. die resultierende Maßnahme).

Im Beispiel stellt man fest, dass von den 500.000 Personen innerhalb von fünf Jahren 30.000 eine Herzkrankheit entwickeln. An dieser Stelle kommen die Verfahren maschinellen Lernens zum Einsatz. Gesucht werden Algorithmen für ein Voraussagemodell, das die Beobachtungsdaten mit den Ergebnisdaten korreliert. Welche Parameter sind in welchem Maße ausschlaggebend dafür, eine Herzkrankheit zu entwickeln oder eben nicht, welche *Muster* ergeben sich? Die Gegenüberstellung von zwei Voraussagemodellen soll dazu dienen, die Problematik der Nachvollziehbarkeit und Debattierbarkeit und letztlich der Verantwortungszuschreibung zu erläutern.

Das erste Voraussagemodell ist eine *Scorecard*, ein lineares Modell, das auf dem statistischen Verfahren der logistischen Regression beruht. Das andere Verfahren ist ein sogenanntes *künstliches neuronales Netz*.

Abbildung 1

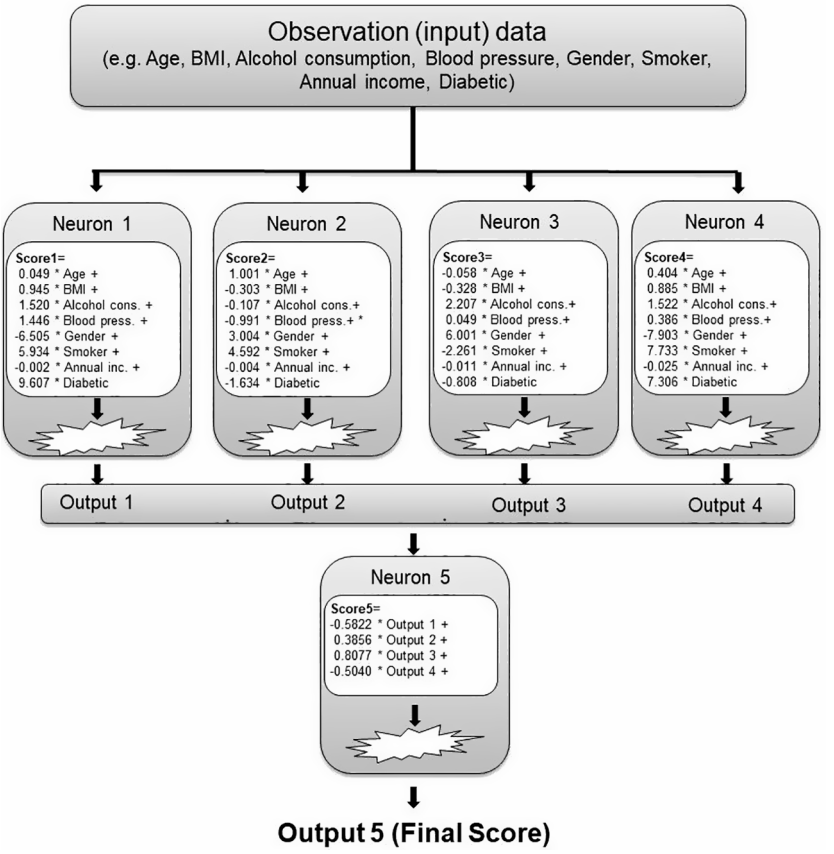
Starting score (constant)	350		
Age (years)		**Gross annual income ($)**	
<23	-57	< $22,000	11
23 - 32	-26	$22,001 - $38,000	6
33 - 41	0	$38,001 - $60,000	0
42 - 48	7	$60,001 - $94,000	-3
49 - 57	15	$94,001 - $144,000	-5
58 - 64	24	>$144,000	-6
65 - 71	31		
>71	65	**Smoker ?**	
		Yes	37
BMI (weight in kg / {height in metres}^{2})		No	0
<19	2		
19 - 26	0	**Diabetic ?**	
27 - 29	8	Yes	21
30 - 32	14	No	0
>32	29		
		Cholesterol level (mg per decilitre of blood)	
Gender		Low (< 160 mg)	-2
Male	2	Normal (160 - 200 mg)	0
Female	-4	High (201 - 240 mg)	19
		Very high (>240 mg)	32
Alcohol consumption (units/week)			
0	4	**Blood pressure**	
1 - 12	0	Low (below 90/60)	3
13 - 24	5	Average (between 90/60 and 140/90)	0
25 - 48	10	High (above 140/90)	36

Quelle: Finlay 2017: 33

Beide Verfahren nutzen dieselben Kriterien, von Alter bis Einkommen. Die Angaben, welche Ausprägung eines Attributs ein Patient hat, z.B. ein Alter von 53 Jahren und Blutdruckwerte von 150 zu 90 liefern den Input für die Verfahren. Bei einem neuronalen Netz, wie es oben abgebildet ist, gehen diese Daten mehrfach in die Berechnung ein. Die Werte der neun Attribute beliefern vier Neuronen bzw. Knoten, in denen die Werte mit einer Gewichtung versehen und zu einem Output-Wert addiert werden. Die Output-Werte von diesen liefern den Input für – in diesem Fall – ein weiteres Neuron, in dem diese vier Werte wiederum gewichtet und addiert werden. Der letzte Schritt besteht in einer Transformation dieser Werte auf eine Zahl aus einem bestimmten Intervall, z.B. auf Werte zwischen 0 und 1. Dieser transformierte Wert gibt die Wahrscheinlichkeit dafür an, in den nächsten fünf Jahren an einem Herzleiden zu erkranken. Im sehr einfachen Beispielnetzwerk kommen dabei 36 (4 x 8 + 4 = 36) Gewichtungen zum Einsatz. In künstlichen neuronalen Netzen, wie sie insbesondere bei der Bilderkennung zum Einsatz kommen, geht die Zahl der Gewichtungen in die Tausende.[9]

9 Für grundlegende Informationen zu neuronalen Netzen, Arten des Lernens sowie der Analogie zu biologischen neuronalen Vorgängen vgl. Mainzer 2016: Kap. 7.2.

Abbildung 2



Quelle: Finlay 2017: 52

Bei der Scorecard handelt es sich um das Resultat eines statistischen Standardverfahrens, der logistischen Regression. Im Vergleich mit den künstlichen neuronalen Netzen kann man dieses als ein einzelnes Neuron auffassen.[10] Mit der logistischen Regression wird der Einfluss der hier genannten Merkmale auf das Eintreten eines Herz-Kreislauf-Ereignisses bestimmt. Im Unterschied zu den künstlichen neuronalen Netzen werden die Werte für die Attribute nur einmal zur Bestimmung

10 Baesens setzt der verbreiteten Auffassung, künstliche neuronale Netzwerke seien eine Nachbildung physiologischer Neuronen, eine »realistischere« Perspektive entgegen: »A first perspective on the origin of neural networks states that they are mathematical representations inspired by the functioning of the human brain. Another more realistic perspective sees neural networks as generalizations of existing statistical models.« (Baesens 2014: 48)

dieser Wahrscheinlichkeit herangezogen. Die Gewichtungen sind als solche erkennbar. Auch bei der Übertragung in das Scorecard-Modell sind die Gewichtungen weiterhin nachvollziehbar: So trägt beispielsweise das Geschlecht weniger zur Erstellung der Prognose bei als das Alter der Patienten.[11]

Beide Verfahren sind so zu interpretieren, dass gefragt wird, welche Gewichte für die einzelnen Attribute angenommen werden müssen, um sich den Ergebnisdaten möglichst weit anzunähern. Nachdem in einem *Trainingsdatenset* ein Voraussagemodell entwickelt wurde, wird dieses an einem *Testdatenset*, das verschieden ist vom Trainingsdatenset, erprobt und eventuell adjustiert. Von den erwähnten Schritten in der Anwendung maschinellen Lernens ist der dritte Schritt mit der Formulierung von Voraussagemodellen abgeschlossen. Die Anwendung der Voraussagemodelle auf neue Fälle ergibt Wahrscheinlichkeitsaussagen für das betrachtete Ereignis.

Wenn man diese Voraussagen mit Maßnahmen verknüpfen möchte, sind weitere Schritte erforderlich, die jenseits des maschinellen Lernverfahrens liegen. Im Beispiel ist unterstellt, dass die Zeit, die Ärzten zu großflächigen Präventionsuntersuchungen zur Verfügung steht, begrenzt ist und somit eine Auswahl unter den Patienten zu treffen ist. Hierzu dient zunächst die Bildung von Risikoklassen, um diejenigen zu identifizieren, die am stärksten von einer Präventivuntersuchung profitieren.

Abbildung 3

Group	Score range From	To	Number of people	% of population	Number with heart disease after	% with heart disease after 5 yrs.
1	0	300	55,950	11.19%	40	0.07%
2	301	320	56,606	11.32%	68	0.12%
3	321	340	59,700	11.94%	129	0.22%
4	341	360	58,706	11.74%	216	0.37%
5	361	380	64,429	12.89%	403	0.63%
6	381	400	52,749	10.55%	575	1.09%
7	401	420	34,089	6.82%	600	1.76%
8	421	440	21,107	4.22%	632	2.99%
9	441	460	17,269	3.45%	878	5.09%
10	461	480	23,364	4.67%	2,020	8.65%
11	481	500	17,477	3.50%	2,553	14.61%
12	501	520	13,554	2.71%	3,366	24.84%
13	521	540	7,103	1.42%	3,463	48.76%
14	541	560	8,260	1.65%	6,587	79.74%
15	561	999	9,637	1.93%	8,469	87.88%
Total	**Total**		**500,000**		**30,000**	**6.0%**

Quelle: Finlay 2017: 41

11 Aus der Formel der logistischen Regression lässt sich – auch wenn eine detailliertere Einsicht statistische Kenntnisse erfordert – ablesen, dass die Gewichte für die Merkmale nur einmal in die Berechnung des Wahrscheinlichkeitswertes eingehen (vgl. Baesens 2014: 48).

Hier sind dies beispielhaft 15 Klassen aus den Scorecard-Werten (vgl. Abbildung 3). Von den Personen aus der Risikoklasse 15, dies sind 9637 Personen, entsprechend einem Populationsanteil von 1,93%, sind 87,88% in den nächsten fünf Jahren erkrankt. Wenn man nun dieses Erfahrungswissen auf neue Fälle anwenden möchte, und die Beschränkung annimmt, dass aufgrund endlicher Zeit der Ärzte nur 5% der Population eingeladen werden können, ist zu fragen, welche Risikoklassen zu einer Vorsorgeuntersuchung eingeladen werden sollen und welche nicht.

Die Angehörigen der Risikoklassen 13–15, also alle Personen mit einer Punktzahl größer gleich 521, entsprechen 5% der Population. Schaut man auf die Krankheitsentwicklung dieser Personen in der zweitletzten Spalte, sieht man, dass 18519 davon erkranken – das sind 62% aller 30000 Erkrankten. Indem man 5% der Population einlädt, kann man somit 62% der potentiell Erkrankten identifizieren.

Geht man davon aus, dass diese großen Datenmengen von Computern in kurzer Zeit verarbeitet werden können, dann sieht man zunächst, dass der Einsatz maschinellen Lernens hier sehr effizient ist, d.h. dass viele Personen von präventiven Maßnahmen profitieren, bei zugleich vertretbarem ärztlichem Aufwand.

Dieses Geschehen als *Entscheidungen* eines Algorithmus zu beschreiben, ist jedoch irreführend.[12] *Ex ante* werden Regeln formuliert, wie mit den algorithmisch ermittelten Prognosen zu verfahren ist, z.B. »Alle Patientinnen, die zu den Risikoklassen 13–15 gehören, sollen zu einer umfassenden Vorsorgeuntersuchung eingeladen werden.« Diese Festlegung auf bestimmte Risikoklassen stellt eine *vorgezogene Entscheidung* dar. Der Algorithmus hat hier keine diskretionären Spielräume wie das bei Entscheidungen von Akteuren im engeren Sinne der Fall ist. So könnte ein Arzt von einer Behandlungsleitlinie abweichen, wenn er bei einem Patienten im zeitlichen Verlauf eine ungewöhnliche Steigerung bei einem Merkmal sieht. Er könnte den Patienten zu einer Präventivuntersuchung überweisen, obwohl er noch unter dem dafür vorgesehenen Wert liegt. Bei einer automatisierten Maßnahme hingegen lautet die Vorgabe, dass ab einer Wahrscheinlichkeit von 49 Prozent, in den nächsten fünf Jahren ein Herzkreislaufereignis zu erleiden, eine Patientin zu einer umfassenden fachärztlichen Untersuchung einzuladen ist. Die Verknüpfung zwischen dem Resultat maschinellen Lernens und einer Maßnahme wird aufgrund *von ex ante* angestellten Überlegungen unter bestimmten Zielen *gesetzt*.

Das Beispiel demonstriert die Leistungsfähigkeit der Verfahren maschineller Mustererkennung. Wenn man solche Verfahren jedoch umfassend einschätzen will, sind nicht nur die dadurch geschaffenen neuen Handlungsmöglichkeiten und ihre

12 Das Stichwort lautet »Automated Decision Making« oder »Automatisiertes Entscheiden«. Eines von vielen Beispielen ist die Kapitelüberschrift »Die nächste Stufe der Automatisierung: Maschinen treffen Entscheidungen« in dem informativen Bändchen von Ramge (Ramge 2018: 13). – Zu einer Kritik an dieser Übertragung von Ausdrücken aus den Kontexten menschlichen Handelns auf Maschinen vgl. Hahn 2024.

Effizienz zu betrachten. Vielmehr ist zu fragen, ob und inwiefern diese Verfahren dort, wo sie eingesetzt werden, negativ beurteilte Konsequenzen für das Handeln haben können.

Vergleicht man die beiden Verfahren unter diesem Gesichtspunkt, dann ist festzuhalten, dass Scorecards einfach handhabbar und transparent sind. Damit ist gemeint, dass jeder – auch der Laie – aus seinen oder den Merkmalsausprägungen einer anderen Person leicht einen Wert berechnen und sehen kann, wie viel ein Parameter zum Gesamtergebnis beiträgt. Diese einfache Kontrollierbarkeit und Nachvollziehbarkeit geht bei künstlichen neuronalen Netzen verloren.[13] Angesichts der 36 Gewichtungen im Beispiel lässt sich nicht im Einzelnen nachvollziehen, warum welcher Wahrscheinlichkeitswert für eine Patientin berechnet wurde. Viele Anwendungen kommen auf mehr Variablen und mehr Schichten, so dass sich leicht eine Zahl von 1000 Gewichtungen ergibt. Wie sich die berechneten Werte im Einzelnen zusammensetzen, ist – auch für die Programmierer des Algorithmus – *nicht mehr nachvollziehbar*. Hier findet die Rede von der *black box* ihre Anwendung.

Neuronale Netzwerke genießen jedoch besondere Wertschätzung, weil sie – anders als Verfahren wie die logistische Regression – feinere Muster in den Daten auffinden können. Man setzt sie bereits sehr erfolgreich bei vielerlei Arten von Mustererkennung, sei es von gesprochener Sprache, Schrift und Bildern, ein.

3. Wer ist verantwortlich? – Das klassische Verantwortungskonzept

Die Zuschreibung von Verantwortung ist eine soziale und auch rechtlich normierte Praxis. Üblicherweise und in einem ersten Zugriff wird Verantwortung zugeschrieben, wenn ein *Schaden* entstanden ist, der wesentlich auf menschliches *Handeln* zurückgeht. Mit dieser ersten Charakterisierung soll nicht ausgeschlossen werden, dass man auch für handelnd herbeigeführte *positive* Zustände Verantwortung zuschreibt; diese Variante ist aber an dieser Stelle weniger interessant. Mit dem Zusatz »normiert« ist gemeint, dass nicht jede faktisch geäußerte Verantwortungszuschreibung auch eine solche ist. Vielmehr haben sich in Gemeinschaften Kriterien für die *korrekte* Zuschreibung von Verantwortung herausgebildet. In der Strafrechtsprechung beispielsweise sind solche Bedingungen für die Zurechnung von Schäden systematisch etabliert. Die folgende Charakterisierung orientiert sich an der

13 Vgl. zu dieser speziellen Opakheit verfahrensinhärenter Art: Burrell 2016. Zu den Vor- und Nachteilen neuronaler Netzwerke in Bezug auf Effizienz und Interpretierbarkeit: Baesens 2014: 48ff.

deutschen Strafrechtsprechung, doch gehören die genannten Elemente anerkanntermaßen zum Kern des neuzeitlichen Verantwortungsbegriffs.[14]

Zunächst ist noch einmal hervorzuheben, dass die Verantwortungszuschreibung dort ansetzt, wo Schäden durch das *Handeln* von Akteuren hervorgerufen wurden. Vorgänge, die nicht durch Handeln beeinflussbar sind, sind kein Gegenstand der Verantwortungszuschreibung. So wird beispielsweise niemand für den Ausbruch eines Vulkans verantwortlich gemacht. Verantwortungszuschreibung ist jedoch immer eine soziale Praxis auf dem Hintergrund von Handlungsmöglichkeiten und Wissen: Wenn es irgendwann möglich wäre, durch technische Maßnahmen einen Vulkanausbruch zu verhindern, könnte eine entsprechende Unterlassung zum Gegenstand der Verantwortungszuschreibung werden.

Das klassische Konzept der Zuschreibung von Verantwortung in Bezug auf einen Schaden enthält die Bedingungen Kausalität, Norm und Normverletzung, Willentlichkeit und Wissentlichkeit.[15] Jemand wird genau dann korrekt für einen Schaden verantwortlich gemacht, wenn er durch sein Handeln diesen Schaden hervorgebracht hat, die Kausalitätsbedingung also erfüllt ist, und wenn er dabei eine bestehende Norm verletzt hat und diese Hervorbringung wissentlich und willentlich erfolgt ist. Es reicht somit nicht, lediglich ein Element in einer Kausalkette zu identifizieren, die zum Schaden führt. Sonst wäre beispielsweise auch der Inhaber eines Messergeschäftes für einen Mord verantwortlich zu machen, der mit einem in diesem Geschäft erworbenen Filetiermesser verübt wurde. Vielmehr ist für die korrekte Verantwortungszuschreibung entscheidend, dass zum Zeitpunkt der Tat eine Norm in Kraft war, die die Herbeiführung eines Tatbestands verbietet und der Akteur diese Norm verletzt hat (vgl. Heinrich 2005: 73ff.).[16] Der Verkauf von Küchenmessern zählt jedoch zu den erlaubten Tätigkeiten.

Dieses klassische Konzept – obwohl bereits in wesentlichen Anteilen in Aristoteles' Nikomachischer Ethik angedeutet (Aristoteles 2006: Drittes Buch) – ist selbst

14 Für die Bedingungen der objektiven und subjektiven Zurechnung vgl. Heinrich 2005: 73ff. Für die Herausbildung des Verantwortungskonzepts vgl. Bayertz 1995.

15 Vgl. für eine Einbettung und die besondere Rolle von Normen sowie weitere Literaturhinweise Hahn 2014.

16 In der – angelsächsischen – philosophischen Diskussion findet sich auch eine bloße »Kausalverantwortung«, bei der schwer vorstellbar ist, wie sich damit die Steuerungswirkung von Verantwortungszuschreibungen realisieren ließe: vgl. z.B. Braham/van Hees 2012. – Hier unterscheidet sich die deutschsprachige Tradition, wie sie bei Bayertz (Bayertz 1995) in ihrer historischen Entwicklung und sozialen Konstruiertheit nachgezeichnet wird und auch in Handbuchartikeln (vgl. Werner 2011) weitergegeben wird. Die Rede von bloßer »Kausalverantwortung« wird als abgeleitete, eher metaphorische Redeweise aufgefasst, der gegenüber die korrekte Zuschreibung retrospektiver Verantwortungskonzept über die Kausalität hinaus weitere Bedingungen erfordert.

das Resultat, oder besser, das Zwischenresultat einer langfristigen kulturellen Entwicklung. So hat es beispielsweise die Zurechnung eines Schadens zu ganzen Gruppen gegeben, denen der eigentliche Täter angehört hat, oder aber, noch bis ins Mittelalter hinein, auch die Verurteilung von Tieren oder Gegenständen.[17] Insgesamt ist festzuhalten, dass die Praxis der Verantwortungszuschreibung eingebettet ist in ein Verständnis vorsätzlichen, kausal wirksamen Handelns durch Akteure und in Vorstellungen, welche Handlungsweisen (nicht) erlaubt sein sollen.

4. Verantwortungszuschreibung beim Einsatz künstlicher Intelligenz in der medizinischen Diagnostik

Inwiefern fordert der Einsatz künstlicher Intelligenz in der medizinischen Diagnostik die Praxis der Verantwortungszuschreibung heraus? Ärztliches Tun, d.h. Handeln und Unterlassen, kann – wie jedes Handeln – nicht nur erwünschte Folgen herbeiführen, sondern eben auch Schädigungen. Typischerweise handelt es sich bei Schädigungen durch ärztliches Tun nicht um vorsätzliche Handlungen. Schädigungen in medizinischen Kontexten sind vielmehr in der Regel darauf zurückzuführen, dass Individuen ihre rollenbezogenen Sorgfaltspflichten nicht oder nur mangelhaft wahrgenommen haben. Im Strafrecht werden sie den sogenannten Fahrlässigkeitsdelikten zugeordnet. Die Verletzung einer ärztlichen Sorgfaltspflicht erhält eine entscheidende Rolle bei der Zuschreibung von Verantwortung (vgl. Heinrich 2005: 66). Wenn Patienten in Folge ärztlichen Handelns geschädigt wurden, ist für die korrekte Verantwortungszuschreibung erforderlich, dass mit dem Handeln oder Unterlassen Sorgfaltspflichten verletzt wurden. Beispiele stellen Behandlungsfehler dar, die auf fehlerhafte oder nicht dem Stand des Wissens entsprechende Diagnosen zurückgehen oder auf Therapien, die nicht dem Stand der Kunst entsprechen.[18] Wenn ein Patient geschädigt wird und sich diese Schädigung auf die Verletzung ärztlicher Sorgfaltspflichten zurückführen lässt, dann wird der Schaden korrekt der ärztlichen Person zugeschrieben.

Die medizinische Diagnostik ist ein wachsendes Einsatzfeld für maschinelles Lernen. Mittlerweile gibt es Studien, die vermuten lassen, dass insbesondere Varianten des sogenannten *Deep Learning*, zu denen auch neuronale Netze gehören, den menschlichen Diagnosefertigkeiten ebenbürtig sind.[19]

---

17 Vgl. Bayertz 1995: 6f. Das Auspeitschen des Meeres wegen der durch Sturm zerstörten Brücken durch den Perserkönig Xerxes, von dem Herodot berichtet, ist ein Beispiel für die Verurteilung eines Objekts. Prozesse gegen Tiere hat es im Mittelalter gegeben.

18 Rechtlich wird das Erfordernis, die aktuellen Standards zu beachten z.B. in § 630a BGB geregelt.

19 Vgl. zu dieser Aussage die Metastudie Liu et al. 2019. Darin finden sich Anhaltspunkte für die Ebenbürtigkeit der maschinellen Lernverfahren. Zugleich weisen die Autoren auf erheb-

Im präsentierten Beispiel geht es um die Prävention von Herzerkrankungen. Die Verfahren der Scorecard und die neuronalen Netze werden diagnostisch eingesetzt, und zwar, um zu bestimmen, welche Personen zu einer Untersuchung eingeladen werden sollen.

Angenommen, Person A, die einen Herzinfarkt erleidet, hatte zwei Jahre zuvor bei ihrem Hausarzt alle Daten angegeben, die für den Präventionscheck erforderlich waren, ohne eine Einladung zu einer Untersuchung zu bekommen. Es ist davon auszugehen, dass der Herzinfarkt mit entsprechenden Maßnahmen vermeidbar gewesen wäre. Person A will ihren Arzt für diesen Schaden verantwortlich machen. Die Korrektheit dieser Zuschreibung hängt davon ab, ob die erwähnten Bedingungen für die Verantwortungszuschreibung erfüllt sind. Wesentlich ist hier die Prüfung, ob der Hausarzt bei seiner Diagnose *eine Sorgfaltspflicht verletzt* hat. Die Zurückweisung der Verantwortungszuschreibung kann nicht lediglich mit dem Verweis erfolgen, dass eine Prognose »Patient A wird in den nächsten Jahren wahrscheinlich kein Herzkreislaufereignis erleiden« als Prognose mit den üblichen Unsicherheiten behaftet ist. Zwar ist festzuhalten, dass, wer etwas prognostiziert, dieses Ereignis nicht als mit Sicherheit eintreffend angibt. Prognosen sind aber nicht beliebig, sondern an *Korrektheitsstandards* zu prüfen. Die ärztlichen Prognosen erfolgen im Rahmen der entsprechenden Diagnostik und den aus der Vergangenheit bekannten Verläufen: Wenn die-und-die Merkmale in der-und-der Ausprägung vorliegen, dann darf man aufgrund aus der Vergangenheit bekannter Verläufe mit dem-und-dem-Verlauf rechnen. Verstöße gegen die Sorgfaltspflicht liegen vor, wenn der Arzt die relevanten Merkmale oder die bekannten Verläufe nicht adäquat berücksichtigt hat, anders gesagt: wenn Ärzte Fehler gemacht haben. Kann der Arzt nachweisen, dass er diesen Sorgfaltspflichten nachgekommen ist, dann ist er für die Schädigung nicht verantwortlich. Er hat eine gemäß den Standards korrekte Prognose abgegeben, die sich – wie bei Prognosen möglich – nicht bewahrheitet hat. Hat eine Ärztin jedoch z.B. nicht adäquat berücksichtigt, dass eine Patientin einen ungünstigen BMI und hohe Cholesterinwerte hat, dann ist die Sorgfaltspflicht verletzt.

Wie ist jedoch zu verfahren, wenn bei der Diagnose Verfahren künstlicher Intelligenz zum Einsatz gekommen sind? Wie ist zu prüfen, ob der an die Ärztin gerichtete Vorwurf der Schädigung korrekt ist? *Welche Sorgfaltspflichten* sind überhaupt zu

liche methodische Mängel der dieser Metastudie zugrunde gelegten Studien hin, die es auszuräumen gilt, um belastbare und Glaubwürdigkeit schaffende Beurteilungen zu ermöglichen. Beispielsweise wurden häufig keine menschlichen Kontrollgruppen vorgesehen; wenn es Kontrollgruppen gab, wurde nicht sichergestellt, dass Algorithmen und Mediziner dasselbe Datenmaterial beurteilten; es wurde häufig nicht angegeben, wie man mit fehlenden Daten umgeht etc. (vgl. Liu et al. 2019: 291ff.). – Eric Topol (Topol 2019) gibt einen umfassenden Einblick in die Möglichkeiten des Einsatzes maschinellen Lernens in der Medizin.

unterstellen, wenn sie die Prognose nicht selbst erstellt hat, sondern diese nach Dateneingabe in ein Voraussagemodell entstanden ist? Welche Rolle spielt dabei die Art des eingesetzten maschinellen Lernens? Wenn man unterstellt, dass die Sorgfaltspflicht nunmehr lediglich in der korrekten Dateneingabe besteht, dann muss die Ärztin dokumentieren, dass sie diese Pflicht erfüllt hat. Kann sie die korrekte Dateneingabe nachweisen und dann darauf verweisen, dass der berechnete Wert für den Patienten unterhalb der für die Einladung gesetzten Grenze zu einer Präventivuntersuchung gelegen hat, ist sie für den Schaden nicht verantwortlich, da keine Pflichtverletzung vorliegt. Dies trifft für den Einsatz der Scorecard genauso zu wie für die künstlichen neuronalen Netze.

Wenn man jedoch eine weitergehende ärztliche Pflicht unterstellt, die neben der korrekten Dateneingabe auch die Einschätzung der Plausibilität der maschinell erstellten Prognose einschließlich einer Abweichung von dieser Einschätzung umfasst, dann unterscheiden sich die beiden Verfahren in Bezug darauf, inwiefern die Ärztin dieser Pflicht nachkommen kann. Hier findet die unterschiedliche Nachvollziehbarkeit ihren Niederschlag: Bei der Scorecard kann die Ärztin die Gewichtungen der Merkmale im Voraussagemodell mit den üblichen Standards, wie sie über die Aus- und Fortbildung in der Medizin sowie in den Leitlinien der Fachgesellschaften vertreten werden, vergleichen. Relativ auf diese Standards kann sie einen möglichen »Fehler« der maschinellen Prognose feststellen und entsprechend darauf reagieren. Wenn die Ärztin im Beispielfall dieser unterstellten Pflicht nicht nachgekommen ist, und tatsächlich eine Abweichung z.B. in der Gewichtung des Body-Mass-Index durch den Algorithmus vorliegt, den sie durch eine abweichende Einschätzung nicht korrigiert hat, dann ist sie für den Schaden verantwortlich. Ist sie hingegen dieser Pflicht nachgekommen und hat keine Abweichung festgestellt, ist sie für den Schaden nicht verantwortlich.

Bezogen auf dieses Verfahren maschinellen Lernens, das auf die logistische Regression zurückzuführen ist, ist festzuhalten, dass das Handeln der Ärztin sowie das Voraussagemodell des Algorithmus – auch für die Ärztin – nachvollziehbar ist. Basiert das Voraussagemodell hingegen auf einem künstlichen neuronalen Netz, ist die Nachverfolgung eines Fehlers für die Ärztin nicht möglich. In der Abbildung des einfachen neuronalen Netzes (vgl. Abbildung 2) sieht man, dass aufgrund der vielen Gewichtungen die Nachvollziehbarkeit der Prognose und damit die Handhabbarkeit dieses Instruments im Detail auf der Strecke bleibt.[20] Diese mangelnde

20 Mit dem Stichwort der Nachvollziehbarkeit ist die Debatte um die Konzepte Opakheit, Transparenz und Interpretierbarkeit aufgeworfen. Dazu gehören sowohl Fragen zu den Begrifflichkeiten als auch zu den damit verknüpften Forderungen (vgl. weiter unten, 6.; sowie Burell 2016; Durán/Jongsma 2021; Rudin 2019; Grote/Berens 2020). – Im Anschluss an begriffliche Klärungen ist auch der Hinweis zu diskutieren, dass die Forderung nach Erklärbarkeit oder Nachvollziehbarkeit ärztlichen Handelns eine überzogene Forderung sei, da es sich beim ärztlichen Handeln ohnehin um eine eher auf impliziten Kenntnissen beruhende Pra-

Nachvollziehbarkeit lässt eine ärztliche Pflicht zur Plausibilitätsüberprüfung des Voraussagemodells im Einzelnen als nicht erfüllbar erscheinen. Ärzte könnten lediglich eine Abweichung von ihrer auf Standards beruhenden Einschätzung feststellen, nicht aber, aufgrund welcher Gewichtung eines Merkmals diese Abweichung beruht und damit auch nicht, wo ein potenzieller Fehler vorliegen könnte. Da im diskutierten fiktiven Beispiel beide Verfahren maschinellen Lernens zur Verfügung stehen, könnte man unter der Voraussetzung, dass bei ihrem Einsatz auch die Nachvollziehbarkeit der Prognose oder Diagnose gesichert sein soll, verlangen, auf künstliche neuronale Netze zu verzichten und stattdessen auf andere Verfahren wie logistische Regression oder Entscheidungsbäume zu setzen. In Anwendungsfeldern, in denen solche Alternativen zur Verfügung stehen, könnte dies eine sinnvolle Forderung sein.[21] Allerdings lassen sich diese nachvollziehbaren Verfahren nicht für die Bilderkennung, die gerade von besonderer Bedeutung für die medizinische Diagnostik ist, nutzen. In diesen Zusammenhängen sind Verfahren wie die künstlichen neuronalen Netze und andere erforderlich.

Welche *Herausforderungen für die Zuschreibung von Verantwortung* für Schäden im Rahmen medizinischen Handelns lassen sich hier identifizieren? Die Bedingungen für die korrekte Zuschreibung von Verantwortung enthalten wesentlich die *Pflichten* des Akteurs sowie die Bedingungen der *Wissentlichkeit* und *Willentlichkeit*. Der Einsatz von Verfahren maschinellen Lernens fordert die sonst üblichen ärztlichen Sorgfaltspflichten heraus. Im Fall neuronaler Netze ist die Forderung nach einer vergleichenden Einschätzung der Resultate des Algorithmus durch den behandelnden Arzt nicht nur im Ergebnis, sondern bezüglich einzelner Merkmale, obsolet. Wegen der mangelnden Nachvollziehbarkeit des Zustandekommens dieser Resultate lässt

xis handle (vgl. London 2019). *Prima facie* kann dieser Einwurf nicht vollständig überzeugen: Die Formulierung von Standards, wie sie sich insbesondere in medizinischen Leitlinien niederschlägt und auch der Verweis auf »objektive Sorgfaltsmaßstäbe« als Grundlage für Sorgfaltspflichten zeigen den Bedarf, auch ärztlichem Handeln die Pflicht zum Korrektheitsnachweis aufzuerlegen. Allerdings muss sich dieser Nachweis korrekten Handelns nicht zwingend darauf beziehen, die Funktionsweise der eingesetzten Technik im Einzelnen nachvollziehen zu können, sondern kann auch darin bestehen, zu zeigen, dass ein bestimmter geforderter Umgang mit dem technischen Medium eingehalten wurde.

21 Rudin weist daraufhin, dass es viele proprietäre Anwendungen gibt, deren Opakheit sich dem Wunsch der Hersteller bzw. Vertreiber verdankt, nicht ohne weiteres nachahmbar zu sein. Am Beispiel des Systems COMPAS, mit dem in vielen US-Staaten die Rückfallprognosen für straffällig gewordene Personen berechnet werden, konnten Informatiker zeigen, dass sich die Prognosen mit einer einfachen Regelprogrammierung, die nur zwei bis drei Merkmale verwendet, nachgebildet werden können. Die Wahl eines nicht-nachvollziehbaren maschinellen Lernverfahrens ist somit vermutlich nicht immer von dem Ziel der besten Leistungsfähigkeit getragen (vgl. Rudin 2019). – Rudin bestreitet, dass es zwingend einen Trade-Off gibt zwischen der Interpretierbarkeit eines Modells und der Leistungsfähigkeit.

sich zudem die Frage aufwerfen, inwiefern weitere Kriterien für korrekte Verantwortungszuschreibungen erfüllbar sind. Kann man davon sprechen, dass jemand wissentlich und willentlich handelt, wenn das Instrument, das er dabei einsetzt, nicht durchschaubar ist? Bei der einzelnen behandelnden Ärztin geht die oben erwähnte Forderung nach der Fehlereinschätzung des Algorithmus zu weit und damit auch die Forderung, im Detail zu wissen, was man tut, wenn man ein bestimmtes Voraussagemodell einsetzt.[22]

Insgesamt ergeben sich damit bei Schäden, die im ärztlichen Handeln mit Unterstützung künstlicher Intelligenz entstehen, *Zurechnungslücken*: Es ist nicht klar, *welchen* Akteuren die Verantwortung für diese Schäden zuzuschreiben ist. Wie sollte man mit dieser Herausforderung für die Verantwortungszuschreibung umgehen? *Verschiedene Optionen* sind als Reaktion möglich: Man könnte aufgrund der unbefriedigenden Zurechnungslücken den Einsatz dieser Technologie in essenziellen Lebenssituationen verbieten. Man könnte das Verantwortungskonzept in der Anwendung künstlicher Intelligenz suspendieren und mit den Zurechnungslücken leben. Schließlich könnte man versuchen, durch eine Umorganisation des Handelns auf diese Herausforderung zu reagieren.[23]

Die folgende Erwägung der dritten Option verdankt sich einer Reflexion der *Zwecke*, die mit der Zuschreibung von Verantwortung verfolgt werden. Zwar spielen dabei auch Aspekte wie die Vergeltung von Taten bzw. die Entschädigung von Opfern eine Rolle; letztlich dient das Verantwortungskonzept vor allem aber einer *Steuerung des Handelns* (vgl. Hahn 2014; Seebaß 2001). Normen und Pflichten werden in Kraft gesetzt, um bestimmtes Handeln zu fördern und anderes zu unterbinden. Die Verantwortungszuschreibung und eine darauf gründende Entschädigungsforderung bzw. eine Strafe dient der Stützung des sozialen Drucks, sich an die Normen zu halten und die Pflichten zu erfüllen. Viele Normen wie z.B. die, andere nicht zu verletzen oder sie nicht zu bestehlen oder zu belügen, dienen dazu, Schäden zu vermeiden. Die Herausforderung der Verantwortungszuschreibung durch künstliche Intelligenz lässt sich als eine – in der Technikgeschichte übliche – Situation auffassen, in der es darum geht, *erlaubtes* Handeln, das zu Schädigungen führt bzw. führen kann, so umzuorganisieren, dass die Vorteile der technischen Handlungsmöglichkeiten möglichst weitgehend erhalten bleiben, und die Nachteile nicht mehr entstehen. Ein Beispiel aus der Technikgeschichte soll dazu dienen, eine solche Umor-

22 Allerdings gilt vermutlich für viele Technikanwendungen, dass die Anwendenden die Funktionsweise nicht durchschauen und in diesem Sinn »nicht wissen, was sie tun«. Dennoch ist das Nicht-Betätigen der Bremse eines Autos eine Handlung bzw. Unterlassung, die den Autolenkern in ihren Wirkungen bekannt ist. Eine Frage lautet, ob dies bei Algorithmen ebenso der Fall ist.

23 Vgl. zum sogenannten »responsibility gap« im Zusammenhang mit künstlicher Intelligenz Matthias 2004; und möglichen Reaktionsweisen Coeckelbergh 2020; Santoni de Sio/Mecacci 2021.

ganisation in den Blick zu nehmen und auf mögliche strukturelle Ähnlichkeiten im Umgang mit künstlicher Intelligenz zu untersuchen.

5. Ein historisches Beispiel für die Herausforderung der Verantwortungszuschreibung

Die erwähnten Bedingungen bei der Zuschreibung von Verantwortung – Kausalität, Norm, Normverletzung und Vorsätzlichkeit – sind auch schon in der Vergangenheit herausgefordert worden. So wird – prägnant von Kurt Bayertz formuliert – gegen Ende des 18. Jahrhunderts ein tiefgreifender Wandel der Struktur des Handelns verortet, der sich auch im Verantwortungskonzept niedergeschlagen hat. Vor allem zwei Elemente kennzeichnen diesen Wandel: Die zunehmende Arbeitsteilung und der Einsatz von Technik.

> »Die Struktur der gesellschaftlichen Arbeit – als einer paradigmatischen Form menschlichen Handelns – wird im Zuge der Industrialisierung vor allem durch zwei Prozesse in immer kürzeren Zeitabständen revolutioniert: zum einen durch die intensivierte Arbeitsteilung, zum zweiten durch den Fortschritt der Technik. *Damit schieben sich zwischen das handelnde Individuum und die durch dieses Handeln bewirkten Effekte vermittelnde Instanzen, die eine Zurechnung der Handlungsfolge auf bestimmte Individuen erschweren oder gar unmöglich machen.* Zwar hat es solche vermittelnden Instanzen immer schon gegeben, sie gewinnen im Zuge der Industrialisierung seit dem 18. Jahrhundert aber ein solches Gewicht und verstärken die Effekte des Handelns in einem solchen Maße, daß man von einer neuen Qualität sprechen muß: Es besteht keine direkte und lineare Beziehung mehr zwischen dem Akteur und der von ihm hervorgerufenen Folge. Damit wird die Reichweite und Effektivität des Handelns in einem bis dahin unbekannten Maße gesteigert.« (Bayertz 1995: 25)

Die angedeutete Änderung der Handlungsstruktur hat jedoch nicht zu einer Verwerfung des Verantwortungskonzepts geführt, sondern zu einer Veränderung bzw. Ergänzung, die im Folgenden skizziert wird.

5.1 Der Dampfkesselbetrieb

Die Rolle, die der Einsatz von Technik bei veränderten Handlungsstrukturen spielt, lässt sich am Beispiel des Dampfkessels illustrieren: In der ersten Hälfte des 19. Jahrhunderts nahm die Zahl der Kesselexplosionen auf Dampfschiffen enorm zu. Das klassische Modell der Zuschreibung von Verantwortung versagt an dieser Stelle, es ergeben sich *Zurechnungslücken*. Die Dampfkesselunfälle sind keine Resultate absichtlicher Schädigungshandlungen von Individuen an Individuen. Es handelt sich

auch nicht um Pflichtverletzungen, da das Betreiben und Benutzen der Technik ein erlaubtes Handeln darstellt. Vielmehr wird der *Betrieb* dieser von Menschen produzierten technischen Mittel von solchen Unfällen »begleitet«. Dennoch will man sich nicht einfach mit diesen schweren Schäden als bloßen Kollateralschäden erlaubten technischen Handelns abfinden, sondern will die Zurechnungslücken schließen, um Opfer entschädigen zu können und zukünftig Schäden zu vermeiden.

In diesem Rahmen entstanden sogenannte Kesselgesetze, mit denen tiefgreifend in den Betrieb solcher Anlagen eingegriffen wurde (vgl. Lueger 1904). Diese Bestimmungen enthalten verschiedene Arten von Vorgaben, die zum einen *Konstruktionsvorgaben* für Kessel und zum anderen *Normen für die Inbetriebnahme* und *für den laufenden Betrieb* formulieren. So werden für die Zulassung einer Anlage Prüfungen durch zuständige Kesselprüfer gefordert. Die weiter spezifizierten berechtigten Personen müssen für die Inbetriebnahme eine Konstruktionsprüfung, eine Wasserdruckprobe und eine Abnahmeprüfung vornehmen, wobei diese Elemente ebenfalls weiter spezifiziert werden. Außerdem werden in der Folge für die Kesselwärter Verhaltensregeln formuliert, die in den Kesselhäusern aufgehängt werden. Die Dokumentationspflicht schlägt sich in dem Gebot der Führung eines Revisionsbuches nieder: »Jeder einem Dampfkesselüberwachungsverein angehörige oder unter staatlicher Aufsicht stehende besitzt ein Revisionsbuch, in dem der Zeitpunkt und die Ergebnisse aller Prüfungen und Untersuchungen einzutragen sind.«[24]

Bezogen auf den Eisenbahnbetrieb wird schließlich der *Gefährdungstatbestand* eingeführt: Unabhängig vom Verschulden haftet der Betreiber für Schäden, die sich aus dem Eisenbahnbetrieb ergeben.

> »Die Gesellschaft ist zum Ersatz verpflichtet für allen Schaden, welcher bei der Beförderung auf der Bahn, an den auf derselben beförderten Personen und Gütern, oder auch an anderen Personen und deren Sachen, entsteht und sie kann sich von dieser Verpflichtung nur durch den Beweis befreien, dass der Schaden entweder durch die eigene Schuld des Beschädigten oder durch einen unabwendbaren äußeren Zufall bewirkt worden ist. Die gefährliche Natur der Unternehmung selbst ist als ein solcher, von dem Schadenersatz befreiender Zufall nicht zu betrachten.« (Preußisches Eisenbahngesetz, 3.11. 1938, § 25; zit. nach von Gadow 2002: 68)

Mit der Einführung von Gefährdungstatbeständen wurde die bisherige Praxis der Verantwortungszuschreibung wesentlich erweitert. Der Verzicht auf den Nachweis, dass der Schaden vorsätzlich (oder fahrlässig) herbeigeführt wurde, kann vermutlich als eine kleine Revolution der Verantwortungszuschreibung betrachtet werden.

---

24 [http://www.zeno.org/Lueger-1904/A/Dampfkesselbetrieb+%5B1%5D] (Zugriff: 12.06.2024).

5.2 Strategien für den Umgang mit Zurechnungslücken: Konstruktionsbedingungen, Rollenpflichten, Gefährdungshaftung

Die Maßnahmen stellen eine Reaktion auf die durch Technik und Arbeitsteilung entstandenen Zurechnungslücken dar. Die erweiterten Handlungsmöglichkeiten bringen nicht nur als positiv bewertete Resultate hervor, nämlich die industrielle Fertigung sowie die neuen, effizienteren Transportmöglichkeiten, sondern auch negative, nämlich Schäden an Leib und Leben.

Ausgehend davon, dass die Zuschreibung von Verantwortung eine soziale Praxis ist, kann man anhand dieses Beispiels die Funktion dieser sozialen Praxis und ihre Realisierung erläutern. Die Funktion liegt *letztlich* in der Handlungssteuerung. Akteure sollen beispielsweise Handlungen unterlassen, die zu bestimmten unerwünschten Folgen führen. In der rechtlichen Systematisierung dieser Praxis wird im Fall von Körperverletzung, Diebstahl, Betrug etc. diese Zielsetzung verfolgt, indem man Tatbestände als rechtswidrig identifiziert und ihre Herbeiführung verbietet. Es wird also eine entsprechende Verbotsnorm in Kraft gesetzt. Führt jemand diesen rechtswidrigen Zustand herbei und erfüllt die Bedingungen für die Verantwortungszuschreibung, ist die Grundlage für eine Schadenersatzforderung und eine Bestrafung gelegt. *Ex ante* wird dieses Handeln somit als unerwünschtes verboten. Kommt es trotzdem zu solchen Handlungen, wird auf dieser Basis *ex post* Verantwortung zugeschrieben und der Täter bestraft. Die Sanktion versieht das handlungssteuernde Verbot mit einem zusätzlichen sozialen Druck. Soweit zum klassischen Modell.

Bei den geschilderten Dampfkesselunfällen liegt die Lage anders: Der Betrieb von Dampfkesseln ist ein *erlaubtes technisches Handeln mit erwünschten Folgen* – allerdings können auch unerwünschte Nebenfolgen eintreten. Wenn die Funktion der Verantwortungszuschreibung, nämlich die Handlungssteuerung, erfüllt werden soll, ist zu fragen: *Welche Handlungsbeschränkungen bzw. Handlungslenkungen des eigentlich erlaubten und prinzipiell erwünschten Handelns sollen eingeführt werden, um die unerwünschten Folgen, die Schäden, zu vermeiden? Welche Akteure können und sollen in ihrem Handeln wie angeleitet werden?* – Hier handelt es sich um eine *ethische* (wenn es um die Umsetzung in Regulierungen geht, auch rechtliche) Fragestellung: Es geht darum, welche Handlungsbeschränkungen im Sinne von zu unterlassenden, aber auch auszuführenden Handlungen man Akteuren auferlegen darf/sollte/könnte, um die Beeinträchtigungen anderer (oder auch gleicher), die sich ohne diese Regulierung einstellen würden, zu vermeiden. Diese Handlungsbeschränkungen durch entsprechende Normen – im Grenzfall auch das vollständige Verbot dieses Handelns, im Beispiel also das Betreiben von Dampfkesseln –, werden den Akteuren gegenüber als Forderung und mit Zwang aufrechterhalten, auch wenn diese weder der Zielsetzung der Regulierung noch ihrer Rechtfertigung zustimmen. Wie diese Fragen entschieden werden, d.h. wie die Abwägung von erzielbaren erwünschten

Folgen gegenüber den potenziellen Schädigungen ausfällt, ist nicht vorgegeben. Es handelt sich um eine verhandelbare Abwägung, bei deren Durchführung wenigstens zwei Elemente entscheidend sind: *Erstens* müssen die jeweiligen Optionen mit ihren abgeschätzten Konsequenzen für die Entscheider klar dargelegt sein: Wer profitiert in welcher Weise, wer hat Einschränkungen und potenzielle Schäden zu gewärtigen? Mit welchem Grad an Gewissheit lassen sich diese Aussagen treffen? – Hierzu ist in den betrachteten Fällen die entsprechende wissenschaftlich-technische Expertise erforderlich. *Zweitens* muss eine normative Analyse die Regulierungsoptionen auf Verträglichkeit mit bereits vorhandenen Rechten, Pflichten und Verfahrensvorschriften prüfen.[25]

Im Fall der erwähnten technischen Entwicklung, d.h. beim Betrieb von Dampfkesseln und Eisenbahnen, hat die Regulierung auf verschiedenen Ebenen angesetzt. Erstens wurde wissenschaftlich-technische Expertise herangezogen, um Vorschriften für die *Konstruktion* der Kessel zu formulieren. Zweitens wurden Personen mit bestimmter Expertise identifiziert, nämlich Ingenieure, die berechtigt sind, Kessel für die Inbetriebnahme und auch den laufenden Betrieb zu prüfen. Drittens wurde eine Personengruppe identifiziert, nämlich die Kesselwärter, und Vorschriften für ihr Handeln formuliert. Die letzten beiden Maßnahmen lassen sich als Etablierung von *Rollenpflichten* lesen. Rollenpflichten können darin bestehen, bestimmte Zustände, z.B. die Funktionsfähigkeit einer Maschine, aufrechtzuerhalten und geben dabei die auszuführenden Handlungen nicht im Einzelnen vor. Vielmehr ist unterstellt, dass diejenigen, an die sich die Pflicht richtet, über die notwendige Expertise verfügen, um die spezifizierte Aufgabe zu erfüllen. In der Debatte um das Verantwortungskonzept wird der Umstand, dass z.B. der Einsatz von Technik die Anwendung des klassischen individualistischen Verantwortungskonzepts verhindert, mit der Forderung nach einem neuen Verantwortungskonzept verbunden. Für den Umgang mit Zurechnungslücken ist jedoch nicht ein neues Verantwortungskonzept erforderlich, sondern es werden vielmehr *neue Pflichten* benötigt. Diese Pflichten bestehen vermehrt in Rollenpflichten, auch um dynamische Entwicklungen des technischen Handlungsfeldes zu erfassen und zu vermeiden, immer wieder dem jeweiligen Entwicklungsstand angepasste Handlungsnormen zu verabschieden. Die Rollenpflichten bauen auf einer Expertise der angesprochenen Akteure auf; diese haben beispielsweise den »Stand der Technik« zu beachten.[26]

Viertens, und das ist die deutlichste Veränderung gegenüber der bisherigen sozialen Praxis, wurden *Gefährdungstatbestände* formuliert und damit zugleich

25 Zur Berücksichtigung empirischer Sachverhalte in normativen Argumentationen vgl. Bayertz 1991: 27ff. – Der Abgleich neuer normativer Forderungen mit bestehenden Rechten und Pflichten lässt sich methodisch durch die Herstellung von Überlegungsgleichgewichten anleiten (vgl. Hahn 2016).

26 Vgl. zur Debatte um das Verantwortungskonzept Bayertz 1995; Hahn 2014.

von den Bedingungen der Verantwortungszuschreibung abgewichen. In diesen Fällen ist die Kausalitätsbedingung nicht mehr relevant bzw. abgeschwächt, d.h. es ist nicht nachzuweisen, dass der Betrieb der Eisenbahn den Schaden konkret hervorgerufen hat. Es reicht, dass der Schaden im Umfeld dieses Betriebes eingetreten ist. Ähnliche Bedingungen finden sich auch bei der später eingeführten Produkthaftung.[27]

Die Formulierung von Konstruktionsbedingungen, von Rollenpflichten sowie vor allem des Gefährdungstatbestands lassen sich im Sinne der Handlungssteuerung lesen: Die externen Bedingungen des Handelns werden durch diese Forderungen und die Androhung von Schadenersatzforderungen bzw. sogar Strafen so verändert, dass gerade die Individuen, die Einfluss auf den Technikbetrieb nehmen können, entsprechende Anreize bekommen, mögliche Schäden zu verhindern. Anders gesagt, handelt es sich um eine Umstrukturierung des Handelns, die auch die Motivationen der Akteure miteinbezieht – eine Vorgehensweise, die Moritz Schlick als zentral für die Verantwortungszuschreibung hervorhebt:

> »Die Frage nach der Verantwortung ist nun die: Wer ist denn im gegebenen Fall eigentlich zu bestrafen? Wer ist als wahrer Täter der Handlung anzusehen? Die Frage ist nicht einfach identisch mit der nach dem Urheber der Handlung überhaupt, denn als solche könnten sich schließlich ebensogut die Urgroßeltern des Täters gelten, denen er durch Vererbung seinen Charakter verdankt, ferner die Staatsmänner, die sein soziales Milieu geschaffen haben usf. – Sondern ›Täter‹ heißt derjenige, *an dem die Motive hätten einsetzen müssen*, um die Tat sicher zu verhindern (bzw. hervorzurufen). […] Die Frage nach dem Verantwortlichen ist die Frage nach dem *richtigen Angriffspunkt der Motive*.« (Schlick 1984[1930]: 161f.)

6. Übertragbarkeit der Strategie auf den Umgang mit Algorithmen?

Lässt sich dieses Vorgehen im Umgang mit Zurechnungslücken auf die Zurechnungslücken übertragen, die durch den Einsatz von Algorithmen in der medizinischen Diagnose entstanden sind? Selbst wenn das nicht der Fall sein sollte, kann die Erörterung möglicherweise dazu dienen, neue Fragen oder neue mögliche Lösungsstrategien zu formulieren.

27 Vgl. dazu unter dem Stichwort »Zurechnungsexpansion« Lübbe 1998. Für eine detailliertere Darstellung des Technikrechts und der durchgreifenden Änderungen im Zeitalter der Industrialisierung vgl. Vec 2011.

6.1 Algorithmen zur Entwicklung von Vorhersagemodellen – Wissenschaftsphilosophische Bemerkungen

Die genannten Fragen erfordern zunächst die Überlegung, welche Funktionen mit dem Einsatz der Algorithmen erfüllt werden sollen. Eine Hauptaufgabe besteht darin, auf der Basis verfügbarer Merkmale die Manifestation einer Krankheit *vorauszusagen*. Grundlage für die Voraussage ist die vorhandene Erfahrung. Die bisherigen Vorkommnisse der Erkrankung werden in einen allgemeinen Zusammenhang zu Merkmalen gesetzt, die die Erkrankten aufweisen. Anders gesagt: Man will durch die Bildung allgemeiner Hypothesen die Mittel bereitstellen, um neue Fälle voraussagen zu können, letztlich um entsprechend intervenieren zu können. Damit ist ein *wissenschaftsphilosophisches Kernthema* berührt, nämlich das von *Erklärung und Voraussage*, und damit auch der Gewinnung von Gesetzeshypothesen durch *Induktion* (vgl. Chalmers 2007: 35ff.; Schurz 2006: 47ff.). Auf diesen Wegen zur Gewinnung von Erkenntnis liegen zentrale wissenschaftsphilosophische Problemstücke, die Gegenstand vieler Überlegungen waren und sind. Dazu gehören neben dem bereits erwähnten Problem des Unterschieds von Korrelation und Kausalität der Umstand der Gehaltserweiterung durch induktive Schlüsse, der Umgang mit Wahrscheinlichkeiten in Anwendung auf Einzelfälle sowie das Problem der Gesetzesartigkeit.[28]

Wissenschaftsphilosophische Reflexionen begleiten die Bemühungen der Wissenschaften, die Erkenntniswege zu sichern. Die Bedingungen für kontrollierte Experimente oder für die Gewinnung statistischer Hypothesen lassen sich als Korrektheitsstandards für die Qualität der allgemeinen Aussagen und der daraus abgeleiteten Voraussagen deuten. Wissenschaftliche Äußerungen mit dem Anspruch auf Erkenntnis sind mit dem Verweis auf das Einhalten der Standards als korrekte Äußerungen zu rechtfertigen. Wer eine Voraussage macht, muss zu ihrem Korrektheitsnachweis auf entsprechende allgemeine Aussagen verweisen und die Gewin-

28 »Daß ein gegebenes Stück Kupfer den elektrischen Strom leitet, erhöht die Glaubwürdigkeit von Aussagen, daß andere Kupferstücke den Strom leiten und damit wird die Hypothese bestätigt, daß alles Kupfer den Strom leitet. Doch die Tatsache, daß ein bestimmter Mann, der sich jetzt in diesem Zimmer befindet, ein dritter Sohn ist, erhöht nicht die Glaubwürdigkeit von Aussagen, daß andere Männer, die sich jetzt in dem Zimmer befinden, auch dritte Söhne sind, und bestätigt also nicht die Hypothese, daß alle Menschen, die sich jetzt in diesem Zimmer befinden, dritte Söhne sind. Doch in beiden Fällen ist unsere Hypothese eine Verallgemeinerung der Datenaussage. Der Unterschied liegt darin, daß im ersten Fall die Hypothese eine gesetzesartige Aussage ist, im zweiten dagegen bloß eine zufällige allgemeine Aussage. Nur eine gesetzesartige Aussage – unabhängig von ihrer Wahrheit oder Falschheit oder ihrer wissenschaftlichen Bedeutung – kann durch einen ihrer Anwendungsfälle bestätigt werden, zufällige Aussagen können es nicht. Offenbar müssen wir uns also nach einer Möglichkeit umsehen, gesetzesartige von zufälligen Aussagen zu unterscheiden.« (Goodman 1975: 97)

nung dieser allgemeinen Aussagen mit dem Nachweis rechtfertigen, beispielsweise kontrollierte Experimente durchgeführt zu haben.[29]

Die unter Einsatz von Algorithmen entwickelten Voraussagemodelle sollen ihren Befürwortern zufolge den menschlichen Prognosen äquivalente Leistungen erbringen. Sie werden im hier betrachteten Beispielbereich zur Erstellung von Diagnosen verwendet und werden somit zum Bestandteil von Handlungsketten, an deren Ende unerwünschte Folgen stehen können. Die Rechtfertigungsbedürftigkeit der Prognosen sowie in einem weiteren Schritt der Bildung von gesetzesartigen Aussagen oder »Muster« führt auf die Frage, auf welche Weise und durch das Handeln welcher Akteure sich die Qualität der Voraussagemodelle sicherstellen lässt.[30] Kann die Strategie des Umgangs mit Zurechnungslücken, die in Bezug auf die Dampfkesselunfälle identifiziert wurde, hier übertragen werden?

6.2 Konstruktionsbedingungen für Algorithmen?

Kern der Herausforderung für die Verantwortungszuschreibung ist die mangelnde Nachvollziehbarkeit der Resultate von Algorithmen maschinellen Lernens. Inzwischen gibt es vielerlei Bemühungen, gerade an diesem Defekt Reparaturmaßnahmen vorzunehmen. Die Erklärbarkeit, Interpretierbarkeit, Nachvollziehbarkeit, Verständlichkeit oder auch Transparenz von Algorithmen werden als von Betroffenen einforderbare Bedingungen in Katalogen zum Umgang mit KI festgeschrieben.[31] Diese Bedingungen ließen sich als Analogon zu den Konstruktionsbedingungen für Dampfkessel diskutieren.[32]

Eine Einschätzung bezüglich der Realisierbarkeit dieser Verständlichkeitseigenschaften scheint aber unter den beteiligten Wissenschaftlern nicht einhellig zu sein. So kommt beispielsweise in der einschränkenden Formulierung »Wenn keine anderen Modelle zur Verfügung stehen, muss die Nachvollziehbarkeit durch

29 Für die Rechtfertigung im statistischen Fall sowie zu Überlegungen zu Korrelation und Kausalität vgl. Schurz 2006: 133ff.

30 Dazu gehört eine Bewertung, wie zuverlässig die Prognosen sind im Vergleich zu menschlichen Voraussagen. Bislang gibt es lediglich Ansätze für einen solchen Vergleich, die die Leistungen der Algorithmen als vielversprechend beurteilen. Allerdings fehlen hierzu noch umfassende und methodisch sorgfältige Vergleichsstudien. (Vgl. Liu et al. 2019) – Eine Zusatzschwierigkeit entsteht für die allgemeine Betrachtung der Leistungsfähigkeit von Algorithmen dadurch, dass die Vorhersagemodelle häufig in privatwirtschaftlichen Arrangements mit entsprechenden Eigentumsrechten entwickelt werden.

31 Vgl. die Übersicht bei Hagendorff 2020.

32 Allerdings wäre hier zunächst eine begriffliche Abgrenzung dieser Eigenschaften und in einem weiteren Schritt eine kritische Betrachtung ihrer Verwendung in normativen Zusammenhängen angezeigt (vgl. Alloa 2019).

den Gebrauch von post-hoc Erklärungen gesteigert werden«[33] zum Ausdruck, dass Verständlichkeit im engeren Sinn einer Vorsehbarkeit und Erklärbarkeit von Input und Output nicht immer realisierbar ist. Dem stehen Aussagen gegenüber, die den häufig behaupteten trade-off zwischen Akkuratheit der Vorhersagen und Transparenz – also der zwingenden Verminderung der Transparenz als Preis der besseren Vorhersagen – als Mythos bezeichnen. Es gebe auch Vorgehensweisen, gute Prognosen oder Klassifikationen zu bekommen, ohne diesen ›Preis‹ zu entrichten.[34] Beim jetzigen Stand lässt sich angesichts dieser Erörterungen lediglich festhalten, dass in der Forschungscommunity keine Einigkeit darüber herrscht, welche Arten der Verständlichkeitsforderungen bezüglich der Konstruktion von KI-Modulen in welcher Form realisierbar sind.

Neben den genannten Bedingungen werden inzwischen aber auch weitere Kriterien für die Qualität von KI-Algorithmen genannt, die das Vertrauen in diese Technologie stärken sollen. Im Anschluss an die wissenschaftsphilosophische Frage, wie sich die Qualität von algorithmischen Voraussagemodellen sichern lässt, und damit letztlich auch das übergeordnete Problem, das Handeln so zu steuern, dass möglichst viel von den erwünschten Konsequenzen des Technikeinsatzes erhalten bleiben, während unerwünschte Folgen vermieden werden, wären diese weiteren Kriterien zu erörtern.

So werden beispielsweise in der erwähnten Richtlinie des Deutschen Instituts für Normung (neben der Nachvollziehbarkeit) Funktionalität und Leistungsfähigkeit einerseits sowie Robustheit andererseits als weitere Bedingungen für einen Qualitätsnachweis von Algorithmen genannt.[35] Dabei werden diese Qualitätsmerkmale erläutert mit Forderungen, die sich auf die adäquate »Problemformalisierung,

33 DIN SPEC 92001–1: Artificial Intelligence – Life Cycle Processes and Quality Requirements – Part 1: Quality Meta-Model. In dieser DIN-Richtlinie wird ein Modell vorgestellt, das den allgemeinen Rahmen für spezielle Qualitätsanforderungen für AI-Entwicklungen bieten soll. Ein Teil dieser speziellen Anforderungen wird in einer weiteren Richtlinie ausgearbeitet: DIN SPEC 92001–2: Artificial Intelligence – Life Cycle Processes and Quality Requirements – Part 2: Robustness. – Die DIN-Richtlinien sind nur ein Beispiel von vielen Initiativen zur Sicherstellung der Qualität von Algorithmen.

34 Vgl. das Plädoyer von Rudin, das sich bereits im Aufsatztitel niederschlägt (Rudin 2019) – Für eine skeptische Position bezüglich der Einlösbarkeit der Verständlichkeitsforderungen vgl. Durán/Jongsma 2021. – Die Beispielbetrachtung der Herz-Kreislauf-Diagnostik sollte illustrieren, dass künstliche neuronale Netze, erst recht, wenn sie mehr Ebenen enthalten als im Beispiel, durch die Vielzahl der beteiligten Gewichtungen nicht mehr hinsichtlich des Weges, den eine Dateneingabe (Merkmalsset) bis zu einem Ergebnis (Prognose) nimmt, nachvollziehbar sind.

35 DIN SPEC 92001–1:2019-04: 21. – Ähnliche Anforderungen, vor allem im Hinblick auf die Einbeziehung des Expertenwissens werden im Übrigen auch von Rudin genannt, wenn es darum geht, die Interpretierbarkeit der Modelle sicherzustellen (vgl. Rudin 2019).

Aufgabenanalyse, Datensammlung, Datenanalyse und Datenverarbeitung«[36] beziehen. Hinzu kommt die Bewertung der Leistungsfähigkeit und die Auswahl des Modells, die wiederum adäquate Kriterien, einschließlich passender Metrisierung voraussetzen. Das Qualitätsmerkmal der Robustheit von Modulen künstlicher Intelligenz soll sicherstellen, dass diese mit irritierenden Daten umgehen können, d.h. z.B. mit solchen Daten, die über die Trainings- und Testdatensets hinausgehen.[37]

Eine ähnliche Richtung schlägt ein Vorschlag ein, der unter dem Titel »Computational Reliabilism« geführt wird.[38] Unter den vier Indikatoren für die Zuverlässigkeit finden sich Verifizierung und Validierung, Robustheitsanalyse, Betrachtung der erfolgreichen bzw. erfolglosen Implementationsgeschichte und die Einbeziehung von Expertenwissen (vgl Durán/Jongsma 2021: 332).

Die Erläuterungen zur sorgfältigen Formulierung der zu lösenden Aufgabe sowie der Datensammlung, -analyse und -verarbeitung im Rahmen des Qualitätsmerkmals »Funktionalität und Leistungsfähigkeit« lassen sich als Pendant zu den Standards bei der »manuellen Entwicklung« von Voraussagemodellen betrachten. Dabei wird deutlich, dass neben dem eigentlichen Einsatz von Verfahren wie logistischer Regression oder künstlichen neuronalen Netzen die *fachliche Expertise* im Umgang mit dem jeweiligen Problem, aus bestehenden Daten und Erfahrungen Voraussagen für neue Fälle abzuleiten, ausschlaggebend sind. Im Beispiel sind es nicht beliebige Daten, die für die Entwicklung des Voraussagemodells herangezogen werden, sondern solche, bei denen Mediziner bereits die Vermutung haben, dass sie mit Krankheitsverläufen korreliert sind. Selbst wenn Verfahren maschinellen Lernens verwendet werden, um erste Hinweise auf Korrelationen zu bekommen, sind diese Vermutungen zum Gegenstand weiterer Analysen und Beurteilungen zu machen.

Die genannten Qualitäts- oder Zuverlässigkeitskriterien lassen sich – jedenfalls u.a. – als Qualitätskriterien für die Gewinnung von Erkenntnis auffassen. So werden datenbezogene Erfordernisse für die Verbesserung der Verallgemeinerung erwähnt. Man kann vermuten, dass z.B. Modelle für die Wirksamkeit von Medikamenten, die in ihrer Datengrundlage nicht repräsentativ sind für die angezielte Patientengruppe, diesen Erkenntnisnormen nicht genügen. Anders als häufig in der öffentlichen Diskussion anzutreffen, ist anzuraten, diese Forderung nicht unter die Rubrik »Ethik« einzusortieren. Nicht jede Norm ist eine moralische Norm. Für die Sicherstellung der Qualität wissenschaftlicher Resultate sind nicht Moralphiloso-

36 DIN SPEC 92001–1:2019-04: 21; Übersetzung – SH.

37 DIN SPEC 92001–1:2019-04: 21.

38 Vgl. Durán/Jongsma 2021. – In diesem Artikel wird auf diesen Ansatz im Kontext medizinischer KI-Anwendungen verwiesen.

phinnen zuständig, sondern die Wissenschaftlerinnen, die sich an den Standards zur Erkenntnisgewinnung orientieren müssen

Mit den aufgeführten Bedingungen, die die Qualität der Vorhersagemodelle sicherstellen sollen, wird zum Teil an Selbstverständlichkeiten wissenschaftlicher Sorgfalt erinnert. An wen richten sich die so formulierten Pflichten? Es scheint – und das ist vermutlich jedenfalls in dieser Umfassendheit eine Neuheit –, dass hier Dateningenieure und Wissenschaftler des jeweiligen Gebiets, für das ein KI-Modul entwickelt wird, in Kooperation verpflichtet werden (vgl. Rudin 2019).

Liegt mit diesen geforderten Merkmalen ein Analogon zu den Konstruktionsbedingungen für Kessel im 19. Jahrhundert vor? Sind Richtlinien solcher Art *geeignet, das Handeln in gewünschter Weise zu steuern*, d.h. so zu steuern, dass man vom Einsatz der KI-Module profitieren kann, ohne ihren Schäden ohne weitere Vorkehrungen ausgesetzt zu sein? Von der Antwort auf diese Frage ist auch die Erörterung des Problems betroffen, inwiefern man den Einsatz der KI-Module in risikoreichen Situationen zumuten kann.

Lassen sich Funktionalität, Leistungsfähigkeit und Robustheit sicherstellen, können diese Anforderungen somit handlungssteuernd wirken? Lässt sich auf diese Weise Vertrauen in die Technologie stärken? Könnten Initiativen wie die TÜV-Zertifizierung von KI-Anwendungen diese Bedingungen nutzen?

Diese Fragen sind zu klären, um einschätzen zu können, ob die genannten Bedingungen ein leistungsäquivalentes Gegenstück zu den Konstruktionsbedingungen für Kessel darstellen. Wenn dies der Fall ist, sie also eine handlungssteuernde Wirkung entfalten können, können auch Verstöße gegen diese Qualitätsstandards festgestellt und zur Grundlage von Verantwortungszuweisungen gemacht werden.

6.3 Rollenpflichten und Gefährdungshaftung

Neben den Konstruktionsbedingungen waren Rollenpflichten für Ingenieure und Kesselwärter sowie das Institut der Gefährdungshaftung Reaktionen auf die entstandenen Zurechnungslücken. Letzteres wird beispielweise inzwischen auch in der Produkthaftung realisiert. Eine Überlegung wäre also, die KI-Module, die in Hochrisikosituationen eingesetzt werden, mit einer solchen Gefährdungshaftung zu verknüpfen. Bezogen auf den Eisenbahnbetrieb hat diese Regulierung offenbar nicht dazu geführt, dass die Gesellschaften ihren Betrieb aus Furcht vor nicht bewältigbaren Entschädigungsleistungen bzw. auch strafrechtlichen Folgen eingestellt haben. Will man die Übertragbarkeit auf KI-Module weiter ausloten, könnten weitere Untersuchungen die Voraussetzungen dieser historischen Verantwortungszuschreibung für Schäden, unabhängig von nachgewiesenem Verschulden eruieren. Es ist allerdings zu vermuten, dass man bei der Regulierung davon ausgegangen ist, dass die Betreiber über das notwendige Know-how verfügen, um Schäden sowohl durch die Konstruktion der Anlagen als auch durch die Normierung der Handha-

bung zu verhindern. Eine entscheidende Frage lautet: Gilt dies auch für Algorithmen? Die Frage verweist erneut auf die Bedingungen für die Handhabbarkeit und damit auf die Funktionsweise.[39]

Wenn man sich gegen eine Gefährdungshaftung entscheidet und stattdessen auf Rollenpflichten setzt, ist zu fragen, an wen sich Rollenpflichten im Fall von KI-Modulen in der medizinischen Diagnostik richten würden? Es liegt vermutlich nahe, hier Pflichten von Dateningenieurinnen und Medizinern bei der Entwicklung der KI-Module zu formulieren, die im Sinne der beiden genannten Qualitätsstandards liegen. Was darüberhinausgehend die Pflicht der einzelnen behandelnden Ärztin angeht, die – möglicherweise gezwungenermaßen – mit den Empfehlungen des Algorithmus umgehen muss, wäre es denkbar zu fordern, dass sie die Empfehlung des KI-Moduls missachtet, wenn sie dies für richtig hält. Ob diese Forderung adäquat ist, ist fraglich – immer auf dem Hintergrund der Zielsetzung, dass man vom Einsatz der Algorithmen profitieren möchte. Zunächst hieße das, dass die behandelnde Ärztin einerseits einem Algorithmus vertrauen, andererseits aber misstrauen soll. Wenn nicht klar ist, aufgrund welcher Merkmale ein Algorithmus zu einem Ergebnis kommt, scheint diese Forderung schwierig einzulösen.[40] Zudem ist auf das allgegenwärtige Problem übermäßigen Vertrauens gegenüber übermäßigem Misstrauen zu verweisen: Häufig stellen sich Routinen ein, in denen das Funktionieren des Algorithmus nicht hinterfragt wird (übermäßiges Vertrauen). Demgegenüber ist auch festzuhalten, dass die Forderung, das Ergebnis des Algorithmus zu missachten, das Ziel, von seinem Einsatz zu profitieren, konterkarieren

39 Ob die Entwicklung, der Betrieb und die Nutzung von KI-Algorithmen und von mit ihnen betriebenen Robotern »nicht mehr mit den Mitteln individualistischer Handlungstheorien zu erfassen sind, sondern auf Theorieangebote zurückgreifen müssen, die imstande sind, neue soziotechnische Ensembles von Menschen als zentrale Phänomene des Gesellschaftlichen zu begreifen« (Gruber 2013: 366) – das soll hier bezweifelt werden: Wie im oben aufgeführten Schlick-Zitat dargelegt, geht es bei der Handlungssteuerung darum, diejenige Person zu identifizieren, die die Tat hätte verhindern können. Solange Maschinen keine Motive aufweisen, an denen Anreize oder Sanktionen ansetzen können, bleibt nur der Ansatz bei den agierenden Personen (oder Personenkollektiven wie Unternehmen). Diesen Weg schlägt Gruber (der interessanterweise in seinem Aufsatz ebenfalls mit dem Vergleich zur Gefährdungshaftung des Eisenbahnbetriebs operiert) mit seiner Überlegung zur Haftungsverteilung selbst ein: Er unterstellt, dass die Personen, die wissen, dass sie haften werden, die entsprechenden Sicherungsmaßnahmen ergreifen bzw. Versicherungen abschließen (vgl. Gruber 2013: 370).

40 Auch die Forderung von Hannah Fry, Mathematikerin, das Beste aus beiden Welten in einer »AI-alliance« zu vereinigen, nämlich die Fähigkeit des Algorithmus, Veränderungen von Gewebe zu entdecken und die Fähigkeit der Ärztinnen, falsch-positive Resultate zu identifizieren, ist auf diesem Hintergrund zu hinterfragen (vgl. Fry 2018: bes. 102ff.). – Für eine Diskussion des Umgangs abweichender Einschätzungen von Algorithmus und Ärztinnen vgl. Grote/Berens 2020.

kann. Es ist gerade nicht so, dass der menschliche Eingriff immer zu besseren Resultaten führt.

7. Zwei Fragen als Resümee

Der Einsatz von Algorithmen in der medizinischen Diagnostik fordert wegen resultierender Zurechnungslücken überkommene Praxen der Verantwortungszuschreibung hinaus. Das dadurch aufgeworfene Problem, ob sie wegen dieser Herausforderung überhaupt zum Einsatz kommen sollen oder dürfen, und wenn ja, ob diese Anwendungen unter der Maßgabe regulatorischer Vorgaben stehen sollen, lässt sich aufgrund der vorangegangenen Analyse in zwei Fragen überführen:

A. *Sollen Handlungsbeschränkungen des eigentlich erlaubten und prinzipiell erwünschten Handelns eingeführt werden, um die unerwünschten Folgen, die Schäden, zu vermeiden?*

Will man diese Abwägungsentscheidung von ethischer Relevanz behandeln, dann gehen die verfolgten Ziele, die auf dem Spiel stehenden Interessen von Individuen aus unterschiedlichen Gruppen, das Verhältnis zu bestehenden Rechten und Pflichten sowie auch Überlegungen zur Umsetzbarkeit entsprechender Regulierungen ein. Der letzte Gesichtspunkt stellt die Verbindung zur zweiten Frage her:

B. *Welche Akteure können und sollen in ihrem Handeln wie angeleitet werden?*

Hierzu ist Auskunft von Informatikern, Datenwissenschaftlern und Fachwissenschaftlern erforderlich:
Lässt sich Transparenz herstellen?
Lassen sich andere Gütekriterien für Algorithmen zur Entscheidungsfindung formulieren?

Antworten auf diese Fragen sind wesentlich, um darüber entscheiden zu können, welche Regulierungen sich umsetzen lassen und zu welchen Zielen sie beitragen können, d.h. inwiefern sich das Handeln in gewünschter Weise lenken lässt. Normative Überlegungen, so eine Lehre daraus, umfassen immer auch Überlegungen nicht-normativer Art.

All das zusammen liefert die Grundlage für aufgeklärte Urteile.

Literatur

Alloa, E. (2019): Das Unbehagen in der Transparenz, in: *Internationales Jahrbuch für Medienphilosophie*, 5(1), 155–182.

Aristoteles (2006): Nikomachische Ethik. übersetzt und herausgegeben von Ursula Wolf, Reinbek bei Hamburg: Rowohlt Verlag.

Baesens, B. (2014): Analytics in a big data world. The essential guide to data science and its applications, Minneapolis: John Wiley & Sons, Inc.

Bayertz, K. (1991): Praktische Philosophie als angewandte Ethik, in: Ders. (Hg.), Praktische Philosophie. Grundorientierungen angewandter Ethik, Reinbek bei Hamburg: Rowohlt Verlag, 7–47.

Bayertz, K. (1995): Eine kurze Geschichte der Herkunft der Verantwortung, in: Ders. (Hg.), Verantwortung: Prinzip oder Problem?, Darmstadt: Wissenschaftliche Buchgesellschaft, 3–71.

Braham, M.; van Hees, M. (2012): An anatomy of moral responsibility, in: *Mind*, 121(483), 601–634.

Bringsjord, S.; Govindarajulu, N.S. (2020): Artificial intelligence, in: Zalta, E.N.; Nodelman, U. (Hg.), The Stanford Encyclopedia of Philosophy, Stanford (CA): The Metaphysics Research Lab. [https://plato.stanford.edu/archives/win2019/entries/artificial-intelligence/] (Zugriff: 22.02.2022).

Burrell, J. (2016): How the machine ›thinks‹. Understanding opacity in machine learning algorithms, in: *Big Data & Society*, 3(1), 1–12.

Chalmers, A.F. (5. Auflage 2007): Wege der Wissenschaft. Einführung in die Wissenschaftstheorie, Berlin: Springer-Verlag.

Coeckelbergh, M. (2020): Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability, in: *Science and Engineering Ethics*, 26(4), 2051–2068.

Durán, J.M.; Jongsma, K.R. (2021): Who is afraid of black box algorithms? On the epistemological und ethical basis of trust in medical AI, in: *Journal of Medical Ethics*, 47(5), 329–335.

Engemann, C. (2018): Rekursionen über Körper. Machine Learning-Trainingsdatensätze als Arbeit am Index, in: Engemann, C.; Sudmann, A. (Hg.), Machine Learning. Medien, Infrastrukturen und Technologien der Künstlichen Intelligenz, Bielefeld: transcript Verlag, 247–268.

Finlay, S. (2017): Artificial intelligence and machine learning for business. A no-nonsense guide to data driven technologies, Lancashire: Relativistic.

Fry, H. (2018): Hello world. How to be human in the age of the machine, New York: W. W. Norton & Company.

von Gadow, O. (2002): Die Zähmung des Automobils durch die Gefährdungshaftung, Berlin: Duncker & Humblot.

Goodman, N. (1975): Tatsache, Fiktion, Voraussage, Frankfurt a.M.: Suhrkamp.

Grote, T.; Berens, P. (2020): On the ethics of algorithmic decision-making in healthcare, in: *Journal of medical ethics*, 46(3), 205–211.

Gruber, M.-C. (2013): Gefährdungshaftung für informationstechnologische Risiken: Verantwortungszurechnung im ›Tanz der Agenzien‹, in: *Kritische Justiz*, 46(4), 356–371.

Hagendorff, T. (2020): The ethics of AI ethics. An evaluation of guidelines, in: *Minds and Machines*, 30(1), 99–120.

Hahn, S. (2014): Norm und Verantwortung, in: *Archiv für Rechts- und Sozialphilosophie*, 100(4), 429–449.

Hahn, S. (2016): From Worked-Out Practice to the Justification of Norms by Producing a Reflective Equilibrium, in: *Analyse & Kritik*, 38(2), 339–369.

Hahn, S. (2024): Algorithmische ›Entscheidungen‹ in der Medizin? Eine Reflexion zu einem handlungsbezogenen Ausdruck, in: Ruschemeier, H.; Steinrötter, B. (Hg.), Der Einsatz von KI & Robotik in der Medizin. Interdisziplinäre Fragen, Nomos: Baden-Baden, 13–26.

Heinrich, B. (2005): Strafrecht – Allgemeiner Teil I, Stuttgart: Kohlhammer Verlag.

Lepri, B.; Oliver, N.; Letouzé, E.; Pentland, A.; Vinck, P. (2018): Fair, transparent and accountable algorithmic decision-making processes. The Premise, the proposed solutions, and the open challenges, in: *Philosophy & Technology*, 31(4), 611–627.

Liu, X.; Faes, L.; Kale, A.U.; Wagner, S.K.; Fu, D.J.; Bruynseels, A.; Mahendiran, T.; Moraes, G.; Shamdas, M.; Kern, C.; Ledsam, J.R.; Schmid, M.K.; Balaskas, K.; Topol, E.J.; Bachmann, L.M.; Keane, P.A.; Denniston, A.K. (2019): A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging. A systematic review and meta-analysis, in: *The Lancet Digital Health*, 1(6), 271–297.

London, A.J. (2019): Artificial intelligence and black-box medical decisions. Accuracy versus explainability, in: *Hastings Center Report*, 49(1), 15–21.

Lübbe, W. (1998): Verantwortung in komplexen kulturellen Prozessen, Freiburg i.Br.: Verlag Karl Alber.

Mainzer, K. (2016): Künstliche Intelligenz – Wann übernehmen die Maschinen?, Berlin: Springer Verlag.

Matthias, A. (2004): The responsibility gap. Ascribing responsibility fort he actions of learning automata, in: *Ethics and Information Technology*, 6(3), 175–183.

Pearl, J. (2018): The book of why: The new science of cause and effect, New York: Basic Books.

Ramge, T. (2018): Mensch und Maschine. Wie Künstliche Intelligenz und Roboter unser Leben verändern, Stuttgart: Reclam Verlag.

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.

Santoni de Sio, F.; Mecacci, G. (2021): Four Responsibility Gaps with Artificial Intelligence. Why they Matter and How to Address them, in: *Philosophy & Technology*, 34(4), 1057–1084.

Schlick, M. (1984[1930]): Fragen der Ethik, Frankfurt a.M.: Suhrkamp.

Schurz, G. (2006): Einführung in die Wissenschaftstheorie, Darmstadt: Wissenschaftliche Buchgesellschaft.

Seebaß, B. (2001): Kollektive Verantwortung und individuelle Verhaltenskontrolle, in: Wieland, J. (Hg.), Die moralische Verantwortung kollektiver Akteure, Berlin: Physica Verlag, 79–99.

Topol, E. (2019): Deep medicine. how artificial intelligence can make healthcare human again, New York: Basic Books.

Vec, M. (2011): Kurze Geschichte des Technikrechts, in: Schulte, M.; Schröder, R. (Hg.), Handbuch des Technikrecht, Berlin: Springer Verlag, 3–92.

Werner, M.H. (2011): Verantwortung, in: Düwell, M.; Hübenthal, C.; Werner, M.H. (Hg.), Handbuch Ethik, Stuttgart/Weimar: J.B. Metzler, 541–548.