Kaminski, Andreas (2019): Gründe geben. Maschinelles Lernen als Problem der Moralfähigkeit von Entscheidungen. In: Klaus Wiegerling, Michael Nerurkar und Christian Wadephul (Hg.): Ethische Herausforderungen von Big-Data. Bielefeld: Springer, [im Druck].

Gründe geben

Maschinelles Lernen als Problem der Moralfähigkeit von Entscheidungen

Andreas Kaminski

Die unter den Stichworten ‚Künstliche Intelligenz' und ‚Big Data' geführten Debatten haben schnell den Eindruck entstehen lassen, dass angesichts neuer Technologien neue moralische Problemlagen entstünden.[1] Den neuen Technologien müsse, so die weitere Annahme, daher in Form einer neuen Ethik entsprochen werden.[2] Szenarien wie das viel beachtete Trolley-Problem bekräftigen diese Vorstellung im Kontext autonomen Fahrens: Wir scheinen auf einmal vor eine unmögliche Wahl gestellt.[3]

Gegenwärtig werden für viele Kontexte informationstechnische Systeme entwickelt, welche entweder eigenständig *Entscheidungen treffen* oder die *Entscheidungsfindung unterstützen* sollen. Solche Systeme können etwa die Funktion haben, Bewerber zu selegieren, die geeignete Therapie für Patienten zu wählen, eine Aktion im Straßenverkehr (Spurwechsel, Bremsen usw.) durchzuführen, Aktien zu handeln, die Rückfallwahrscheinlichkeit von Straftätern zu prognostizieren oder die Glaubwürdigkeit der Aussagen von einreisenden Personen an der Grenze zu bestimmen, also ein KI-Lügendetektor.[4]

Entscheidungen verweisen in einem begrifflichen Sinne auf Gründe.[5] Es gehört zur Form von Entscheidungen, dass nach ihren Gründen gefragt werden kann; in einem begrifflichen Sinn gibt es keine Entscheidung, die nicht einen Bezug zu Gründen aufweist. Das heißt also nicht, dass vorausgesetzt wird, Entscheidungen müssten *faktisch* immer begründet werden; auch nicht, dass es immer *einfach* ist, eine solche Begründung zu geben; noch, dass solche Begründungen nicht scheitern können oder umstritten sind, wenn sie versucht werden. Vielmehr bedeutet es, dass, wenn etwas als eine Entscheidung betrachtet

---

[1] Vgl. dazu die Kontroverse, welche im *Jahrbuch Technikphilosophie* 2019 erscheint (Beck et al. 2019.), ob eine Roboterethik erforderlich sei; in vielen Beiträgen zur Debatte ist nicht erkennbar, was die Robotik von anderen Bereichen der Informationstechnik in einer Weise unterscheidet, dass sie eigene, neuartige ethische Probleme stellt, denen in Form einer Roboterethik zu begegnen wäre.

[2] Vgl. dazu die von Phillip Otto und Eike Gräf herausgegebene Edition *Ethik der digitalen Zeit* (Otto und Gräf 2018), welche in der Schriftenreihe der Bundeszentrale für politische Bildung 2018 erschien. In deren Einleitung wird schnell und suggestiv von der Neuartigkeit digitaler Technologien auf das Erfordernis einer neuen Ethik geschlossen: „Die Welt verändert sich mit einer unglaublichen Geschwindigkeit." (S. 8f.) Daher lautet die programmatische Losung: „Die Neuerfindung der Ethik ist unsere Aufgabe!" (S. 6). Vgl. für eine kritische Diskussion dieser Edition Brenneis 2019.

[3] Gleichwohl handelt es sich dabei bislang um ein Gedankenexperiment. Und das Gedankenexperiment ist älter als autonome Fahrzeuge.

[4] Vgl. dazu https://www.iborderctrl.eu/

[5] Für die Idee praktischer Vernunft ist das Gründe-geben-Können so zentral, dass die meisten ethischen Begriffe wie ‚Selbstbewusstsein', ‚Handeln', ‚Rechtfertigen', ‚Allgemeinheit' in einem engen explikativen Zusammenhang damit stehen. Vgl. etwa Korsgaard 2009; Bittner 2005; Bieri 2003; Halbig 2007.

wird, damit die Möglichkeit eröffnet ist, nach den Gründen, welche der Entscheidung zugrunde lagen, zu fragen. Gründe sind dabei nicht Ursachen, da Gründe eine wesentlich normative Dimension aufweisen. Gründe können reflektiert und bewertet werden. Sie können als gute oder schlechte Gründe betrachtet werden. Anders Ursachen, welche nie gute oder schlechte Ursachen, sondern hinreichend oder notwendig sein können.

Stehen Entscheidungen und Gründe in einem so engen Zusammenhang, dann ergibt sich die Frage, wie Systeme, die entworfen werden, um Entscheidungen zu treffen oder die Entscheidungsfindung zu unterstützen, diesen Zusammenhang berühren mögen. Die Gründe für Entscheidungen sind von zentraler Bedeutung für ihre ethische Bewertung. Verändert sich mit solchen Systemen die Gegebenheit und die Form von Gründen, dann stellt sich die Frage, welche ethische Bedeutung dieser Wandel haben könnte.

Diese Frage geht mit einem Perspektivwechsel einher. Statt die neue Technik im Lichte der Frage zu betrachten, ob sie eine neuartige Ethik erfordert, betrachte ich die neue Technik unter dem Gesichtspunkt, wie sie praktische Begriffe rekonfiguriert und wie man sich ethisch zu ihr verhalten kann. Es geht also weniger um die Frage einer *neuartigen Ethik* als vielmehr um die Frage nach den *Möglichkeiten ethischen Verhaltens zu einer (neuartigen) Technologie, welche unsere moralische Praxis affiziert*. Diese Perspektive ist von Christoph Hubig und Philipp Richter für die Technikethik im Allgemeinen entwickelt worden.[6] Sie ist bedeutsam, um die ethische Relevanz von Entscheidungssystemen, welche auf lernenden Algorithmen beruhen, zu analysieren.

Die Frage im Folgenden ist: Verändern solche Entscheidungssysteme die Rolle, welche das Gründe-geben-Können für die moralische Bewertung von Entscheidungen spielt, in einer Weise, die ethisch problematisch sein könnte? Meine Argumentation wird in folgender Weise erfolgen:

1. Die Entscheidungen solcher Systeme beruhen auf Modellen, die von lernenden Algorithmen gebildet werden.
2. Diese Modelle sind in der Regel mathematisch opak (intransparent).
3. Personen können dann das Zustandekommen der Inferenzen und damit der Entscheidung nicht nachvollziehen.

Das führt zu zwei Konsequenzen, welche gleichsam die zwei Seiten der opaken Medaille darstellen:

A. Die Gründe für Entscheidungen von lernenden Algorithmen bleiben für Personen internalistisch opak. Die Form der Begründung solcher Systeme erfolgt stattdessen externalistisch: Sie werden durch Reliabilität gerechtfertigt. Damit stellt sich die Frage nach der *Güte der Gründe*.
B. Die Rechtfertigung durch Reliabilität könnte in einigen Situationen jedoch eine falsche Art von Gründen darstellen. Damit stellt sich die Frage nach der *Angemessenheit der Gründe*.

Meine Argumentation und ihren Geltungsbereich möchte ich sogleich spezifizieren und einschränken: Sie gilt nicht für alle Lernstrategien in gleichem Maße. Je nach Lernstrategie (und natürlich auch Aufgabenstellung und Komplexitätsgrad des Problems) wird die Opazität mehr oder weniger groß und unumgänglich ausfallen. Entscheidungsbäume mögen in vielen Fällen noch nachvollziehbar sein,

[6] Vgl. dazu Hubig 2015, S. 193-205; Hubig und Richter 2015.

künstliche neuronale Netze (mit mindestens zwei ‚hidden layers') oder evolutionäre Algorithmen sind jedoch selbst in relativ einfachen Fällen vielfach auch für Expertinnen mathematisch nicht mehr im Detail nachvollziehbar. Eine zweite Einschränkung: Es gibt keinen Beleg dafür, dass zukünftig nicht mathematische Techniken entwickelt werden könnten, welche es erlaubten in viel größerem Detailgrad die Entscheidungsfindung eines neuronalen Netzes nachzuvollziehen; auch wenn es aktuell nicht absehbar sein mag, wie dies gelingen könnte.

Der Aufbau ist wie folgt: Zunächst werde ich einige der technikphilosophisch bedeutsamen Eigenschaften lernender Algorithmen darstellen. Anschließend zeige ich, wie diese Eigenschaften das Verhältnis von Personen zu Algorithmen verändern und dadurch die Rolle, Gründe geben zu können, betreffen. Schließlich betrachte ich die Bedeutung der beiden Konsequenzen A und B in ethischer Perspektive.

1. Lernende Algorithmen: eine transklassische Technik

Lernende Algorithmen erfahren erst seit kurzem eine größere Aufmerksamkeit in der Öffentlichkeit, der Technikfolgenabschätzung, der Philosophie und den Sozialwissenschaften. Das mag auch daran liegen, dass die Technologie nicht eigenständig in Erscheinung tritt. Lernende Maschinen werden als Subsysteme in technische Systeme integriert: in Spracherkennungssysteme, in medizinische Assistenzsysteme, in Recommender-Systeme, welche Nutzern Produkte zum Kauf vorschlagen, in Finanzsysteme, die mit Aktien handeln, oder in Großforschungsanlagen wie den Large Hadron Collider am CERN. Lernende Algorithmen können überall dort eingesetzt werden, wo große Datenmengen gegeben sind. Um welche Daten es sich handelt, spielt dafür keine Rolle. Daher rührt auch die enorme Anwendungsbreite von Machine Learning-Systemen. Ihr Einsatz beruht auf einer Art Wette: der Annahme nämlich, dass sich in den Daten eine Struktur finden lässt, welche die Ordnung dieser Daten beschreibt und die für praktische Zwecke nutzbar gemacht werden kann:

> Our belief is that behind all this seemingly complex and voluminous data, there lies a simple explanation. That although the data is big, it can be explained in terms of a relatively simple model with a small number of hidden factors and their interaction. Think about millions of customers who buy thousands of products online or from their local supermarket every day. This implies a very large database of transactions; but what saves us and works to our advantage is that there is a pattern to this data. People do not shop at random. A person throwing a party buys a certain subset of products, and a person who has a baby at home buys a different subset—there are hidden factors that explain customer behavior. It is this inference of a hidden model—namely, the underlying factors and their interaction—from the observed data that is at the core of machine learning. (Alpaydin 2016, XI)

Lernende Algorithmen sollen Modelle bilden, welche die einer Datenmenge zugrundeliegende Struktur beschreiben. Die Modelle sind per se weder schlecht noch gut; die Güte eines Modells lässt sich erst in

der Hinsicht auf einen praktischen Zweck beurteilen.[7] Solche Zwecke können sein, handschriftlich notierte Ziffern zu erkennen, die Signale im Large Hadron Collider vom Hintergrundrauschen zu unterscheiden, oder eine Person einer Risikoklasse zuzuweisen. Der Lernvorgang, welcher diesen Algorithmen ihren Namen gibt, besteht genau darin: die Modelle, welche die Datenstruktur beschreiben, zu bilden und, gemessen an einem praktischen Zweck, zu verbessern. Dies bedeutet dann, weniger Fehler in der Erkennung der Ziffern zu machen; die Signale besser vom Hintergrundrauschen zu unterscheiden; genauere Risikoklassen zu bilden etc.

Lernende Algorithmen bilden die Modelle in bestimmten Hinsichten eigenständig. Diese Eigenständigkeit in der Modellbildung muss jedoch präzise bestimmt werden. Die diesen Algorithmen zugeschriebenen Eigenschaften wie Lernen, Autonomie, Adaptivität, Smartness evozieren sonst allzu suggestive Assoziationen mit den Fähigkeiten von Personen.[8] Dann erscheinen lernende Algorithmen als quasi-subjektgewordene Maschinen. Von einer Lernfähigkeit solcher Algorithmen kann jedoch nur in einem eingeschränkten Sinne gesprochen werden. Wichtiger aber noch: Sie beruht in hohem Maße auf der schieren Rechenkraft von Computern und raffinierten mathematischen Verfahren.[9] Dadurch können bestimmte Aspekte der Leistungen von Personen *funktional äquivalent* nachgebildet werden, ohne dass der Lernprozess selbst personales Lernverhalten imitiert.

Es gibt verschiedene Einteilungsprinzipien zur Klassifikation dieser Algorithmen (wie überwacht/unüberwacht) und verschiedene Strategien des Lernens (wie Entscheidungsbäume, künstliche neuronale Netze, evolutionäre Algorithmen usw.). Bevor es aber um die Frage geht, in welcher Weise lernende Algorithmen prozessieren, soll ein Verständnis davon gewonnen werden, inwiefern sich mit lernenden Algorithmen Technik in einer Weise wandelt, die es rechtfertigt, von transklassischer Technik zu sprechen.[10]

Klassische Technik ist weitgehend davon bestimmt, dass sie die Reproduzierbarkeit von Effekten sichert. Immer, wenn ein bestimmtes Ereignis eintritt, soll ein anderes daraufhin folgen. Wird etwa eine bestimmte Aktion initiiert, soll ein bestimmter Effekt auftreten: Wenn ein Schalter gedrückt wird, soll etwa das Licht angehen, der Motor anspringen, eine Tür sich öffnen oder ein Programm sich schließen usw. Klassische Technik verwandelt daher einen Input durch eine Transformationsregel in einen Output.[11] Diese Transformationsregel ist im Fall klassischer Technik konstant. Genauer: Sie *soll* im Fall klassischer Technik unveränderlich sein. Dies ist die normative Anforderung, welche an klassische

[7] Aus der Wissensgeschichte ist bekannt, dass sich zahllose Ordnungen bilden lassen (Foucault 1991). Praktische Zwecke begrenzen die Anzahl möglicher Ordnungen und stellen ein Gütekriterium für die Bewertung und Auswahl der Modelle dar, welche sie beschreiben.

[8] Für eine präzise und kontrollierte Verwendung des Autonomiebegriffs in Beziehung zu lernenden Algorithmen vgl. etwa Wiegerling 2011; Hubig und Harrach 2014.

[9] Vgl. für eine einführende Darstellung aus dem Bereich der Informatik Görz et al. 2003; Lämmel und Cleve 2008; Russell und Norvig 2007; Alpaydın 2008; Alpaydın 2016. Für eine philosophisch orientierte Darstellung Harrach 2014; Kaminski und Glass 2018.

[10] Vgl. dazu auch Hubig 2008.

[11] Heinz von Foerster hat klassische Technik daher als triviale Maschine bezeichnet und von nichttrivialen unterschieden. Vgl. Foerster 1993, 2002 Zu den grundlegenden Problemen dieser Unterscheidung und zur nötigen begrifflichen Umarbeitung, welche eine Anwendung auf lernende Algorithmen ermöglicht, vgl. Kaminski 2014.

Technik gestellt wird; und die Sicherungsbemühungen[12], die unternommen werden, zielen exakt darauf. Die Erwartbarkeit klassischer Technik und damit die Interaktion, ihre Sicherheit, Planbarkeit und Verlässlichkeit beruhen darauf.[13] Immer, wenn ein *m* als Input erfolgt, soll *p* als Output erfolgen, solange die Maschine funktioniert und nicht kaputt ist. Die Transformationsregel bleibt konstant.

Transklassische Technik setzt an dieser Stelle an: der Transformationsregel.[14] Die Transformationsregel wird transformiert. Mit Blick auf lernende Algorithmen heißt dies: ihr Lernvorgang besteht darin, das Modell, welches die Datenstruktur beschreibt, in Hinsicht auf einen gegebenen praktischen Zweck zu verbessern, damit dieser besser erreicht wird. Durch die Veränderung des Modells wird die Verbindung zwischen In- und Output verändert. Auf *m* folgt einmal *p*, das nächste Mal jedoch *q* oder *r*. Dadurch wird aufgegeben, was für klassische Technik zentral ist: die Wiederholbarkeit, Erwartbarkeit und Planbarkeit, welche Voraussetzung der verlässlichen Interaktion mit klassischen Maschinen sind.

2. Opazität, epistemisch und praktisch

Die Vorstellung klassischer Technik ist an die Konstanz der Transformationsregel gebunden, welche garantierte: Immer, wenn *m*, dann *p*. Das Verfolgen praktischer Zwecke, die Aneignung von und das Erlernen, Technik effektiv zu nutzen, erfolgte unter dieser Voraussetzung. Die mechanistische Maschine als Prototyp von Technik erfüllt diese Voraussetzung par excellence. Wenn bei einer mechanistischen Maschine auf *m* nicht *p* folgt, ist man berechtigt, anzunehmen, dass mit der Maschine etwas nicht stimmt; dass sie womöglich kaputt ist.

An der Transformationsregel hängt damit das *praktische* Verstehen von Technik. Man hat in praktischer Perspektive verstanden, wie eine jeweilige Technik funktioniert, wenn man ihre Transformationsregel erkannt hat. Ändert sich die Transformationsregel jedoch aufgrund von Lernvorgängen, so wird es schwieriger, diese zu verstehen.

Von diesem praktischen Verstehen unterschieden ist die Begründung der Funktionsweise. Dies sei hier *epistemisches* Verstehen genannt. Während das praktische Verstehen also darin besteht, zu wissen, was auf *m* folgt, besteht das epistemische Verstehen im Nachvollzug der Begründung, warum die Technik so funktioniert, wie sie funktioniert; warum und wie Input und Output miteinander verbunden sind. In der Regel benötigt man im Alltag nur ein praktisches Verständnis, um Technik nutzen zu können.[15] Ein epistemisches Verständnis ist anlässlich besonderer Situationen jenseits der unmittelbaren Interaktion erforderlich; insbesondere dann, wenn man die Leistung von Technik bewerten möchte, weil etwa in Frage steht, ob und wie gut sie funktioniert, oder weil etwaige Technikfolgen eingeschätzt werden sollen.

---

[12] Vgl. hierzu Hubig 2006, S. 79, 101 et passim.
[13] Vgl. Kaminski 2010.
[14] Zwischen klassischer und transklassischer Technik gibt es einen Übergang; denn klassische Technik kann feine Unterschiede zwischen Eingaben treffen (die unter Umständen differenzierter sind als Personen sie treffen) und auf diese mit differenzierten Ausgaben antworten. Dies wird hier nicht berücksichtigt. Für eine Analyse, die dem nachgeht, vgl. Kaminski 2014.
[15] Hier lässt sich auf Husserls treffende Analyse verweisen, dass in praktischer Perspektive uns Technik davon entlastet, den Begründung- und Entstehungskontext einer Technik kennen zu müssen. Vgl. Husserl 1976, 1989; Kaminski 2013.

Klassische Technik erfordert also nur ein praktisches Verständnis, um sie nutzen zu können. Von einem epistemischen Verständnis werden die Nutzerinnen entlastet. In diesem Sinne wird auch von einem Black Boxing gesprochen. Entscheidend dabei ist, dass im Falle klassischer Technik die Black Box im Prinzip geöffnet werden könnte, wenn man sich für die Funktionsweise interessiert und diese verstehen will. So könnte man sich an Expertinnen wenden, welche eine Einsicht in diese Funktionsweise haben und für die es sich deshalb um eine White Box handelt.

Transklassische Technik hingegen verändert die Möglichkeit des praktischen *und* epistemischen Verständnisses. Hinsichtlich des praktischen Verstehens ist die Situation nicht eindeutig. Zum einen zielt transklassische Technik darauf, Nutzer auch vom nötigen praktischen Verständnis zu entlasten, indem situationsadäquate Entscheidungen für diese getroffen werden. Zum anderen bestehen verschiedene Abstraktionsgrade. Auf einer relativ abstrakten Ebene können transklassische Maschinen weiterhin klassisch agieren (das Übersetzungsprogramm übersetzt weiterhin auf einen Knopfdruck: *m* → *p,* auch wenn sich die Übersetzung des gleichen Satzes durch Lernfortschritte wandelt; Nutzerinnen mögen daher den transklassischen Charakter der Technik nur indirekt erleben).

Wichtiger in unserem Zusammenhang ist die Frage des epistemischen Verstehens. Denn dieses besteht darin, Gründe für die Funktionsweise einer Technik geben zu können. Einige Lernstrategien von lernenden Algorithmen führen nämlich dazu, dass auch Experten nur mehr auf einer abstrakten Ebene, aber nicht mehr im Detail in der Lage sind, die Funktionsweise zu begreifen. Zur Erläuterung wende ich mich zwei Lernstrategien zu: Entscheidungsbäumen einerseits und neuronalen Netzen andererseits. Zwischen ihnen besteht ein relativ großer Kontrast hinsichtlich der Möglichkeit, sie epistemisch zu verstehen.

Entscheidungsbäume können von Expertinnen nicht nur auf prinzipieller Ebene, sondern häufig auch im Detail nachvollzogen werden. Bei dieser Lernstrategie sind die Daten vollständig durch Attribute beschrieben. Das Lernen hat zum Ziel, eine Klassifikation aufzubauen: Dazu wird ausgehend von einem ersten Attribut eine Verzweigungsstruktur entlang weiterer Attribute entwickelt. Auf diese Weise entsteht der Entscheidungs*baum*. So liegen etwa die Kundendaten in einer Bank als Attribute vor: Alter, Geschlecht, Familienstand, Beschäftigungsart, Schufa-Auskunft, Spareinlagen. Zu den Attributen gehört auch, ob ein vergangener Kreditnehmer den Kredit fristgerecht zurückgezahlt hat oder nicht. Der praktische Zweck sei nun die Prognose, ob eine Kreditvergabe an einen neuen Kunden ein hohes oder geringes Risiko darstellt; es handelt sich um eine Klassifikationsaufgabe. Die Daten der Bank werden dazu in Test- und Trainingsdaten geteilt. Der Lernalgorithmus erhält die Testdaten und beginnt nun Entscheidungsbäume zu bilden, welche die Personen korrekt und (möglichst) vollständig in kreditwürdig/nicht-kreditwürdig aufteilen sollen. Der lernende Algorithmus beginnt dazu mit einem Attribut, etwa der Schufa-Auskunft oder der Spareinlage. Es ist dabei prinzipiell nicht festgelegt, mit welchem Attribut der Algorithmus beginnen muss. Da die Attribute einen unterschiedlich großen Informationswert haben, entstehen unterschiedliche Modelle, je nachdem, mit welchen Verzweigungen der Algorithmus den Baum aufbaut.[16] Nachdem der Algorithmus ein Modell gebildet hat, werden die

[16] Bei Entscheidungsbäumen werden durch die Attribute Verzweigungen gebildet. Eine solche Verzweigung kann durch das Attribut ‚Geschlecht' gebildet werden. Ein anderes Attribut wäre beispielsweise, ob eine Person in der

Testdaten verwendet, um dessen Güte zu prüfen. Wird diese positiv bewertet, kann der Algorithmus auf neue Fälle angewandt werden.

Entscheidungsbäume können zu unterschiedlichen Zwecken eingesetzt werden; etwa, ob ein eingelieferter Patient auf die Intensivstation verlegt werden soll oder nicht. Entscheidungsbäume können zwar (je nach Anzahl der Attribute und Attributwerte) ungemein komplex werden. Insbesondere ist es schwer zu bestimmen, ob ein gelerntes Modell das beste Modell ist. Ein Vorzug ist jedoch, dass die Klassifikation und damit die Entscheidung von Personen auch im Detail gut nachvollzogen werden kann.

Dies ist anders bei neuronalen Netzen. Betrachten wir dazu ein einfaches Beispiel: ein neuronales Netz, welches handschriftlich notierte Ziffern erkennen soll. Ein solches Netz wird in unserem einfachen Beispiel als ein „multilayer perceptron" aufgebaut. Die erste Schicht enthält z.B. 784 Neuronen, welche einem Gitter mit der Auflösung von 28x28 Pixeln entspricht. Dieses Gitter wird unter die handschriftlich notierte Ziffer gelegt, so dass für jede Box der jeweilige Licht- und Schattenwert bestimmt werden kann. Jedes der 784 Neuronen weist daher einen Wert zwischen 0 und 1 auf; dieser Wert entspricht dem Farbwert (von 0 für weiß bis 1 für schwarz). Die letzte Schicht weist 10 Neuronen auf, denen die zu erkennende Zahlenwerte (0-9) entsprechen. Diese letzte Schichte kann ebenfalls wieder Werte zwischen 0 und 1 aufweisen; dem entspricht die Wahrscheinlichkeit, mit der der Zahlenwert der handschriftlichen Ziffer (von 0-9) korrekt erkannt wurde. Dazwischen finden sich, in diesem simplen Beispiel, zwei so genannte „hidden layers" mit jeweils 16 Neuronen. Jedes Neuron der einen Schicht ist mit allen Neuronen der nächsten Schicht verbunden. Zudem werden so genannte Gewichte und Biaswerte eingeführt.[17] In diesem Netz ergeben sich so 13.002 Parameter. Der Lernvorgang besteht darin, *die Werte für die Parameter so zu verändern*, dass die handschriftlichen Ziffern bestmöglich in ihren Zahlenwerten erkannt werden. Dafür werden wiederum Trainingsdaten verwendet, bei denen der richtige Zahlenwert annotiert ist; der lernende Algorithmus kann so entsprechend der Abweichungen sein Modell verbessern, bis es den richtigen Zahlenwert erkennt (überwachtes Lernen).

Der Entscheidungsbaum und das neuronale Netz unterscheiden sich eklatant in einem Punkt: der Nachvollziehbarkeit im Detail. Zwar können, wie gesagt, auch Entscheidungsbäume komplex werden. Aber das neuronale Netz ist einer Weise mathematisch verfasst, welche die Einsicht in die detaillierte Funktionsweise stark limitiert. Entscheidend ist dabei (1), dass die Gewichte und Biaswerte nur eine

---

Vergangenheit einen ‚Kredit beglichen' hat. Der Entscheidungsbaum könnte im ersten Knoten also entweder mit dem Attribut ‚Geschlecht' (männlich/weiblich) oder mit dem Attribut ‚Kredit beglichen (ja/nein) beginnen. Je nachdem, mit welchem Attribut der Entscheidungsbaum beginnt, werden unterschiedlich viele Personen auf die jeweiligen Zweige verteilt. Die jeweiligen Attribute können daher einen unterschiedlich großen Informationswert haben und so etwa die erforderliche Rechenzeit beeinflussen; außerdem könnte die Prognosequalität dadurch betroffen sein. Zum Beispiel mag die Schufa-Auskunft in Bezug darauf, ob eine Kundin ihren Kredit beglich oder nicht, die Daten besser zu Beginn (am ersten Knoten) teilen als das Geschlecht und daher einen größeren Informationswert haben.

[17] Dazu wird eine sogenannte Kostenfunktion verwendet. Es ist an dieser Stelle nicht möglich, aber auch nicht nötig, auf die Gründe hierfür einzugehen. Wer sich dafür interessiert, dem sei neben der Literatur zu dem Thema (hier wurden bereits einige Titel genannt), ein Videotutorial empfohlen, von dem das hier dargestellte Beispiel stammt. Zu finden ist dieses auf https://www.patreon.com/3blue1brown, wo drei Videos zum Deep Learning verfügbar sind.

sehr abstrakte Interpretation zulassen, sie stehen zunächst für nichts anderes als dafür, Gewichte oder Biaswerte zu sein; (2) die Anzahl der Parameter (hier 13.002) begrenzt die Möglichkeit, ihren jeweiligen Einfluss zu verstehen; (3) die Art der Verbindung, insbesondere ihre wechselseitige, nicht-lineare Abhängigkeit führen dazu, dass eine Einsicht in den Einfluss, den einzelne Elemente haben, in der Regel nur in geringem Maße möglich ist.[18] Mit anderen Worten: Das erlernte neuronale Modell ist nicht oder nur in Grenzen mathematisch verstehbar.

Um diese Grenze einer mathematischen Einsicht in das neuronale Modell zu erläutern, können wir uns einer Überlegung des Physikers Richard Feynman zuwenden. Feynman hatte für den Kontext der Physik einen Vorschlag unterbreitet, was es heißt, eine Gleichung mathematisch zu verstehen:

> 'I understand what an equation means if I have a way of figuring out.' So if we have a way of knowing what should happen in given circumstances without actually solving the equations, then we ›understand‹ the equations, as applied to these circumstances. (Feynman 2010, 2-1)[19]

Mathematisches Verstehen erläutert sich für Feynman also am Unterschied von Einsicht und Ausrechnen. Eine Gleichung wird verstanden, wenn eine Einsicht in die Abhängigkeit von Elementen eines Gleichungssystems gewonnen wird. Man erkennt dann, wie ein oder mehrere Elemente das globale Verhalten der Gleichung bestimmen. In diesem Fall muss die Gleichung nicht mehr ausgerechnet werden, um ihre Dynamik zu antizipieren.

In diesem Sinne sind neuronale Netze nur in einem geringen Maße verstehbar; und zwar auch für Expertinnen.[20] Mit anderen Worten: Sie sind ‚epistemisch opak'. Der Begriff epistemischer Opazität wurde von dem Philosophen Paul Humphreys eingeführt, um eine methodische Neuartigkeit von Computersimulationen zu markieren. Computersimulationen sind opak, sofern nicht alle methodisch an ihnen relevanten Elemente einsichtig nachvollzogen werden können.[21] Passenderweise erläutert Humphreys seine Überlegungen aber im Kontext von Beispielen aus der KI-Forschung. Epistemische Opazität ist in diesem Sinne methodische Opazität.

Die soziale Epistemologie und die Technikphilosophie haben gezeigt, dass die Sozialität von Wissenschaft und ihre technische Verfasstheit die Einsicht, welche einzelne Wissenschaftlerinnen haben, begrenzt. Diese lassen sich als *soziale* und als *technische* Opazität bezeichnen, welche aspektual unterschieden sind: Die soziale Arbeitsteilung in den Wissenschaften führt dazu, dass

---

[18] Entsprechend besteht ein Lernschritt nicht darin, dass ein einzelner Parameter angepasst wird, vielmehr werden in einem Lernschritt in der Regel alle Parameter zugleich adaptiert.

[19] Johannes Lenhard hat auf diese wichtige Passage aufmerksam gemacht. Vgl. dazu Lenhard 2015, S. 99.

[20] Die Argumentation erfolgt hier mit Blick auf neuronale Netze im Unterschied zu Entscheidungsbäumen, da der Kontrast in diesem Vergleich am stärksten ausfällt. In unterschiedlichem Maße weisen jedoch Lernstrategien ähnliche Opazitätseffekte auf. Dieser Punkt wird auch vonseiten der Informatik geteilt. Ein kompensatorischer Versuch ist, dass lernende Algorithmen ihren Lernprozess Personen in Form eines Dialogs erläutern. Siehe dazu Irving et al. 2018.

[21] Vgl. Humphreys 2004, 2009 – Epistemische Opazität betrifft daher nicht die Intransparenz der Natur als dem klassischen Problem der Epistemologie, sondern sie ist methodische Opazität. Da Wissenschaftlichkeit primär auf Methodizität basiert, ist mit der Behauptung von epistemischer Opazität eine starke These über die Veränderung von Wissenschaft verbunden.

Wissenschaftlerinnen in ihrer Expertise füreinander wie eine Black Box erscheinen. Die Expertise wird jedoch in der Regel in technischer Form weitergegeben (als Daten, als Algorithmus, als Modell, als mathematische Technik etc.). Technik entlastet von der Notwendigkeit des Nachvollzugs; die Black Box kann geschlossen weitergegeben und verwendet werden. Mit den Überlegungen zu den Grenzen mathematischen Verständnisses tritt nun aber eine neuartige Form auf: *mathematische* Opazität.[22] Anders als bei sozialer und technischer ist bei mathematischer Opazität nicht ausgemacht, dass es Experten gibt oder, beim gegenwärtigen Stand der Mathematik, geben kann, welche in der Lage sind, sie aufzuhellen. Dies bedeutet zwar nicht, dass sich Ergebnisse oder Verfahren nicht rechtfertigen lassen, aber dass die Rechtfertigung nicht durch Einsicht erfolgen kann. Kurz: Im Bereich der Computersimulation wie auch im Bereich avancierter maschineller Lernverfahren wie den neuronalen Netzen können häufig nur externalistische (Reliabilität), aber keine internalistischen Rechtfertigungen (Einsicht, Beweis) gegeben werden.

3. Güte und Angemessenheit von Gründen

Die bisherige Argumentation sollte zeigen: Lernende Algorithmen werden zur Entscheidungsfindung verwendet. Sie sollen entweder eigenständig Entscheidungen treffen oder Personen unterstützen, die richtige Entscheidung zu treffen. Entscheidungen sind *normativ* auf Gründe bezogen. Internalistisch sind die Gründe von Entscheidungssystemen jedoch entzogen. Der internalistische Entzug der Gründe verfehlte daher eine wesentliche normative Anforderung an Entscheidungen. Dass nach ihren Gründen gefragt werden können muss, bedeutet, dass eine Antwort prinzipiell gegeben werden können muss. Die internalistische Leerstelle würde daher eine wesentliche Norm, unter der wir Entscheidungen betrachten, verletzen: das Gründe-geben-Können.

Die Gründe für Entscheidungen von lernenden Systemen mögen zwar internalistisch verdeckt bleiben; das bedeutet jedoch nicht, dass sie keine Gründe für ihre Entscheidung bieten, nämlich externalistische. Es handelt sich um Reliabilitätsargumente. Der Algorithmus kommt zu einer probabilistischen Einschätzung. Gründe für Entscheidungen lassen sich bewerten. Diese Bewertung kann mit Blick auf Verlässlichkeitsgründe in zumindest zweierlei Hinsicht erfolgen: Es lässt sich zum einen fragen, ob das Maß an Reliabilität ausreichend sei, sodass die Entscheidung in diesem Sinne gut begründet ist. Zum anderen kann gefragt werden, ob und wann Verlässlichkeitsargumente die richtige Art von Gründen für Entscheidungen darstellen:

(1) Wann bieten uns Reliabilitätsschlüsse *ausreichend gute Gründe*, um auf ihrer Grundlage eine Entscheidung zu treffen? In Frage steht dann die Güte der Gründe.
(2) Wann bieten Reliabilitätsschlüsse die *richtige Art der Gründe*, um Entscheidungen durch sie begründen? In Frage steht dann die Angemessenheit der Gründe.

Ich halte beide Fragen für zentral. Üblicherweise erfolgt die Kritik von (lernenden) Algorithmen unter dem Gesichtspunkt des Grades an Verlässlichkeit, also der Güte der Gründe. Insbesondere die

[22] Zur Unterscheidung von sozialer, technischer und mathematischer Opazität vgl. Kaminski et al. 2018; Kaminski 2018.

wertvollen Studien zum ‚machine bias' stehen in dieser Perspektive.[23] Mir scheint jedoch, dass die Frage nach der Angemessenheit der Gründe ebenso bedeutsam ist. Im Folgenden werde ich daher mein Augenmerk vor allem auf diese Frage richten.

Dazu kehre ich zu einem der eingangs genannten Beispiele zurück: In einem EU-Pilotprojekt wird gegenwärtig ein System zur Bewertung der Vertrauenswürdigkeit von einreisenden Personen aus Nicht-EU-Staaten entwickelt und getestet: „The iBorderCtrl system unifies different interdisciplinary modules and converges into an overall system to speed up the procedure of crossing the borders especially for bona fide travellers."[24] Dazu gehören biometrische Module (u.a. Fingerabdrücke), ein Modul zur Gesichtserkennung, eine algorithmenbasierte Dokumentenprüfung, ferner ein Modul namens „External Legacy and Social interfaces system (ELSI)", das auch Informationen aus sozialen Netzwerken über die Person sucht. Mein primäres Interesse gilt dem „Automatic Deception Detection System (ADDS)". Über die Funktionsweise und Zielsetzung des Systems heißt es:

> [It] performs, controls and assesses the pre-registration interview by sequencing a series of questions posed to travellers by an Avatar. ADDS quantifies the probability of deceit in interviews by analysing interviewees non-verbal micro expressions. This, coupled with an avatar, moves this novel approach to deception detection to the pre-registration phase resulting in its deployment without an impact to the time spend at the border crossing by the traveller. The avatar also allows for a consistent and controllable stimuli across interviews in terms of both the verbal and non-verbal from the direction of the avatar agent to the traveller personalized to gender and language of the traveller, reducing variability compared to human interviewers and potentially improving the accuracy of the system.

Die Person, deren Vertrauenswürdigkeit geprüft werden soll, wird also von einem Avatar interviewt. Dabei werden Mikroexpressionen, insbesondere minimale Regungen im Gesicht, von Sensoren erfasst. Ein Modell schließt daraufhin auf die Wahrscheinlichkeit von Täuschungen. Das Modell wird von lernenden Algorithmen gebildet. Die Forscherinnen und Forscher haben dazu in Zusammenarbeit mit Psychologinnen und Psychologen mögliche Indikatoren im nonverbalen Verhalten von Personen, welche andere zu täuschen versuchen, identifiziert. Daraufhin haben sie ein maschinelles Lernsystem mittels eines Datensets trainiert, Personen, die täuschen, anhand nonverbaler Verhaltensweisen zu erkennen.

Die Rechtfertigung der Entscheidung (hier: ob eine Person vertrauenswürdig ist) lässt sich mit Blick auf die beiden genannten Fragen prüfen. Es kann erstens die *Güte der Gründe* für Entscheidungen betrachtet werden. Die Güte der Gründe wird anhand der Reliabilität des Systems bestimmt: *Wie verlässlich ist das Modell?* Zweitens kann die *Angemessenheit der Gründe* untersucht werden: Handelt es sich um die richtige Art von Gründen? Im ersten Fall können Entscheidungen schlecht begründet sein, dann nämlich,

---

[23] In Mittelstadt et al. 2016 findet sich eine Übersicht zu Studien, welche vor allem unter Gesichtspunkt der Verlässlichkeit solche Systeme betrachten.
[24] Die folgenden Zitate stammen aus der Beschreibung des Technical Framework von iBorderCtrl, zu finden auf der Projektwebsite: https://www.iborderctrl.eu/Technical-Framework

wenn das System unverlässlich ist. Im zweiten Fall kann das System zwar verlässlich sein, aber es handelt sich um die falsche Art von Gründen.

Im Folgenden geht es mir primär um die Frage nach der Angemessenheit von Verlässlichkeitsargumenten im Kontext maschineller Entscheidungen. Daher nehmen wir an, ADDS würde verlässlich in dem Sinne sein, dass die Detektionsrate bei 95% liegen würde, was ein (unrealistisch) hoher Wert wäre.[25] Wir nehmen ferner an, dass Personen weniger verlässlich darin wären, täuschende Personen zu erkennen. ADDS würde daher in dem spezifizierten und damit eingeschränkten Sinne *gut* begründete Entscheidungen darüber treffen, ob eine Person vertrauenswürdig ist. Aber wäre das verlässlich genug? Und insbesondere wäre es eine *angemessen* begründete Entscheidung? Dies wird in drei Schritten untersucht. Und zwar jeweils mittels eines Gedankenexperiments, welches auf einer Analogie beruht.

1. Individuum und Verlässlichkeitsgründe: Stellen wir uns folgende Situation vor: Eine eines Verbrechens beschuldigte Person erscheint vor Gericht. Es gibt keine zwingenden Beweise, dass die beschuldigte Person die ihr zur Last gelegte Tat begangen hat. Doch der Richter verfügt über eine Statistik. Diese gibt an, wie viele der angeklagten Personen tatsächlich die Tat begangen haben. Wie diese Statistik erstellt worden ist, braucht uns hierbei nicht zu kümmern. Im Gegenteil: Dass wir keine Einsicht in das Zustandekommen der Statistik haben, entspricht der methodischen Opazität lernender Algorithmen.

Der Richter trifft in dieser Situation eine Entscheidung auf statistischer Grundlage. Er schaut in der Statistik nach, die angibt, dass 95% der angeklagten Personen (sei es generell, sei es dieser Gruppe) schuldig sind. Wir nehmen ferner an, dass diese Statistik korrekt ist. Sie bietet insofern eine relativ verlässliche Grundlage.

Gleichwohl, ich würde den Schuldspruch als nicht gerechtfertigt erachten. Warum? Eine Statistik bietet einen Überblick über eine Gruppe von Entitäten, in Bezug auf diese *Gruppe* kann sie die Basis für Prognosen sein. Diese Prognosen haben Gültigkeit nur in Bezug auf die Gruppe, nicht in Bezug auf einen singulären Fall. Daher sind die statistischen Schlüsse des Richters in Bezug auf die Vertrauenswürdigkeit des jeweilig singulären Falls unangemessen. Sie sind gültig nur als Aussagen über Elemente *der Gruppe*.

Mir erscheint die Situation analog zum KI-Lügendetektor zu sein. Die Statistik bietet reliable Gründe für Schlüsse auf die Schuld von angeklagten Personen an. Wir wissen nicht, wie die Inferenz zustande gekommen ist, jedoch dass sie korrekt ist. Sie ist methodisch opak, aber verlässlich. Denn tatsächlich sind 95% der angeklagten Personen schuldig. Gleichwohl erscheint das Urteil des Richters auf der statistischen Grundlage als unangemessen. Der KI-Lügendetektor bietet ebenfalls ein statistisches Modell an, um wahrhaftige und täuschende Personen zu bestimmen. Ist die Urteilsbegründung des Richters unangemessen, so ist es auch der KI-Lügendetektor.

[25] In parallelen Studien wurden Interviews mit vertrauenswürdigen und täuschenden Probanden *simuliert*. Vgl. OrShea et al. 2018; Rothwell et al. 2007. Damit steht die Validität von Training und Test in Frage. Jemand, der im Rahmen einer Studie gebeten wird zu lügen, mag sich anders verhalten als jemand, für den die Einreise in die EU eine hohe, vielleicht sogar existentielle Bedeutung hat.

Das bedeutet nicht, dass Verlässlichkeitsüberlegungen stets unangemessen wären. So wäre es beispielsweise eigenartig, wenn von einer Person gefordert würde, sie müsste nachvollziehen können, warum ein Thermometer so funktioniert, wie es funktioniert, um es zu verwenden. Oder warum ihr Auto fährt. Es genügt, wenn sie sich darauf verlässt und verlassen kann, weil es sich wiederholt bewährt hat. In Fall des juristischen Urteils ist, anders als beim Start des verlässlichen Autos, der singuläre Fall entscheidend. Es geht um die Schuld *dieses* Individuums. Ebenso wäre es eigenartig von einem Objekterkennungssystem zu sagen, dass es unangemessen wäre, wenn es Objekt hochverlässlich identifiziert (etwa um Informationen für einen Reisenden über seine Umgebung zu bieten). Würde das Objekterkennungssystem jedoch verwendet, um über die Schuld des Angeklagten zu bestimmen, würden die gleichen Zweifel auftreten. Denn es handelte sich wiederum um probabilistische Aussagen über eine Gruppe von Elementen. In Frage steht aber die Schuld dieses Individuums.

Daher würde auch eine Verbesserung der Statistik die Angemessenheit ihrer Verwendung nicht berühren. Nehmen wir an, dass 98% der Personen, welche vor Gericht erscheinen, schuldig sind. Meines Erachtens würde es die Unangemessenheit nicht verringern, eine solche Statistik zur Rechtfertigung des Schuldspruchs zu verwenden. Dabei würde es sich im oben genannten Sinne um einen guten Grund handeln. Das Problem bleibt: Individuum und Statistik stehen in einem kategorialen Missverhältnis zueinander. Der statistische Schluss ist unangemessen für die Feststellung einer individuellen Schuld.

2. Erstaunliche Fehlleistungen: Ich komme zum zweiten Gedankenexperiment. In diesem ist die Polizei in der Lage mit einer hohen Verlässlichkeit von 99,5% zu bestimmen, wer eine Tat begangen hat. Ferner wissen wir, dass die Resultate korrekt sind in 99,5% der Fälle. Wir sind nicht in der Lage nachvollzuziehen, wie sie zu ihren Ermittlungsergebnissen kommt. Aber: Die 0,5% der Fälle, in denen die Polizei zu einem Fehlurteil kommt, sind erstaunliche Fehlleistungen. Jeder, der den Fall nachvollziehen könnte, würde auf einen Blick sehen, dass der Schluss auf den Täter völlig daneben geht. Der Fehlschluss gleicht einer unglücklichen Lotterie: Irgendeine unschuldige Person wird der Tat bezichtigt. Das Problem ist jedoch, dass die Schlüsse der Polizei nicht nachvollziehbar und daher auch nicht kritisierbar sowie revidierbar sind. Wir wissen nur, dass es in 0,5% der Schlüsse zu erstaunlichen Fehlleistungen kommt. Wir wissen jedoch nicht, welche das sind, und wir sind auch nicht in der Lage, diese im Nachhinein aufzuklären. Mir erscheint es, eine solche Situation würde die Frage aufwerfen, ob die Methode angemessen ist, um Schuldige zu bestimmen.
Wiederum handelt es sich um eine Analogie. Denn es gibt Hinweise darauf, dass lernende Algorithmen solche erstaunlichen Fehlleistungen vollziehen. In der jüngsten Zeit erregten Studien im Bereich der Objekterkennung durch neuronale Netze Aufmerksamkeit. In den Studien zeigten die algorithmischen Modelle eine hohe Verlässlichkeit darin, Objekte richtig zu klassifizieren. Doch wenn die Bilder in einer Weise verändert wurden, die Personen weiterhin keinerlei Schwierigkeit bereitet, die visuell dargebotenen Objekte korrekt zu identifizieren, zeigten die algorithmischen Modelle erstaunliche Fehlschlüsse. In einer Studie wurden feine Strukturen, die für Personen kaum sichtbar waren, in die Bilder eingeführt. Das zuvor korrekt als Waschmaschine klassifizierte Objekt wurde nun als Safe oder Lautsprecher fehlklassifiziert.[26] In einer anderen Studie mit dem schönen Titel „The Elephant in the

[26] Kurakin et al. 2017.

Room“ wurde in ein Bild, auf dem eine Person, ein Notebook, eine Tasse, Bücher und ein Stuhl zu sehen war, ein zweites Bild eingefügt: ein Elefant.[27] Wurden zuvor alle Objekte korrekt klassifiziert, wurden, nachdem der Elefant eingefügt worden war, einzelne Objekte nicht mehr oder falsch erkannt. Das geschah jedoch nicht immer, sondern hing von geringen Positionsveränderungen des Elefanten ab. Erstaunlich daran war ferner, dass der Elefant in keiner räumlichen Nähe zu den nun nicht erkannten oder fehlerkannten Objekten stand. Diese Studien mit sogenannten „adversarial images“ wurden von den Entwicklern unternommen, um zu prüfen, ob lernende Algorithmen überlistet und manipuliert werden können. Doch die Ergebnisse der Studien sind noch in anderer Weise aufschlussreich: Sie demonstrieren, dass die Weise, in der neuronale Netze ‚lernen‘ und ‚erkennen‘, grundlegend anders ist, als die von Personen. Die Fehlschlüsse sind deshalb erstaunlich, weil sie Personen kaum unterlaufen dürften. In den Studien können wir, da es sich um eine simple Bilderkennung handelt, leicht überprüfen, ob der Algorithmus korrekt klassifiziert. Wir sehen nach und wissen, was das richtige Ergebnis ist. Werden lernenden Algorithmen jedoch in Bereichen angewandt, in denen wir nicht über die Möglichkeit verfügen, ihre Leistung eigenständig, leicht und schnell zu überprüfen (Rückfallwahrscheinlichkeit, Bewerberauswahl, medizinische Diagnose, Vertrauenswürdigkeit von Personen etc.), können wir uns nicht mehr sicher sein, ob und wann erstaunliche Fehlleistungen vorkommen. Wenn es um die Beurteilung von Personen geht und eine Nachvollziehbarkeit und also Korrektur des Urteils ausgeschlossen ist, derartige Systeme jedoch krasse Ausreißer produzieren, dann erscheint es mir unangemessen, Verlässlichkeit zur Grundlage des Urteils zu machen. Das zeigt das Gedankenexperiment. Die Situation ist analog bei einem KI-Lügendetektor. Es steht in Frage, ob eine Nachvollziehbarkeit und Korrigierbarkeit des Urteils gegeben ist. Und es kann nicht ausgeschlossen werden, dass das System keine krassen Fehlleistungen hervorbringt, selbst wenn es in den allermeisten Fällen hochverlässlich funktioniert.

3. Die fehlende Verbindung von Tat und Person: Doch was wäre, wenn die Polizei 100% korrekt in der Identifikation des Täters wäre? Die vorherigen Probleme bestanden ja, weil die Systeme eine zwar extrem hohe Verlässlichkeit aufwiesen, aber nicht fehlerfrei waren. Wir nehmen nun an, dass keine Fehler mehr auftreten. Dass wir aber auch nicht nachvollziehen könnten, wie eine Person mit einer Tat verbunden ist bzw. wie die Polizei zu ihren Urteilen kommt. Dabei haben wir jedoch keinen Grund anzunehmen, dass es Fehlurteile gibt. In diesem Fall stellten sich weder das oben genannte Singularitätsproblem noch das der erstaunlichen Fehlleistungen. Mir scheint es aber, dass selbst im Falle einer einhundertprozentig verlässlichen Statistik ein Bedenken an einer solchen Rechtfertigungsbasis bliebe. Weil die Methode völlig opak bleibt, können wir die behauptete Schuld nicht in eine Verbindung zur beschuldigten Person bringen. Wir würden damit in eine verwirrende Situation geraten, in der wir einerseits aufgrund der absolut verlässlichen Methode von der Schuld der Person überzeugt sind und diese Schuld dennoch äußerlich[28] und gleichsam leer bliebe – weil die handelnde Person nicht im Zentrum der Tat, als Urheber und ‚Autor‘ des Geschehens erscheint. Wir können nicht nachvollziehen,

[27] Rosenfeld et al. 2018.

[28] Und zwar auf andere Weise äußerlich, als wenn etwa ein Fingerabdruck oder DNA eine Person überführt. Dann können wir nämlich eine Verbindung herstellen zwischen Person und der Tat, derer sie beschuldigt wird; selbst wenn wir nicht im Einzelnen wissen, wie diese Methoden funktionieren, haben wir eine Vorstellung davon, wie die Methode Person und Tat miteinander verbindet.

welche Verbindung zwischen der Tat und der Person bestünde, sondern sind nur mit der leeren Identifikation der Person konfrontiert.

Analog dazu wären KI-Lügendetektoren, die fehlerfrei funktionierten und schuldige Personen identifizierten, uns dabei jedoch keinen Aufschluss über die Beziehung zwischen Tat und Person geben könnten. Selbst wenn dann Zweifel an der Verlässlichkeit des Algorithmus nicht mehr gerechtfertigt wären, bliebe der Eindruck eines Mangels; diesen interpretiere ich so, dass zwar gute Gründe gegeben sind, sie aber gleichwohl unangemessen für ein Urteil erscheinen, in dem es um die Schuld einer Person geht.

Der Fall ist anders, wenn wir einen Fingerabdruck oder eine DNA-Analyse verwenden. Nehmen wir an, dass diese Methoden von den beteiligten Personen vor Gericht auch nicht im Detail nachvollzogen werden können; gleichwohl bietet die Methode eine Vorstellung von deren Verbindung: der Fingerabdruck verweist darauf, dass die Personen den Gegenstand, auf dem der Abdruck festgestellt wurde, berührt hat und je nach Kontext kann dadurch eine mögliche Verbindung zur Tat (etwa über den Tatort) rekonstruiert werden. Diese Verbindung zwischen Tat und Person wird aufgrund der methodischen Opazität der Lernstrategie nicht gewährt. Es spräche aus Verlässlichkeitsgründen nichts gegen die Verwendung des Systems, aber der Mangel verweist m.E. darauf, dass die bloße Identifikation der Schuld einer Person sie in keinen nachvollziehbaren Zusammenhang mit der Tat bringt.

4. Mögliche Einwände

Die bisherige Argumentation lässt sich knapp, wie folgt, skizzieren: Entscheidungen beruhen auf Gründen; lernende Algorithmen, welche zur Entscheidungsfindung eingesetzt werden, sind zum Teil epistemisch und praktisch opak. Daher bieten sie primär reliabilistische Rechtfertigungen an. Reliabilistische Rechtfertigungen mögen nicht in allen Kontexten angemessen sein; selbst wenn die Güte der Verlässlichkeit hoch ist, könnte es sich um die falsche Art von Gründen handeln. Die Opazität stellt dabei in Frage, inwiefern es möglich ist, die Moralfähigkeit solcher Systeme zu beurteilen.[29]

Das Beispiel, welches ich auswählte, sollte einen möglichen Fall vorstellen, in dem reliabilistische Argumente unangemessen sind. Ich behaupte nicht, dass sie in allen Fällen unangemessen sind. Reliabilistische Überlegungen sind praktisch unverzichtbar und theoretisch in vielen Kontexten selbst wiederum rechtfertigbar. Man denke nur an medizinische Diagnosen mittels Laboruntersuchungen. Der behandelnde Arzt muss dafür die Methode nicht im Detail nachvollziehen können. Doch es ist für mich fraglich, ob das gleiche gilt, wenn ein medizinisches Diagnosesystem, das auf lernenden Algorithmen beruht, Entscheidungen trifft; sofern nämlich solche Entscheidungen mit Blick auf fachliche, ökonomische und rechtliche Aspekte getroffen werden. Ist die Entscheidungsfindung aufgrund der Opazität nicht nachvollziehbar, steht für mich in Frage, ob meta-stufig der Einsatz solcher Systeme rechtfertigbar ist; also die Entscheidung dafür, die Entscheidungsfindung an solche zu delegieren. In welchen Kontexten und unter welchen Bedingungen es rechtfertigbar ist, lernende Algorithmen zur Entscheidungsfindung einzusetzen, ist meines Erachtens ein relativ neues Forschungsgebiet, bei dem wir am Anfang stehen.

[29] Vgl. Hubig, Richter 2015.

Zwei mögliche Erwiderungen auf die Problematisierungen sind naheliegend. Der erste Einwand könnte die Funktion der Systeme betreffen. Es wird häufig betont, dass sie lediglich zur *Unterstützung* bei der Entscheidungsfindung eingesetzt werden sollen, aber die *Entscheidung nicht an sie delegiert* werden soll. Das mag *normativ* zutreffend sein, aber es bedeutet nicht, dass der *faktische* Einsatz der Systeme auch so erfolgt. Systeme, welche die Entscheidungsfindung nur unterstützen sollen, können faktisch zunehmend in die Rolle geraten, die Entscheidung zu treffen. Untersuchungen zur Verwendung von Systemen, welche die Rückfallwahrscheinlichkeit von Straftätern in den USA prognostizieren sollen, deuten in diese Richtung.[30] Auch ist aus Untersuchungen von medizinischen *Assistenz*systemen bekannt, dass das medizinische Personal, welches von den Empfehlungen abweicht, unter Rechtfertigungsdruck geraten kann.[31]

Dass solche Systeme, welche zur Unterstützung eingesetzt werden, eine Tendenz dazu haben, dass die Entscheidung an sie delegiert wird, mag auch mit ihrer technischen Anmutung zusammenhängen. Die Systeme beruhen auf technisch-mathematischen Verfahren, die eine größere Transparenz, Nachvollziehbarkeit, Verlässlichkeit und damit etwa auch Gerechtigkeit versprechen. Die hier vorgeführte Problematisierung von Entscheidungssystemen setzt jedoch an der methodischen Opazität lernender Algorithmen an.

Ein naheliegender zweiter Einwand gegen die Kritik an solchen Systemen besteht folglich im Verweis auf die Opazität von Personen. Die Entscheidungen, welche von IT-Systemen auf der Basis lernender Algorithmen getroffen werden sollen, werden andernfalls von Personen getroffen. Personen sind aber keinesfalls weniger opak in ihrer Entscheidungsfindung als Algorithmen. Im Gegenteil weisen Personen die Fähigkeit auf, über die Gründe ihrer Entscheidung zu lügen. Algorithmen dagegen sind implementiert und damit zumindest teilweise nachvollziehbar. Wenn auch bei avancierten Lernstrategien nicht im Detail nachvollzogen werden kann, wie es zu einer Entscheidung kommt, so ist doch zumindest das Lernprinzip bekannt. Dieser Einwand verfehlt, worum es mir geht, macht meinen Punkt dadurch aber sichtbar: Personen mögen über ihre Gründe, welche die faktische Basis ihrer Entscheidung gewesen sind, lügen. Oder die Gründe mögen ihnen unklar sein; sie mögen Schwierigkeiten haben, diese zu formulieren und zu erklären. Sie sind jedoch in dem, was sie sagen, selbst in dem, was sie nur vage und undeutlich sagen oder mutmaßlich verschweigen, moralisch befrag- und kritisierbar. Die Entscheidungen von lernenden Algorithmen bleiben in dieser Hinsicht der Einsicht entzogen, weil sie epistemisch unverständlich sind. Verlässlichkeitsgründe stellen, in zumindest einigen Kontexten, nur einen ungeeigneten Ersatz für internalistische Gründe dar. Dass ein Algorithmus in den allermeisten Fällen richtig liegt, ist kein Grund, der gegenüber einem Individuum geltend gemacht werden kann. Ein lernender Algorithmus verhält sich hier sonst wie ein Richter, der eine statistische Aussage über die Schuld eines Angeklagten zur Begründung anführt, um ein Urteil über die Schuld dieses Angeklagten zu sprechen.

Lernende Algorithmen, die epistemisch opak sind, verändern daher die Moralfähigkeit von Entscheidungen – nicht, weil die Gründe unzureichend sind, sondern weil die Form der Begründung für einige Entscheidungen nicht adäquat ist. Dies gilt selbst dann, wenn der Algorithmus keinen Fehler

---

[30] Vgl. dazu die Ergebnisse der Studie von ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (Datum des Aufrufs 10.11.2018)
[31] Vgl. Manzei 2018.

macht; etwa, wenn er einen Fall richtig prognostiziert. Das Problem ist nicht die Güte der Prognose, sondern die Form der Begründung, die kategorial inadäquat ist. Damit ist nicht gesagt, dass auf lernende Algorithmen verzichtet werden sollte. Ich habe keine Bilanzierung vorgenommen, sondern eine Veränderung im Sprachspiel der Begründung von Entscheidungen markieren wollen, die mir grundlegend zu sein scheint. In diesem Sinne ist damit die Forderung nach einer Ermöglichung der Ethik verbunden.[32]

Literaturverzeichnis

Alpaydın, Ethem (2016): Machine Learning. The New AI. Cambridge: The MIT Press.

Alpaydın, Ethem (2008): Maschinelles Lernen. München: Oldenbourg.

Bieri, Peter (2003): Das Handwerk der Freiheit. Über die Entdeckung des eigenen Willens. Lizenzausg. Frankfurt am Main: Fischer.

Bittner, Rüdiger (2005): Aus Gründen handeln. Berlin, New York: Walter de Gruyter.

Brenneis, Andreas (2019): „Ethik zwischen Front- und Back-End?!“. In: *Jahrbuch Technikphilosophie* 5, [im Druck].

Foerster, Heinz von (1993): Prinzipien der Selbstorganisation im sozialen und betriebswirtschaftlichen Bereich. In: Heinz von Foerster: Wissen und Gewissen. Franfurt am Main: Suhrkamp, S. 233–268.

Foerster, Heinz von (2002): Entdecken oder Erfinden. Wie läßt sich Verstehen verstehen? In: Heinz von Foerster, Ernst von Glasersfeld, Peter M. Hejl, Siegfried J. Schmidt und Paul Watzlawick (Hg.): Einführung in den Konstruktivismus. München: Piper (5), S. 60–67.

Foucault, Michel (1991): Die Ordnung der Dinge. Eine Archäologie der Humanwissenschaften. Frankfurt am Main: Suhrkamp.

Görz, Günther; Rollinger, Claus-Rainer; Schneeberger, Josef (Hg.) (2003): Handbuch der künstlichen Intelligenz. 4., korrigierte Aufl. München: Oldenbourg.

Halbig, Christoph (2007): Praktische Gründe und die Realität der Moral. Frankfurt am Main: Klostermann (Philosophische Abhandlungen, 94).

Harrach, Sebastian (2014): Neugierige Strukturvorschläge im maschinellen Lernen. Eine technikphilosophische Verortung. Bielefeld: Transcript (Edition panta rhei).

Hubig, Christoph (2006): Die Kunst des Möglichen I. Technikphilosophie als Reflexion der Medialität. Bielefeld: Transcript.

Hubig, Christoph (2008): Der technisch aufgerüstete Mensch –. Auswirkungen auf unser Menschenbild. In: Alexander Roßnagel, Tom Sommerlatte und Udo Winand (Hg.): Digitale Visionen. Zur Gestaltung allgegenwärtiger Informationstechnologien. Berlin, Heidelberg: Springer-Verlag (Springer-11774 /Dig. Serial]), S. 165–175.

[32] Vgl. hierfür Hubig und Richter 2015.

Hubig, Christoph; Harrach, Sebastian (2014): Transklassische Technik und Autonomie. In: Andreas Kaminski und Andreas Gelhard (Hg.): Zur Philosophie informeller Technisierung. Darmstadt: Wissenschaftliche Buchgesellschaft, S. 37–53.

Hubig, Christoph (2015): Die Kunst des Möglichen III. Macht der Technik. 1. Aufl. Bielefeld: transcript-Verl.

Hubig, Christoph; Richter, Philipp (2015): Technikethik als Ethik der Ermöglichung des Anwendungsbezuges. In: Regina Ammicht Quinn und Thomas Potthast (Hg.): Ethik in den Wissenschaften. 1 Konzept, 25 Jahre, 50 Perspektiven. Tübingen, S. 209–214.

Humphreys, Paul (2004): Extending ourselves. Computational science, empiricism, and scientific method. New York: Oxford University Press.

Humphreys, Paul (2009): The philosophical novelty of computer simulation methods. In: *Synthese* 169 (3), S. 615–626. DOI: 10.1007/s11229-008-9435-2.

Husserl, Edmund (1976): Die Krisis der europäischen Wissenschaften und die transzendentale Phänomenologie. Eine Einleitung in die phänomenologische Philosophie. 2. Aufl., photomech. Nachdr. Den Haag: Nijhoff (Husserliana, : gesammelte Werke / Edmund Husserl. Aufgrund des Nachlasses veröffentlicht vom Husserl-Archiv (Leuven) unter Leitung von Rudolf Bernet … ; Bd. 6).

Husserl, Edmund (1989): Fünf Aufsätze über Erneuerung. In: Edmund Husserl: Aufsätze und Vorträge (1922-1937). Dordrecht: Kluwer Acad. Publ, S. 3–124.

Irving, Geoffrey; Christiano, Paul; Amodei, Dario (2018): AI safety via debate. paper in progress. Online verfügbar unter https://www.researchgate.net/publication/324908324_AI_safety_via_debate.

Kaminski, Andreas (2010): Technik als Erwartung. Grundzüge einer allgemeinen Technikphilosophie. Bielefeld: Transcript.

Kaminski, Andreas (2013): Husserl: Die Krisis der europäischen Wissenschaften und die transzendentale Phänomenologie. In: Christoph Hubig, Alois Huning und Günter Ropohl (Hg.): Nachdenken über Technik. Die Klassiker der Technikphilosophie und neuere Entwicklungen. Darmstädter Ausgabe. 3., neu bearb. u. erw. Aufl. Berlin: Edition Sigma, S. 186–192.

Kaminski, Andreas (2014): Lernende Maschinen: naturalisiert, transklassisch, nichttrivial? Ein Analysemodell ihrer informellen Wirkungsweise. In: Andreas Kaminski und Andreas Gelhard (Hg.): Zur Philosophie informeller Technisierung. Darmstadt: Wissenschaftliche Buchgesellschaft, S. 58–81.

Kaminski, Andreas (2018): Der Erfolg der Modellierung und das Ende der Modelle. Epistemische Opazität in der Computersimulation. In: Andreas Brenneis, Oliver Honer, Sina Keesser und Silke Vetter-Schultheiß (Hg.): Technik - Macht - Raum. Das Topologische Manifest im Kontext interdisziplinärer Studien. Wiesbaden: Springer.

Kaminski, Andreas; Glass, Colin W. (2018): Interaktion mit lernenden Maschinen. In: Kevin Liggieri und Oliver Müller (Hg.): Mensch-Maschine-Interaktion. Handbuch zu Geschichte – Kultur – Ethik. 1. Auflage 2018. Stuttgart, [Forthcoming].

Kaminski, Andreas; Resch, Michael; Küster, Uwe (2018): Mathematische Opazität. Reproduzierbarkeit in der Computersimulation. In: *Jahrbuch Technikphilosophie* 4, 253-277.

Korsgaard, Christine M. (2009): Self-Constitution. Agency, identity, and integrity. Oxford, New York: Oxford University Press.

Kurakin, Alexey; Goodfellow, Ian; Bengio, Samy (2017): Adversarial examples in the physical world. Under review as a conference paper at ICLR 2017.

Lämmel, Uwe; Cleve, Jürgen (2008): Künstliche Intelligenz. Mit … 50 Tabellen, 43 Beispielen, 208 Aufgaben, 89 Kontrollfragen und Referatsthemen. 3., neu bearb. Aufl. München: Hanser.

Lenhard, Johannes (2015): Mit allem rechnen - zur Philosophie der Computersimulation. Berlin/Boston: de Gruyter (Ideen & Argumente).

Manzei, Alexandra (2018): Sind Standards objektiv und neutral? Zur Ambivalenz von Standardisierungsprozessen in der Medizin. In: Sebastian Klinke und Martina Kadmon (Hg.): Ärztliche Tätigkeit im 21. Jahrhundert - Profession oder Dienstleistung. Unter Mitarbeit von Günther Jonitz und Hans Michael Piper. Berlin, Germany: Springer (Springer-Lehrbuch), S. 208–229.

Mittelstadt, Brent Daniel; Allo, Patrick; Taddeo, Mariarosaria; Wachter, Sandra; Floridi, Luciano (2016): The ethics of algorithms: Mapping the debate. In: *Big Data & Society* 3 (2), 205395171667967.

OrShea, James; Crockett, Keeley; Khan, Wasiq; Kindynis, Philippos; Antoniades, Athos; Boultadakis, Georgios (2018): Intelligent Deception Detection through Machine Based Interviewing. In: 2018 International Joint Conference on Neural Networks (IJCNN). 2018 proceedings. International Joint Conference on Neural Networks (IJCNN). Rio de Janeiro, 7/8/2018 - 7/13/2018. IJCNN; IEEE Computational Intelligence Society; International Neural Network Society; Institute of Electrical and Electronics Engineers; International Joint Conference on Neural Networks; IEEE World Congress on Computational Intelligence (IEEE WCCI). Piscataway, NJ, USA: IEEE, S. 1–8.

Otto, Philipp; Gräf, Eike (Hg.) (2018): 3TH1CS. Die Ethik der digitalen Zeit. Bonn: Bundeszentrale für politische Bildung.

Rosenfeld, Amir; Zemel, Richard; Tsotsos, John K. (2018): The Elephant in the Room. Online available http://arxiv.org/pdf/1808.03305v1.

Rothwell, Janet; Bandar, Zuhair; O'Shea, James; McLean, David (2007): Charting the behavioural state of a person using a backpropagation neural network. In: Neural Computing & Application 16 (4-5), S. 327–339.

Russell, Stuart; Norvig, Peter (2007): Künstliche Intelligenz. Ein moderner Ansatz. 2. Aufl., [Nachdr.]. München: Pearson Studium.

Wiegerling, Klaus (2011): Philosophie intelligenter Welten. Paderborn: Fink.