7 Transparenz

In diesem Kapitel wird erarbeitet, wieso das Fehlen von Transparenz überhaupt ein unerwünschtes Ereignis wäre, warum dieses Fehlen eine verbotene Handlungsfolge wäre und wie man die Transparenz schützen bzw. fördern kann. Dafür ist es zuerst einmal notwendig zu definieren, was genau mit Transparenz im Kontext von Datenökosystemen gemeint ist – insbesondere wie Transparenz im Datenökosystem hergestellt werden kann. Ich zeige auf, dass das Vorgehen des Algorithmus bekannt sein muss, damit seine Ergebnisse auf Verzerrungen überprüft werden können. Zu diesem Zweck können Post-hoc Erklärungen von Algorithmen das Verhältnis zwischen Input und Output offengelegen und dadurch aufzeigen, warum der Algorithmus eine bestimmte Ausgabe erzeugt. Damit dieser Zusammenhang verständlich und das Ergebnis damit transparent ist, schlage ich vor, dass die Ergebnisse im zum Verständnis notwendigen Kontext dargestellt werden sollten. Die moralische Pflicht zur transparenten Darstellung algorithmischer Ergebnisse durch Datenverarbeitende leite ich aus Kants erster Formulierung des Kategorischen Imperativs ab. Daraus resultiert für Datennutzende ein Recht auf Transparenz, welches sich zum Teil schon in europäischen und deutschen Gesetzen wiederfinden lässt. Abschließend lassen sich aus der Pflicht zur Transparenz zwei administrative Anforderungen ableiten, welche über das bestehende Recht hinausgehen.

Transparenz ist das Gegenteil von Undurchsichtigkeit (opacity) und wird im Folgenden, wie bei Mittelstadt et al., als Zugänglichkeit und Verständlichkeit verstanden.[759] Bei Datennutzenden entsteht ein Risiko durch die fehlende Transparenz der algorithmischen Ergebnisse, weil es bei fehlender Transparenz möglich ist, dass den Datennutzenden verzerrte Ergebnisse zur Verfügung gestellt werden, auf deren Grundlage sie zu falschen Überzeugungen gelangen. Transparenz fehlt deshalb, weil Datennutzenden sowohl der Zugang zu als

759 Mittelstadt et al. 2016:6

auch das Verständnis von Informationen über die Entstehung der Ergebnisse fehlt.[760] Dies ist möglich auf Grund der epistemischen Asymmetrie im Datenökosystem. Denn Datenverarbeitende sind zu einem gewissen Grad dazu in der Lage, die für Datennutzende verfügbaren Beweise in Bezug darauf, wie das Ergebnis zustande gekommen ist, zu kontrollieren. Es handelt sich dabei nicht um eine vollkommene epistemische Dominanz, aber es besteht eine gewisse Abhängigkeit der Datennutzenden von den Datenverarbeitenden.[761] Auf Grund dieser Abhängigkeit besteht bei den Datennutzenden ein Interesse an Transparenz.

Damit die Datennutzenden die ihnen zur Verfügung gestellten Ergebnisse auf ihre Richtigkeit prüfen können und so möglichst nicht zu falschen Überzeugungen gelangen, benötigen sie Zugang zu und Verständnis von Informationen über die Entstehung der Ergebnisse. Diese Informationen können nur durch die Datenverarbeitenden zur Verfügung gestellt werden, aber auch die kennen auf Grund von der Nutzung komplizierter Algorithmen nicht unbedingt den gesamten Entstehungsprozess. In Bezug auf algorithmische Ergebnisse bedeutet Transparenz also, dass man den Mechanismus, nach dem der eingesetzte Algorithmus funktioniert, in gewisser Weise versteht.[762] Es ist also unklar, wie Transparenz im Datenökosystem hergestellt werden kann, es ist jedoch klar, dass Transparenz im Datenökosystem wichtig ist, um falsche Ergebnisse und darauf gegründete falsche Überzeugungen vermeiden zu können – dieses systematische Problem soll in diesem Kapitel vertieft und gelöst werden.

Da in Datenökosystemen viele Daten verarbeitet werden, nutzen Datenverarbeitende dafür oft Hilfsmittel in Form von Algorithmen. Hier gilt es zu unterschieden zwischen Wenn-Dann-Algorithmen und Black Box KI-Algorithmen (auch KI-Systeme genannt). Wenn-Dann-Algorithmen können als „White Box" bezeichnet werden, weil es sich um einen vorgegebenen Satz von Anweisungen handelt.[763] Solche Algorithmen sind transparent, da deren Ergebnisse problemlos zugänglich gemacht und verständlich dargestellt werden können. KI-Systeme hingegen sind nicht deterministisch – wenn sich ihre Ei-

760 Mittelstadt et al. 2016:6
761 Harris 2022:134, 137
762 Lipton 2018:42
763 Tutt 2017:107

genschaften dennoch leicht vorhersagen und erklären lassen, nennt man das „Grey Box". Sobald Algorithmen emergente Eigenschaften aufweisen, die es schwierig oder unmöglich machen, ihre Merkmale vorherzusagen oder zu erklären, handelt es sich um eine „Black Box".[764] So ein Black Box Algorithmus verbirgt Annahmen und Vereinfachungen. Aufgrund dieser Undurchsichtigkeit können Datennutzende die in den Algorithmus eingebettete Weltanschauung nicht erkennen und seine Glaubwürdigkeit nicht überprüfen. Black Box KI-Systeme sind also nicht inhärent transparent, da es schwierig bis unmöglich ist, die Entstehung ihrer Ergebnisse verständlich darzustellen. Da es darum geht herauszufinden, wie die Entstehung der Ergebnisse von KI-Systemen dennoch transparent dargestellt werden kann, ist im Folgenden mit Algorithmus immer ein Black Box KI-Algorithmus gemeint.

Das größte Problem der fehlenden Transparenz beim Einsatz von Algorithmen ist, dass ihre Ergebnisse, auch bezeichnet als Entscheidungen, nur mit einer gewissen Wahrscheinlichkeit richtig sind. Sie können falsch sein oder anders, als ein Mensch sie getroffen hätte. Insbesondere, wenn auf Grundlage einer solchen Entscheidung Klassifizierungen vorgenommen werden, ist eine andere Entscheidung, als ein Mensch sie getroffen hätte, nicht notwendigerweise schlecht. Denn Menschen haben Vorurteile, und es fällt ihnen oft schwer, objektive Entscheidungen zu fällen. Die Hoffnung wäre, dass Algorithmen objektiv entscheiden und damit Ungerechtigkeiten vermeiden. Aber das stimmt leider nicht, denn auch algorithmische Ergebnisse sind häufig von menschlichen Vorurteilen beeinflusst. Einerseits, weil Menschen zum Beispiel die Trainingsdaten vorklassifizieren sowie Schwellenwerte und Parameter anpassen und andererseits, weil die Trainingsdaten immer aus menschlichen Kontexten entstehen und deshalb auch menschliche Vorurteile widerspiegeln.[765] Auf diese Weise kann das Ergebnis auf Entscheidungskriterien basieren, die Verzerrungen aus bisherigen Entscheidungen reproduzieren, zum Beispiel bei dem Algorithmus, der von Amazon im Rekrutierungsprozess eingesetzt wurde. In der Vergangenheit waren es vor allem Männer, die in Amazons Bewerbungsverfahren erfolgreich waren. Dieses Muster hatte zur Folge, dass ein Entschei-

764 Tutt 2017:107
765 Burrell 2016:3

dungskriterium des Algorithmus das Geschlecht der Bewerbenden war. Deshalb wurden Bewerbungen von Frauen vom KI-System abgelehnt – selbst wenn das Geschlecht nicht direkt genannt wurde, sondern nur z.B. durch feminin konnotierte Hobbys erkennbar war.[766] Das KI-System wurde bei Amazon zum Identifizieren der besten Bewerbungen eingesetzt, bis anhand der Ergebnisse eine Verzerrung deutlich wurde.

Solch eine Verzerrung nennt sich im Englischen „Bias", sie kommt zustande, wenn bestehende Diskriminierungsstrukturen Eingang in die Trainingsdaten eines Algorithmus finden. So schreiben der Technologiephilosoph Bryce Goodman und der Informatiker Seth Flaxman: „Indeed, machine learning can reify existing patterns of discrimination — if they are found in the training data set, then by design an accurate classifier will reproduce them."[767] Wird der Begriff „Bias" im Kontext mit Computersystemen verwendet, dann ist laut der bereits erwähnten Informationswissenschaftlerin Batya Friedman und dem Psychologen Peter H. Kahn Jr. gemeint, dass diese Computersysteme systematisch bestimmte Personen oder Personengruppen zugunsten anderer diskriminieren: „Bias refers to systematic unfairness perpetrated on individuals or groups."[768] Eine solche Diskriminierung ist unfair, wenn einer Person oder einer Gruppe von Personen aus unvernünftigen oder unangemessenen Gründen eine Chance oder ein Gut vorenthalten oder ihnen ein unerwünschtes Ergebnis zugewiesen wird. Eine unfaire Diskriminierung führt aber erst dann zu einem „Bias", wenn sie systematisch erfolgt und mit einem ungerechten Ergebnis einhergeht.[769]

Friedman und Nissenbaum unterscheiden drei verschiedene Formen von Verzerrungen: Bestehende, technische und emergente Verzerrungen. Bestehende Verzerrungen (preexisting bias) existieren unabhängig und in der Regel schon vor der Entwicklung des Systems. Bereits bestehende Vorurteile können in ein System einfließen, entweder durch ausdrückliche und bewusste Bemühungen von Einzelpersonen oder Institutionen oder implizit und unbewusst, sogar

766 Stahl et al. 2023:10f.
767 Goodman, Flaxman 2017:53
768 Friedman, Kahn 2008:1253
769 Friedman, Nissenbaum 1996:332

trotz der besten Absichten.[770] Es können sich sowohl gesellschaftliche Strukturen als auch individuelle Voreingenommenheit von Personen in Computersystemen widerspiegeln. Auf der gesellschaftlichen Ebene beeinflussen Organisationen, Institutionen oder die Kultur die Entstehung von Verzerrungen, während es auf der individuellen Ebene auf die Einstellungen von Personen ankommt, die einen wesentlichen Einfluss auf die Gestaltung des Systems haben.

Technische Verzerrungen (technical bias) ergeben sich aus technischen Zwängen oder Überlegungen. So kann eine technische Verzerrung zum Beispiel durch die Beschränkungen von Computerhardware und -software entstehen. Eine solche technische Beschränkung ist zum Beispiel die begrenzte Größe des Bildschirms, wodurch auf einer Website mit Flugangeboten nur einige Angebote direkt sichtbar sein können – die weiteren Angebote befinden sich auf der nächsten Seite. Wenn der Ranking-Algorithmus systematisch Flüge ausgesuchter Fluggesellschaften zuoberst und Flüge anderer Fluggesellschaften später platziert, weist das System eine technische Verzerrung auf. Technische Verzerrungen können ebenfalls entstehen, weil ein Algorithmus angewendet wird, der nicht alle Gruppen unter allen wichtigen Bedingungen gleichbehandelt oder weil Pseudozufallszahlen unzulänglich generiert werden, sodass eine bestimmte Gruppe von Menschen systematisch bevorzugt wird, oder weil der Versuch unternommen wird, menschliche Konstrukte wie Diskurse, Urteile oder Intuitionen für Computer zugänglich zu machen.[771] Darunter fällt zum Beispiel auch die von Goodman identifizierte Unsicherheitsverzerrung (uncertainty bias). Diese entsteht, wenn eine gesellschaftliche Gruppe im Datensatz unterrepräsentiert ist und der Algorithmus gleichzeitig risikoavers entscheidet. Denn dann bevorzugt der Algorithmus zum Beispiel bei einer Kreditvergabe die Gruppen, die in den Trainingsdaten besser repräsentiert sind, da mit diesen Vorhersagen weniger Unsicherheit verbunden ist.[772]

Emergente Verzerrungen (emergent bias) entstehen erst bei der Benutzung des Computersystems. Wenn zum Beispiel neue wissenschaftliche Erkenntnisse nicht in das Computersystem aufgenommen werden, entsteht eine emergente Verzerrung. Auch wenn sich

770 Friedman, Kahn 2008:1253; Friedman, Nissenbaum 1996:332–335
771 Friedman, Nissenbaum 1996:332, 334–336; Friedman, Kahn 2008:1253
772 Goodman 2016:5; Goodman, Flaxman 2017:54

die Bevölkerungsgruppe, die das System nutzt, von der Gruppe unterscheidet, die bei der Systemgestaltung als Nutzende angenommen wurde, entsteht eine emergente Verzerrung. So kann ein System zum Beispiel eine Lese- und Schreibfähigkeit voraussetzen, aber von Menschen angewandt werden, welche diese Fähigkeit nicht besitzen.[773]

Alle Verzerrungen können durch besondere Aufmerksamkeit minimiert oder sogar vermieden werden. Bestehende Verzerrungen (preexisting bias) können laut Friedman und Nissenbaum durch ein gutes Verständnis der herrschenden Vorurteile minimiert werden. Das heißt, bei der Gestaltung eines Computersystems müsste darauf geachtet werden, dass keine Voreingenommenheit in Bezug auf kulturelle Identität, Klasse, Geschlecht, Lese- und Schreibfähigkeit und Behinderung reproduziert werden. Hierfür könne es nützlich sein, eine diverse Gruppe von potenziell Nutzenden einzubeziehen, um das System zu testen und bewerten. So können auch unbeabsichtigte Verzerrungen aufgedeckt werden.[774]

Um technische Verzerrungen (technical bias) zu vermeiden, muss laut Friedman und Nissenbaum bei der Gestaltung schon die Nutzung des Systems antizipiert werden, sodass technische Entscheidungen nicht im Widerspruch zu moralischen Werten stehen. Bei der technischen Umsetzung eines Systems sollte es also nicht darum gehen, immer die einfachste Lösung zu realisieren, sondern auch darum, gerechte Ergebnisse zu produzieren.

Emergente Verzerrungen (emergent bias) können laut Friedman und Nissenbaum minimiert werden, indem antizipiert wird, in welchen Kontexten das System genutzt werden könnte. Wenn es für bestimmte Nutzungskontexte nicht möglich ist, das System passend zu gestalten, sollte das System in diesen Kontexten nicht angewendet werden.[775]

Laut Friedman und Nissenbaum liegt also die Verantwortung für die Vermeidung oder Minimierung von Verzerrungen bei denjenigen, die das System gestalten. Aber von den Verzerrungen betroffen sind vorrangig diejenigen, die das System nutzen und sich auf die Ergebnisse verlassen. Diese Gruppe der Datennutzenden ist nur

773 Friedman, Nissenbaum 1996:332, 335–336; Friedman, Kahn 2008:1253
774 Friedman, Kahn 2008:1253; Friedman, Nissenbaum 1996:343f.
775 Friedman, Nissenbaum 1996:344; Friedman, Kahn 2008:1253

dann in der Lage zu überprüfen, ob eine Verzerrung durch das System reproduziert oder ausgelöst wird, wenn die Vorgänge im System transparent sind. Transparenz setzt sich laut Mittelstadt et al. aus zwei Komponenten zusammen: Zugänglichkeit und Verständlichkeit.[776] Ohne Transparenz, also ohne zugängliche und verständliche Informationen über das System und sein Vorgehen, können Datennutzende nicht überprüfen, ob die Entscheidung des Systems auf Verzerrungen beruht und dadurch zusätzlich ein normatives Risiko für sie darstellt. Denn auf Grund ihrer Komplexität ist es nahezu unmöglich, im Voraus zu wissen, wann und wie Algorithmen versagen werden.[777]

Ob Computersysteme systematisch und unfair bestimmte Personen oder Personengruppen zugunsten anderer diskriminieren, lässt sich manchmal schon bei der Anwendung erkennen, wenn zum Beispiel die gute Sehfähigkeit der Nutzenden vorausgesetzt wird durch kleine Schrift oder geringe Kontraste. Andere Verzerrungen ließen sich auch beim Vergleich von Ergebnissen erkennen, wenn zum Beispiel Frauen systematisch und aus unvernünftigen oder unangemessenen Gründen andere Ergebnisse zur Verfügung gestellt bekommen als Männer und dies zu einem ungerechten Resultat führt. Allerdings ist es aufwendig, eine große Menge an Ergebnissen zu vergleichen, um eine Systematik feststellen zu können. Einfacher wäre es, wenn transparent wäre, mit Hilfe welcher Entscheidungskriterien das Computersystem zu den Ergebnissen kommt.

7.1 Umsetzung von Transparenz im Datenökosystem

In Bezug auf jegliche Algorithmen bedeutet Transparenz, dass zugänglich und verständlich sein muss, wie das Ergebnis zustande gekommen ist. Dies scheint jedoch schwer zu realisieren zu sein, da das Vorgehen von komplexen Algorithmen undurchsichtig ist.[778] Der Gedanke findet sich auch bei der KI-Ethikerin Kate Vredenburgh: „Some types of algorithms, such as decision trees, are usually judged by experts to be functionally transparent. Other algorithms, such

776 Mittelstadt et al. 2016:6
777 Tutt 2017:86
778 Burrell 2016:1; Humphreys 2009:619

as neural nets, are so complex that it is difficult to understand the rules by which inputs are related to outputs."[779] Durch die Komplexität der Algorithmen ist der Entstehungsprozess der Ergebnisse nicht transparent – er ist sogar undurchsichtig. Wie bereits unter 3.1 erwähnt, ist dem Epistemologen Paul Humphreys zufolge ein Prozess für eine Person epistemisch undurchsichtig (epistemically opaque), wenn die Person nicht alle epistemisch relevanten Elemente des Prozesses kennt.[780] Für Datennutzenden wäre ein epistemisch relevantes Element des Prozesses im Datenökosystem zu wissen, wie die Ergebnisse, die ihnen zur Verfügung gestellt werden, zustande gekommen sind.

Bei der Realisierung dieses Anspruchs gibt es zwei Probleme. Das eine ist das einseitige Interesse an Transparenz. Während Datennutzende ein Interesse daran haben, haben Datenverarbeitende eher ein Interesse an wirtschaftlicher Rentabilität. Dazu schreiben Mittelstadt et al.: „Information about the functionality of algorithms is often intentionally poorly accessible."[781] Der Zugang wird oft deshalb erschwert, da die Algorithmen den Unternehmen einen Wettbewerbsvorteil verschaffen. Eine vollständige Offenlegung der Algorithmen würde den Wettbewerb zwischen verschiedenen Anbietern beenden und wäre dementsprechend auch zum Nachteil der Nutzenden, weil ein Wettbewerb die Qualität verbessern kann. Außerdem würde die vollständige Transparenz des Algorithmus diesen womöglich mehr böswilligen Angriffen und Manipulationsversuchen aussetzen. Denn wer das System versteht, kann es zu seinem eigenen Vorteil ausnutzen.[782] Transparenz kann also auch Nachteile haben, aber trotzdem sind Datennutzende daran interessiert zu verstehen, wie die ihnen zur Verfügung gestellten Ergebnisse erstellt werden und wie sie Entscheidungen beeinflussen, die in datengesteuerten Verfahren getroffen werden.[783]

Außerdem ermöglicht Transparenz den Datennutzenden auch, mögliche wissentliche Täuschung, wie zum Beispiel die Umgehung von Vorschriften oder die Manipulation der Verbraucher durch Da-

779 Vredenburgh 2022:221
780 Humphreys 2009:618
781 Mittelstadt et al. 2016:6
782 Granka 2010:366; Kitchin 2016:7; Burrell 2016:3
783 Mittelstadt et al. 2016:6

tenverarbeitende, aufzudecken.[784] Bei der Abwägung, welche Interessen überwiegen – der Wettbewerbsvorteil durch Intransparenz oder mehr Kontrolle über datengesteuerte Ergebnisse durch Transparenz – haben Datenverarbeitende Vorteile, weil sie über die gefragten Informationen verfügen und sie dementsprechend vorenthalten können. Laut der Informatikerin und Soziologin Jenna Burrell bräuchte man unabhängige, vertrauenswürdige Prüfende, welche die Geheimhaltung für die Unternehmen wahren können und gleichzeitig dem öffentlichen Interesse dienen, indem sie Verbrauchermanipulationen oder Verstöße gegen Rechtsvorschriften durch Lesen des Codes erkennen.[785]

Das andere Problem ist, dass die Entscheidungsfindung von Algorithmen oft nicht eindeutig nachvollziehbar ist. Hier ergibt sich die Undurchsichtigkeit von Algorithmen für Datennutzende schon dadurch, dass nicht jede Person die Fähigkeit hat, Code zu schreiben und zu lesen.[786] Zudem werden die Methoden zur Programmierung von KI-Systemen immer komplexer, bis ihr Vorgehen selbst den Entwickler:innen nicht mehr verständlich ist. Der Rechtswissenschaftler Frank Pasquale formuliert das so: „Deconstructing the black boxes of Big Data isn't easy."[787] So ist auch das Identifizieren von Fehlern und ihren Ursachen erschwert.[788] Und es passieren laut den Technologiephilosoph:innen Thomas Grote, Konstantin Genin und Emily Sullivan tatsächlich Fehler beim Einsatz von Algorithmen: „Machine learning models often achieve impressive accuracy under training conditions, but fail in spectacular or unexpected ways when they are deployed in real-world settings."[789] Auch wenn Algorithmen im Test exakt arbeiten, können sie beim Einsatz in der Praxis versagen.

Das Vorgehen der Algorithmen ist so schwer zu verstehen auf Grund ihrer Undurchsichtigkeit (opacity). Algorithmen sind insofern undurchsichtig, als man als Empfänger der Ausgabe des Algorithmus keine konkrete Vorstellung davon hat, wie oder warum

784 Pasquale 2015:2; Burrell 2016:1f.
785 Burrell 2016:1f.
786 Ebd.
787 Pasquale 2015:6
788 Ertel 2021:343
789 Grote et al. 2024:1

die Ausgabe auf der Grundlage der Eingaben zustande gekommen ist.[790] Um dem entgegen zu wirken, könnte man fordern, dass die Öffentlichkeit im Verstehen von Code geschult wird, damit sie Algorithmen bewerten und kritisieren kann. Allerdings ist Burrell der Meinung, dass dabei die Komplexität der Algorithmen unterschätzt wird. Die Datensätze können überschaubar und der Code klar geschrieben sein, aber ihr Zusammenspiel im Algorithmus macht die Komplexität aus und sorgt dafür, dass es zeitaufwändig ist, ihn zu verstehen.[791] Die Berechnungen der Algorithmen sind so schnell und komplex, dass Menschen die Prozesse nicht reproduzieren oder verstehen können.[792] Selbst, wenn die Methoden öffentlich zugänglich wären, wäre es trotzdem noch schwer, sie zu verstehen.[793] Denn Algorithmen können für Personen undurchsichtig sein, weil eine Diskrepanz besteht zwischen der komplexen mathematischen Optimierung von Algorithmen und menschlichem Denken.[794]

Datennutzende sind typischerweise epistemisch von Datenverarbeitenden und den von ihnen eingesetzten Algorithmen abhängig. Die epistemische Abhängigkeit von algorithmischen Ergebnissen ist dem Technologieethiker Jeroen van den Hoven zufolge vergleichbar mit der epistemischen Abhängigkeit von Expert:innen. Denn es ist nicht möglich, alles aus eigener Erfahrung selbst zu wissen und jede Behauptung auf ihren Wahrheitsgehalt hin zu überprüfen – deshalb verlassen wir uns auf Expert:innen: „Many things we ordinarily claim to know, we claim to know because we believe that others possess direct and indefeasible (empirical) evidence for them."[795] Zum Beispiel behaupten wir zu wissen, dass Rauchen Lungenkrebs verursacht, obwohl wir das selbst nicht beweisen können. Um aber trotzdem Stellung nehmen zu können, verlassen wir uns auf Expert:innen. Die Rolle der Expert:innen übernehmen nun immer öfter Algorithmen, wenn sie zum Beispiel im medizinischen Kontext eingesetzt werden oder auch wenn sie entscheiden, ob eine Bewerberin abgelehnt wird. Datennutzende sind dann epistemisch von

790 Burrell 2016:1
791 Ebd.:4f.
792 Humphreys 2009:619
793 Pasquale 2015:6
794 Burrell 2016:1f.
795 Van den Hoven 1998:105

algorithmischen Ergebnissen abhängig, wenn es für sie schwierig ist, die Ergebnisse unabhängig zu verifizieren. Da die Schwierigkeit einer unabhängigen Verifizierung variieren kann, gibt es verschiedene Grade der epistemischen Abhängigkeit.[796]

Wenn Datennutzende sowohl eng eingebettet sind in so ein Datenökosystem als auch epistemisch abhängig von den algorithmischen Ergebnissen sind, bezeichnet van den Hoven sie als „epistemische Sklaven" des Systems, da es schwierig ist, dem System Output ohne zusätzliche Informationen zu widersprechen: „(…) if the system tells you that a candidate is unsuitable and has a record of high accuracy, it will be difficult to disagree (and defend the disagreement later on) if little other information is available."[797] Hier wenden Grote, et al. ein, dass die Vorhersagen des Modells mit großer Wahrscheinlichkeit nicht die einzigen Beweise sind, die den menschlichen Entscheidungstragenden zur Verfügung stehen, und sie sich unabhängig vom Algorithmus ein Urteil bilden können.[798] Van den Hoven würde jedoch widersprechen, denn seiner Meinung nach verwenden Datennutzende den Output der Algorithmen im Datenökosystem als direkten Input für ihre praktischen Überlegungen und stützen ihre Entscheidungen hauptsächlich darauf. Dadurch können Datennutzende für ihr eigenes Handeln womöglich keine vom Algorithmus unabhängigen Gründe mehr vorbringen, wenn zum Beispiel Bewerberinnen abgelehnt werden, weil der Algorithmus das empfiehlt.[799]

Auch bei menschlichen Empfehlungen kann man nicht davon ausgehen, dass sie frei von Verzerrungen sind. Es wäre also ein sehr hoher und womöglich unerfüllbarer Anspruch zu verlangen, dass man als Empfänger:in einer Empfehlung ausschließlich mit verzerrungsfreien und wahren Ergebnissen konfrontiert wird. In der Interaktion mit Menschen können wir nicht davon ausgehen, dass ihre Empfehlungen unbedingt wahr sind. Deshalb stellen wir Rückfragen oder prüfen die Empfehlungen mit Hilfe von anderen Quellen. Der Unterschied ist jedoch, dass Datennutzende bezüglich

796 Rooksby 2009:82; van den Hoven 1998:104f.; Buijsman, Veluwenkamp 2023:546
797 Buijsman, Veluwenkamp 2023:545
798 Grote et al. 2024:5
799 Van den Hoven 1998:104f.; Buijsman, Veluwenkamp 2023:543

der algorithmischen Ergebnisse oft keine Rückfragen stellen können (wenn es sich nicht um Chatbots handelt), und es zudem schwierig sein kann, alternative Quellen zu finden.[800] Wenn zum Beispiel ein Wohngeldantrag bearbeitet wird und ein Algorithmus darüber entscheidet, ob der Antrag akzeptiert oder abgelehnt wird, gibt es für die Antragstellenden keine Möglichkeit, das Ergebnis zu überprüfen oder Rückfragen zu stellen. Um eine Überprüfung oder Rückfragen zu ermöglichen, müsste entweder einer der Verwaltungsmitarbeitenden den Antrag händisch bearbeiten (dann könnte man sich allerdings den Einsatz des Algorithmus sparen), oder das algorithmische Ergebnis müsste einen Hinweis darauf enthalten, warum ein Wohngeldantrag abgelehnt wird.

Ein weiterer Unterschied zwischen menschlichen und algorithmischen Empfehlungen ist, dass das algorithmische Ergebnis objektiver sein kann als eine menschliche Empfehlung, wenn es gelingt, den Algorithmus mit verzerrungsfreien Daten zu trainieren. Zum Beispiel für eine Berufsempfehlung oder die Auswahl von neuem Personal sollten Identifikatoren wie Geschlecht und Herkunft irrelevant sein. Ein Mensch kann solche persönlichen Indikatoren seines Gegenübers nicht ignorieren und kann auch seine eigenen unbewussten Vorurteile nicht leicht ignorieren; ein Algorithmus hingegen kann so trainiert werden, dass Vorurteile gar keine Rolle spielen. Das würde jedoch eine sorgfältige Selektion der zum Training verwendeten Datensätzen erfordern und ist in manchen Anwendungsfällen gar nicht möglich.[801] Im Falle einer Berufsempfehlung wäre es möglich Identifikatoren wie Geschlecht oder Herkunft einer Person aus den Datensätzen zu löschen und somit bei dem Erstellen der Empfehlung zu ignorieren. Entscheidende Faktoren könnten dann unter anderem das persönliche Interesse und Leistungsdaten wie z.B. Schulnoten sein – allerdings sind Schulnoten in den meisten Fällen auch schon nicht frei von Verzerrungen.

Wenn also Datennutzende mit algorithmischen Ergebnissen konfrontiert werden, besteht die Gefahr, dass diese verzerrt sind und gleichzeitig keine Rückfragen gestellt werden können. Die algorithmischen Ergebnisse haben aber in den meisten Anwendungsfällen einen großen Einfluss auf die Datennutzenden, da diese womöglich

800 Buijsman, Veluwenkamp 2023:545; van den Hoven 1998:104f.
801 Mehr dazu unter 7.2.3

sogar epistemisch abhängig sind von den Ergebnissen. Deshalb stellt sich an dieser Stelle die Frage, wie es vermeidbar ist, dass Datennutzende möglicherweise verzerrten Ergebnissen ausgesetzt werden, denen sie Glauben schenken müssen, weil die Ergebnisse nicht überprüfbar sind. Denn so können womöglich falsche Überzeugungen entstehen. Es gibt in der Epistemologie verschiedene Ansprüche an Überzeugungen: Sie sollen wahr sein[802], oder sie sollen gerechtfertigt sein[803], oder sie sollen sowohl wahr als auch gerechtfertigt sein und somit als Wissen gelten können.[804] Doch an dieser Stelle ist es gar nicht notwendig, sich in diese sehr umfangreiche Debatte zu vertiefen. Denn für diese Arbeit ist lediglich relevant, dass Datennutzende nicht mit algorithmischen Ergebnissen konfrontiert werden sollen, die verzerrt (biased) sind. Dafür gibt es zwei Optionen: Entweder die Algorithmen sind verlässlich, oder die Ergebnisse können auf mögliche Verzerrungen überprüft werden.

7.1.1 Verlässlichkeit

Es gibt verschiedene Ansätze, um zu entscheiden, wann eine Überzeugung gerechtfertigt ist. Einer dieser Ansätze ist der Prozess-Reliabilismus. Einer der Hauptvertreter des Prozess-Reliabilismus, der Epistemologe Alvin Goldman, formuliert den Ansatz folgendermaßen: „The justificational status of a belief is a function of the reliability of the process or processes that cause it, where (…) reliability consists in the tendency of a process to produce beliefs that are true rather than false."[805] Das heißt, eine Überzeugung ist dann gerechtfertigt, wenn sie durch einen zuverlässigen Prozess erzeugt wird. Ein solcher Prozess ist zuverlässig, wenn er dazu neigt, wahre Überzeugungen zu produzieren.[806]

Das Konzept des Prozess-Reliabilismus bleibt vage in Bezug darauf, wie verlässlich ein Prozess der Überzeugungsbildung sein

802 Weiner 2005
803 Littlejohn 2009
804 Hawthorne, Stanley 2008:578; Ichikawa, Steup 2018; Buijsman, Veluwenkamp 2023:547
805 Goldman 1979:10
806 Ebd.:10, 14, 18

muss, um gerechtfertigte Überzeugungen hervorzubringen: „It does seem clear, however, that *perfect* reliability isn't required. Belief-forming processes that *sometimes* produce error still confer justification. It follows that there can be justified beliefs that are false."[807] Nach diesem Ansatz ist es also möglich, falsche Überzeugungen zu haben und trotzdem ist es gerechtfertigt, dass man daran glaubt. Goldman bleibt auch vage in Bezug auf die Frage, ob sich die Zuverlässigkeit eines Prozesses nur auf „(...) das Verhältnis von wahren zu falschen Überzeugungen in allen bisherigen und zukünftigen Anwendungen dieses Prozesses (...)"[808] bezieht, oder ob auch die Anwendungen in Situationen relevant sind, die hätten eintreten können.[809] Der Metaphilosoph Steffen Koch plädiert für einen Kompromiss, bei dem „(...) aktuale Anwendungen sowie Anwendungen in bestimmten nahen möglichen Welten Beachtung finden."[810] So stützt sich die Einschätzung der Verlässlichkeit eines Prozesses der Überzeugungsbildung nicht nur auf zu beobachtende Ergebnisse, sondern auch auf mögliche Ergebnisse beim Einsatz des Prozesses zu anderen Zwecken.

Der Prozess-Reliabilismus wird oft von einem Externalismus begleitet. Externalist:innen sind sich darüber im Klaren, dass die Erfahrungen eines Individuums teilweise durch seine Umwelt bestimmt werden. Diese Erfahrungen wiederum sind Glaubensquellen und beeinflussen daher den Glaubensbildungsprozess. Die Rechtfertigung einer Überzeugung hängt hier von der Zuverlässigkeit der Quellen der Überzeugung ab. Deshalb ist es wichtig zu fragen, ob Quellen wie Erinnerungen, Wahrnehmungen und introspektive Zustände zuverlässig sind.[811]

Laut Goldman sehen wir Überzeugungen als gerechtfertigt an, wenn sie durch Prozesse gebildet werden, die wir für zuverlässig halten. Unsere Annahmen darüber, welche glaubensbildenden Prozesse zuverlässig sind, können fehlerhaft sein, aber das habe keinen Einfluss auf deren Angemessenheit.[812] Um eine Überzeugung als

807 Goldman 1979:11
808 Koch 2019:170
809 Goldman 1979:11
810 Koch 2019:171
811 Comesaña 2010:571; Koch 2019:170; Steup, Neta 2020
812 Goldman 1979:18

Wissen bezeichnen zu können, reiche es jedoch nicht, von falschen Überzeugungen mit Hilfe verlässlicher Begründungsverfahren zu einer Schlussfolgerung zu kommen. Damit von Wissen gesprochen werden kann, müsse auch die gesamte Geschichte des Prozesses verlässlich sein und damit also auch die Überzeugungen, von denen man ausgeht.[813] Laut Goldman ist aber eine Überzeugung auch dann noch gerechtfertigt, wenn man sich an die ursprüngliche Begründung, die einen auf verlässliche Weise zu dieser Überzeugung gebracht hat, nicht mehr erinnern kann.[814]

Interessanterweise beschreibt Goldman einen solchen Prozess der Überzeugungsbildung (belief-forming process) so, wie wir heute den Prozess eines Algorithmus beschreiben würden: „Let us mean by a ‚process' a *functional operation* or procedure, i.e., something that generates *mapping* from certain states – 'inputs' – into other states – 'outputs'."[815] Mit „outputs" meint Goldman nicht etwa algorithmische Ergebnisse, sondern Zustände, in denen man zu einem bestimmten Zeitpunkt an eine Aussage glaubt. Eine solche gerechtfertigte Überzeugung resultiert aus Operationen der kognitiven Fähigkeiten, das heißt: „(...) 'information-processing' equipment *internal* to the organism."[816] Gemeint ist hier der Mensch als Organismus, der im Prozess der Überzeugungsbildung Informationen kognitiv verarbeitet. Aber genauso gut könnte sich dieser Begriff der Informationsverarbeitung heute auf den Vorgang eines Algorithmus beziehen.

Ebenso lässt sich der folgende Satz vom Menschen auf Maschinen übertragen: „A reasoning procedure cannot be expected to produce true belief if it is applied to false premises."[817] Wenn Menschen von falschen Annahmen ausgehen, können sie nicht zu wahren Überzeugungen gelangen – auch nicht mit Hilfe eines verlässlichen Prozesses. Gleichermaßen kann auch ein Algorithmus nicht auf Basis von verzerrtem Input einen nicht-verzerrten Output erzeugen – selbst wenn der Prozess der Informationsverarbeitung verlässlich ist. Goldman nennt einen Prozess bedingt verlässlich, wenn ein ausreichen-

813 Goldman 1979:15f.
814 Ebd.:15
815 Ebd.:11
816 Ebd.:13
817 Ebd.

der Anteil seiner Output-Überzeugungen wahr ist, vorausgesetzt, seine Input-Überzeugungen sind wahr.[818]

Die Technologiephilosophen Juan Manuel Durán und Nico Formanek haben, angelehnt an den Prozess-Reliabilismus, einen computergestützten Reliabilismus (computational reliabilism) entwickelt. Dieser besagt, es sei gerechtfertigt, dass Forscher von den Ergebnissen ihrer Simulationen überzeugt sind, wenn es einen zuverlässigen Prozess (d.h. die Computersimulation) gibt, der in den meisten Fällen wahre Ergebnisse hervorbringt.[819] Die Autoren machen die Zuverlässigkeit von Computersimulationen an Verifikations- und Validierungsmethoden, Robustheitsanalysen für Computersimulationen, Geschichten (un)erfolgreicher Implementierungen sowie Expert:innenwissen fest.[820] Die Zuverlässigkeit wird also durch formale Methoden und Leistungsvergleiche sichergestellt. Für erfolgreiche Implementierungen ist es außerdem wichtig, dass der Algorithmus in Übereinstimmung mit den epistemischen Standards eines Bereichs verwendet wird.[821] Während es Durán und Formanek um die Verlässlichkeit von Computersimulationen geht, übertragen Durán und Jongsma die Idee der Verlässlichkeit auf den Einsatz von KI in der Medizin. Von den formalen Methoden für die Feststellung der Zuverlässigkeit übernehmen Letztere nur das Expert:innenwissen und fügen die Transparenz als Methode hinzu, verstanden als ein Prozess, der das Innenleben von Black-Box-Algorithmen offenlegt.[822]

Gegen die neue Methode von Durán und Jongsma würde die Kantianerin Maya Krishnan vermutlich einwenden, dass durch die Transparenz der Vorteil des Reliabilismus verloren geht. Denn der Reliabilismus liefert eine plausible Erklärung dafür, warum man sich auf die Ergebnisse der (kognitiven) Informationsverarbeitung verlassen kann, auch wenn es keine detaillierte wissenschaftliche Theorie gibt: weil das System zuverlässig das richtige Ergebnis produziert: „In some cases, a record of success can stand in for a precise account

818 Goldman 1979:13
819 Durán, Formanek 2018:654
820 Ebd.:656ff.
821 Grote et al. 2024:6
822 Durán, Jongsma 2021:332

of the inner workings of an instrument."[823] Es kann also gerade dank des Reliabilismus eine genaue Beschreibung der inneren Funktionsweise eines Instruments ersetzt werden durch seine Erfolgsbilanz.

Laut Grote et al. kann die Zuverlässigkeit von Algorithmen bewertet werden, indem ihre Robustheit getestet, ihre Vorhersagegenauigkeit überprüft und ihre internen Entscheidungsprozesse bewertet werden. Ein Algorithmus gilt dabei als robust, wenn seine Leistung relativ stabil ist. Seine Vorhersagegenauigkeit kann anhand seiner Ergebnisse getestet werden. Doch das reicht laut Grote et al. noch nicht aus, da es möglich ist, dass ein Algorithmus eine hohe Vorhersagegenauigkeit erreicht, dies aber auf der Grundlage von Merkmalen, die wir Menschen für falsch halten. So gibt zum Beispiel ein von den bereits erwähnten Informatikern Ribeiro et al. programmierter Algorithmus aus, dass auf einem Bild ein Wolf zu sehen ist, wenn Schnee (oder ein heller Hintergrund im unteren Bildbereich) vorhanden ist, und wenn auf dem Bild kein Schnee zu sehen ist, gibt der Algorithmus an, es sei ein Husky zu sehen, unabhängig von der Farbe, der Position oder der Haltung des Tieres. Dies liegt daran, dass der Algorithmus ausschließlich mit Bildern von Wölfen im Schnee und solchen von Huskys ohne Schnee trainiert wurde.[824] Grote et al. ziehen daraus folgenden Schluss: „In the case of the wolf v. Husky classifier, the model is unreliable, because it relies on the wrong features (…)."[825] Solange auf allen neuen Bildern von Wölfen Schnee im Hintergrund ist und auf allen Bildern von Huskys kein Schnee, wäre die Vorhersagegenauigkeit des Algorithmus sehr hoch – allerdings auf Grund von Merkmalen, die nichts mit den jeweiligen Tieren zu tun haben. Diese unerwünschten Korrelationen, die der Algorithmus während des Trainings erkannt hat, sind sehr schwer zu identifizieren, wenn man sich nur die Rohdaten und Vorhersagen ansieht.[826] Deshalb ist laut Grote et al. die Bewertung des internen Entscheidungsprozesses von Algorithmen ein zentraler Ansatz zur Beurteilung ihrer Zuverlässigkeit.[827]

823 Krishnan 2020:496
824 Ribeiro et al. 2016:1142
825 Grote et al. 2024:6
826 Ribeiro et al. 2016:1142
827 Grote et al. 2024:6

Die Erfolgsbilanz von Algorithmen ist für Datennutzende, die mit einem algorithmischen Ergebnis konfrontiert werden, nicht leicht zu überprüfen, deshalb verlassen sie sich diesbezüglich auf die Entwickler:innen des Systems. Der Reliabilismus geht in Bezug auf algorithmische Ergebnisse davon aus, dass es gerechtfertigt wäre, wenn die Datennutzenden die Ergebnisse des Systems in ihre praktischen Überlegungen einbeziehen – angenommen den Datennutzenden wurde von den Entwickler:innen des Systems korrekterweise mitgeteilt, dass das System insgesamt zuverlässig ist.[828] Die Epistemologin Katherine Dormandy wendet ein: „Allerdings ist anzumerken, dass es ohne Belege alles andere als klar ist, ob eine gewisse Überzeugung überhaupt zuverlässig gebildet wurde."[829] Die Entwickler:innen brauchen mindestens Informationen darüber, ob das System in verschiedenen Anwendungsfällen mehr wahre als falsche Ergebnisse erzeugt hat, um das System als zuverlässig einstufen zu können.

Die Fokussierung auf die Erfolgsbilanz eines Prozesses bedeutet zudem, dass es bei einem Algorithmus, der als zuverlässig gilt, gerechtfertigt wäre, von seinen Ergebnissen überzeugt zu sein, selbst wenn sie falsch sind. Das Gleiche gilt andersherum: Gilt ein Algorithmus als unzuverlässig, wäre es nicht gerechtfertigt, wenn Datennutzende von den algorithmischen Ergebnissen überzeugt wären.[830] Wenn man also herausfinden möchte, ob man algorithmischen Ergebnissen gerechtfertigter Weise glauben kann, ist man angewiesen auf das Urteil der Programmierenden und Systembetreibenden über die Verlässlichkeit des Algorithmus. Die Programmierenden und Systembetreibenden wissen bei hochkomplexen Algorithmen aber auch nicht genau, wie das System funktioniert und machen dementsprechend seine Zuverlässigkeit an seiner Erfolgsbilanz fest.

Es fragt sich nun, ob es ein Problem ist, von algorithmischen Ergebnissen überzeugt zu sein, wenn sie falsch sind. Wenn im Datenökosystem Transparenz fehlt, können Datennutzende durch algorithmische Ergebnisse zu falschen Überzeugungen gelangen. Gelangen Datennutzende auf diesem Wege zu falschen Überzeugungen, können sie Schwierigkeiten haben, epistemisch wertvolle Zustände

828 Buijsman, Veluwenkamp 2023:548
829 Dormandy 2019:184
830 Buijsman, Veluwenkamp 2023:548

zu erreichen.[831] Epistemisch wertvoll ist das, was in einer relevanten Beziehung zum Ziel der Wahrheit steht. Dabei ist es das epistemische Ziel, genaue und umfassende Überzeugungen zu haben.[832] Wahre Überzeugungen sind laut dem Epistemologen William P. Alston notwendig zum Überleben und zum Vermeiden von Chaos: „(…) it is important for human flourishing to be guided by correct rather than incorrect suppositions about how things are, where this is of interest or importance to us (…)"[833] Da es für das menschliche Wohlergehen wichtig ist, sich von richtigen und nicht von falschen Überzeugungen leiten zu lassen, sollte das Erzeugen von falschen Überzeugungen durch algorithmische Ergebnisse möglichst vermieden werden. Deshalb ist es ein Problem des Reliabilismus, dass es gerechtfertigt sein kann, von algorithmischen Ergebnissen überzeugt zu sein, selbst wenn sie falsch sind.

Ein weiteres Problem bei diesem Ansatz ist, dass der Reliabilismus äußeren Faktoren und Gründen gegenüber blind ist. Denn der Prozess der Informationsverarbeitung ist für externe Faktoren empfindlich, aber er bringt Überzeugungen hervor, ohne dass das Subjekt diese Empfindlichkeit bewusst reflektiert.[834] Wenn also Überzeugungsquellen zuverlässig waren und der Prozess der Überzeugungsbildung wahre Überzeugungen hervorgebracht hat, wird erwartet, dass dieser zuverlässig ist und automatisch immer wahre Überzeugungen hervorbringt. Dagegen spricht laut dem Technologiephilosophen Stefan Buijsman vor allem eins: „(…) algorithms will perform worse in outlier cases, even when they are generally accurate."[835] So mag ein Algorithmus als verlässlich gelten, weil er oft genug richtig lag. Aber es bleibt trotzdem unklar, ob der Algorithmus in allen Fällen gute Entscheidungen trifft. Um Einzelfälle zu erkennen, in denen der Algorithmus verzerrte Ergebnisse bereitstellt, müssten sehr viele Tests für alle möglichen Ausnahmen geprüft werden. Der Algorithmus von Amazon galt als verlässlich, bis deutlich wurde, dass Bewerberinnen systematisch aussortiert wurden. Buijsman hat einen Lösungsvorschlag für das Dilemma: „Understanding why the algo-

831 Harris 2022:137
832 Foley 1993:102, 198
833 Alston 2005:30
834 Comesaña 2010:571
835 Buijsman 2023:203

rithm presents a certain output may help to gain a more fine-grained justification, and avoid mistakes that would arise if the algorithm is given blanket trust."[836] Deshalb soll im Folgenden der Ansatz des Evidentialismus betrachtet werden.

7.1.2 Überprüfbarkeit

Ein anderer Ansatz um zu entscheiden, wann eine Überzeugung gerechtfertigt ist, ist der Evidentialismus. Zwei der Hauptvertreter der Position, die Epistemologen Earl Conee und Richard Feldman, formulieren diese so: „What we call evidentialism is the view that the epistemic justification of a belief is determined by the quality of the believer's evidence for the belief."[837] Die epistemische Rechtfertigung einer Überzeugung hängt bei diesem Ansatz nicht von den kognitiven Fähigkeiten der Menschen oder von den kognitiven Prozessen oder Praktiken der Informationsbeschaffung ab, die zu der Überzeugung geführt haben – sie hängt allein von den Beweisen ab, die man für sie hat. Eine Überzeugung ist laut Conee und Feldman dann „gut begründet" (well-founded), wenn sie sowohl epistemisch gut gestützt als auch auf die richtige Art und Weise zustande gekommen ist. Ob eine Überzeugung „gut begründet" ist, hängt ab von der Evidenz, die man hat, und der Evidenz, die man bei der Bildung der Überzeugung verwendet. Man muss also „(...) *aufgrund der entsprechenden Belege* (...)"[838] zu einer Überzeugung kommen, damit die Überzeugung gemäß dem Evidentialismus gerechtfertigt ist.[839]

Der Evidentialismus bietet auch eine Erklärung dafür, woher wir wissen, wann wir zu wahren Überzeugungen gelangt sind. Es reicht nicht, an die Wahrheit der Überzeugung zu glauben, um ihre Wahrheit festzustellen.[840] Um herauszufinden, ob eine Überzeugung wahr ist, braucht man Beweise für diese Überzeugung.[841] Als Beweis reicht es jedoch nicht, eine Rechtfertigung für den Glauben an die

836 Buijsman 2023:203
837 Conee, Feldman 2004:83
838 Dormandy 2019:179
839 Conee, Feldman 2004:93
840 Ebd.:251
841 David 2001:154, 160, 163; Conee, Feldman 2004:249

Wahrheit der Überzeugung vorzubringen.[842] Denn man könnte gerechtfertigter Weise den Aussagen von Ärzt:innen Glauben schenken und dadurch dennoch zu falschen Überzeugungen gelangen – weil sich die Ärzt:innen zum Beispiel irren. Laut Conee und Feldman brauchen wir stattdessen eine Rechtfertigung für die Überzeugung selbst: „A proposition is epistemically justified to someone when it is evident to the person that the proposition is true."[843] Was epistemisch gerechtfertigt werden muss, um zu wissen, ist nicht „die Überzeugung" im Sinne von: der Glaube an die Aussage. Es ist „die Überzeugung" im Sinne von: die geglaubte Aussage.[844] Um herauszufinden, ob eine Überzeugung im Sinne einer geglaubten Aussage wahr ist, brauchen wir also Beweise für die Wahrheit dieser Überzeugung.

Solche Beweise können in Form von Propositionen angeführt werden. Eine Proposition kann mit Hilfe der Vernunft oder aus dem Zeugnis von anderen abgeleitet werden.[845] Da wir nicht alles selbst erfahren können und uns nicht alles nur aus unserer Vernunft ableiten können, müssen wir uns manchmal auf das Zeugnis anderer verlassen. Allerdings reicht das Zeugnis anderer nur dann als Beweis für die Wahrheit einer Überzeugung, wenn man diese nicht selbst vernünftig herleiten kann oder die notwendigen Erfahrungen nicht selbst machen kann.[846] Außerdem ist das Zeugnis anderer nur dann eine Beweisquelle, wenn die anderen eine zuverlässige Quelle darstellen.[847] Wenn man zum Beispiel mit chronischen Bauchschmerzen zu einer Ärztin geht, macht man die Erfahrung selbst, kann sich aber die Ursache als Laie nicht vernünftig herleiten – deshalb muss man sich für eine Erklärung der chronischen Bauchschmerzen auf die Ärztin verlassen. Damit ihr Zeugnis als Beweisquelle für eine wahre Überzeugung dienen kann, muss sie als Quelle zuverlässig sein. In diesem Beispiel bedeutet das, dass die Ärztin durch ein Medizinstudium ausgebildet worden sein muss, ihr die Approbation erteilt wurde und sie sich an den Ärzte-Kodex hält.

842 David 2001:160; Conee, Feldman 2004:251
843 Conee, Feldman 2004:252f.
844 Ebd.:252
845 Ebd.:238; Steup, Neta 2020
846 Steup, Neta 2020
847 Conee, Feldman 2004:225; Steup, Neta 2020

Wenn allerdings Beweise nur solche für wahr oder wahrscheinlich gehaltene Propositionen sein können, entsteht das Problem des infiniten Regresses, weil laut Dormandy „(...) Belege selber durch weitere Belege gestützt werden müssen."[848] Eine Möglichkeit zur Auflösung dieses Problem ist Conee und Feldman zufolge das Akzeptieren von repräsentationalen Erfahrungen, wie zum Beispiel Wahrnehmungen und Erinnerungen, als Beweise: „(...) I do not wish to rule out the possibility that experiences or perceptual states can count as evidence as well."[849] Solche Erfahrungen sind selbst nicht begründbar, aber leisten trotzdem einen wichtigen Beitrag zum Sammeln von Beweisen und zum Bilden einer Überzeugung. Über Wahrnehmungen kann laut Dormandy die Annahme gemacht werden, „(...) dass ihnen relativ viel epistemisches Gewicht zugeschrieben werden kann, sofern es keine Gegenbelege (*defeaters*) gibt."[850] Belege stützen eine Überzeugung dann, wenn „(...) diese Überzeugung die beste verfügbare Erklärung für die Belege bietet."[851] Das heißt, man nimmt zum Beispiel wahr, dass in der Praxis der Ärztin eine Approbationsurkunde mit dem Namen der Ärztin an der Wand hängt. Dieser Beleg stützt die Überzeugung, dass die Ärztin ihr Medizinstudium erfolgreich abgeschlossen hat. Die Approbationsurkunde könnte auch gefälscht sein, aber die plausibelste Erklärung für diese Wahrnehmung (Beleg) ist, dass die Ärztin ihr Medizinstudium abgeschlossen hat und dann einen erfolgreichen Antrag auf Approbation gestellt hat.

Da repräsentationale Erfahrungen auch als Belege gelten, geht der Evidentialismus mit einem Internalismus einher. Es gibt einerseits den mentalistischen Internalismus (mentalism) und andererseits den Zugangsinternalismus (accessibilism). Denn die Dinge, die eine Überzeugung rechtfertigen, sind entweder mentale Zustände (mentalistischer Internalismus) oder sie sind bei Reflexion erkennbar (Zugangsinternalismus). Den Zugangsinternalismus definieren Conee und Feldman folgendermaßen: „What we shall call "accessibilism" holds that the epistemic justification of a person's belief is determined by things to which the person has some special sort

848 Dormandy 2019:180
849 Conee, Feldman 2004:225f.
850 Dormandy 2019:183
851 Ebd.:181

of access."[852] Gegen diese zweite, logisch stärkere Version des Internalismus spricht jedoch, dass er den Menschen zu viel abverlangt. Conee und Feldman plädieren auf Grund der Einfachheit und Klarheit für den mentalistischen Internalismus: „(...) [mentalist] internalism is the view that a person's beliefs are justified only by things that are internal to the person's mental life."[853] Die Überzeugungen einer Person sind demnach durch Beweise gerechtfertigt, welche die Person erlebt oder reflektiert.[854]

Das Problem beim mentalistischen Internalismus ist, dass unklar ist, wann die erlebte oder reflektierte Evidenz geeignet ist, eine Überzeugung zu rechtfertigen. Ohne Regeln für die Qualität der Evidenz könnte jede Erinnerung oder Intuition als Beweis für die Wahrheit eines Glaubens gelten. Dies macht die Beweise und den gesamten Rechtfertigungsprozess sehr subjektiv und schwer überprüfbar.[855] Wenn auf der Grundlage von algorithmischen Ergebnissen Überzeugungen gebildet werden, dann werden diese Ergebnisse dabei als Belege für diese Überzeugung betrachtet. Dieser Umstand verdeutlicht das Problem, dass es so kein Qualitätskriterium für Belege gibt. Denn algorithmische Ergebnisse können verzerrt sein, ohne dass Datennutzende und sogar auch Datenverarbeitende sich dessen bewusst sind. In dem Fall kommen Datennutzende „aufgrund der entsprechenden Belege"[856] zu einer Überzeugung, sodass die Überzeugung gemäß dem Evidentialismus gerechtfertigt ist.[857] Doch auch bei falschen Belegen in Form von verzerrten algorithmischen Ergebnissen wäre die Überzeugung gerechtfertigt, da es scheint, als hätte man Belege für die Wahrheit der Überzeugung, und doch ist sie womöglich nicht wahr.

852 Conee, Feldman 2004:55
853 Ebd.:55
854 Dormandy 2019:179; Steup, Neta 2020
855 Comesaña 2010:571
856 Dormandy 2019:179
857 Conee, Feldman 2004:93

7.1.3 Ein Kompromiss

Der Epistemologe Juan Comesaña fasst zusammen, dass der Reliabilismus nicht ohne Beweise kann und der Evidentialismus nicht ohne Verlässlichkeit: „Reliability without evidence is blind, evidence without reliability is empty."[858] Um eine Lösung für die unterschiedliche Qualität der Evidenz im Evidentialismus und den Mangel an Reflexion im Reliabilismus zu finden, hat Comesaña beide Ansätze miteinander kombiniert. Sein sogenannter evidentialistischer Reliabilismus kann wie folgt definiert werden: „(...) to be justified (…) is to believe in accordance with one's evidence, and one's beliefs accord with one's evidence if and only if that evidence is reliably connected to the truth of those beliefs."[859] Eine Überzeugung ist also gerechtfertigt, wenn sie auf der Grundlage von Beweisen gebildet wird. Die Evidenz wiederum ist geeignet, einen Glauben zu rechtfertigen, wenn sie zuverlässig ist.[860] Auf diese Weise wird das Problem des Evidentialismus gelöst, da es jetzt eine Bedingung für die Qualität der Evidenz gibt. Gleichzeitig wird das Problem des Reliabilismus gelöst, weil wir uns nicht auf einen automatischen Glaubensbildungsprozess verlassen müssen, sondern bewusst über die Evidenz nachdenken, die wir erfahren. Indem wir über die Beweise und ihre Zuverlässigkeit nachdenken, können wir die Überzeugungen, die wir bilden, rechtfertigen.

So muss man sich nicht darauf verlassen, dass der Algorithmus meistens wahre Ergebnisse erzeugt, sondern man sammelt Belege dafür. Gleichzeitig muss man sich nicht fragen, welche der Belege qualitativ hochwertig genug sind, um eine Überzeugung rechtfertigen zu können, da das Kriterium die Verlässlichkeit ist. Der evidentialistische Reliabilismus löst allerdings in der Praxis auch nicht alle Probleme. Denn wenn wir nun also mit Hilfe eines Algorithmus zu einer Überzeugung gelangen, brauchen wir Beweise für die Wahrheit dieser Überzeugung, und diese Beweise müssen verlässlich sein. Es ist aber auf Grund der Komplexität des Algorithmus oft schwierig, Beweise außerhalb der algorithmischen Ergebnisse selbst zu finden oder sie mittels unserer Vernunft abzuleiten. Die Tech-

858 Comesaña 2010:571
859 Ebd.:595
860 Ebd.:584

nologiephilosophen Stefan Buijsman und Juan M. Durán schreiben über Machine Learning Models: „(…) their epistemic opacity makes it difficult to extract causal relations and even to determine the reliability and robustness of their predictions."[861] Der algorithmische Prozess ist somit im Wesentlichen epistemisch undurchsichtig, weil es Menschen auf Grund ihrer Natur unmöglich ist, alle epistemisch relevanten Elemente des Prozesses zu kennen.[862] Deshalb müssen wir uns laut dem Epistemologen und Bioethiker John Hardwig in solchen Fällen auf das Zeugnis anderer verlassen: „(...) one can have good reasons for believing a proposition if one has good reasons to believe that *others* have good reasons to believe it (…)."[863] Laut Grote et al. generieren die Algorithmen selbst dieses Zeugnis in Form von ihren Ergebnissen[864], aber oft ist nicht klar, ob dieses Zeugnis auf „guten Gründen" beruht. Über die Gründe der Algorithmen müsste dementsprechend eine Person Zeugnis ablegen, die deren Vorgehen verstehen und nachvollziehen kann, wie die Ergebnisse zustande gekommen sind. Ein Algorithmus selbst gibt nicht von sich aus an, auf welchem Weg sein Output entstanden ist, und sowohl die Programmierenden als auch die Betreibenden des Algorithmus wissen das bei erhöhter Komplexität auch nicht genau.

Bei einem einfachen Wenn-Dann-Algorithmus wissen die Programmierenden und die Betreibenden, auf Grund welcher Entscheidungskriterien der Algorithmus zu seinem Ergebnis kommt, weil sie die Kriterien vorgegeben haben. Sie können also verlässliches Zeugnis darüber ablegen, dass der Algorithmus den Wohnantrag von Person A abgelehnt hat, weil Person A nicht alle gesetzlich notwendigen Bedingungen erfüllt. Programmierende und Betreibende wissen in dem Fall, dass weder Geschlecht noch Religion von Person A ausschlaggebend für die Entscheidung gewesen sind. Sie haben dem Algorithmus die Kriterien, zum Beispiel Wohnungsgröße und Einkommen, vorgegeben und können demnach Zeugnis darüber ablegen, wie die Ergebnisse des Algorithmus zustande kommen. Mit Hilfe dieses verlässlichen Zeugnisses können Datennutzende Beweise für die Wahrheit ihrer durch den Algorithmus verursachten Überzeugungen sammeln.

861 Buijsman, Durán 2024:466
862 Humphreys 2009:618
863 Hardwig 1985:336
864 Grote et al. 2024:5

Handelt es sich aber um ein komplexes KI-System, bei dem selbst für die Programmierenden nicht mehr nachvollziehbar ist, wie der Algorithmus zu seinen Ergebnissen kommt, gibt es folglich niemanden, der verlässlich Zeugnis darüber ablegen kann. Verlässlich sind die Algorithmen nur, wenn jemand nachvollziehen kann, wie die Ergebnisse zustande kommen – das ist nur bei Wenn-Dann-Algorithmen einfach möglich. Die Überprüfung der Ergebnisse auf mögliche Verzerrungen, als alternative Absicherung der durch sie gebildeten Überzeugungen, ist ebenfalls deutlich leichter, wenn das Vorgehen des Algorithmus bekannt ist. Deshalb ist es an dieser Stelle wichtig, sich mit der Erklärbarkeit und Interpretierbarkeit von Algorithmen auseinanderzusetzen. Denn dem Informationswissenschaftler Michael Winikoff zufolge ist eine der wesentlichen Voraussetzungen für angemessenes Vertrauen in ein autonomes System die Fähigkeit zu erklären, warum das System eine Entscheidung getroffen hat.[865]

7.2 Erklärbarkeit und Interpretierbarkeit

Um Beweise zu sammeln für die Verlässlichkeit von algorithmischen Ergebnissen können diese erklärt oder interpretiert werden. Sowohl Erklärung als auch Interpretierbarkeit sind erkenntnistheoretische Begriffe. Wie unter 2.1 erörtert, sind sie jedoch nicht gleichzusetzen. Ein Algorithmus ist interpretierbar, wenn seine Ergebnisse bei der Inspektion des Algorithmus verstanden werden können. Ein Algorithmus ist erklärbar, wenn die Ergebnisse durch einen separaten Prozess verstanden werden können, indem ein anderer Algorithmus zusammenfasst, wie der zu erklärende Algorithmus vorgeht.[866] Interpretierbarkeit ist vor allem bei Wenn-Dann-Algorithmen möglich, da hier zumindest die Entwickler:innen nachvollziehen können, wie die Ergebnisse zustande gekommen sind. Deshalb empfiehlt die Informatikern Cynthia Rudin hingegen, statt Black-Box Algorithmen interpretierbare Algorithmen zu verwenden.[867] Wenn aber doch komplexere Black-Box Algorithmen verwendet wurden, werden Erklärungen gebraucht.

865 Winikoff 2018:8
866 Russel, Norvig 2021:729
867 Rudin 2019; Buijsman 2023:206

Buijsman zufolge gibt es keine allgemein anerkannte Definition von Erklärung (explanation) und auch keine Möglichkeit, die Qualität verschiedener Erklärungen zu vergleichen.[868] Aber er meint, wenn wir Erklärungen fordern, fragen wir „Warum?" – also zum Beispiel „Warum gibt der Algorithmus diese Ergebnisse aus?". Buijsman wird noch konkreter und behauptet, dass Erklärungen Was-wäre-wenn-es-anders-gewesen-wäre-Fragen sind. Er nennt sie kontrastive Warum-Fragen (contrastive why-questions) und meint damit, dass den Datennutzenden ein alternatives Ergebnis präsentiert werden sollte und eine Eingabe, die zu diesem Ergebnis führen würde – nur so ließe sich die Funktionsweise eines Algorithmus vollständig erklären. Je mehr solche kontrastiven Warum-Fragen beantwortet würden, desto besser sei die Erklärung.[869]

Neben diesem interventionistischen Ansatz, der sich darauf konzentriert zu klären, wie das Ergebnis aussehen würde, wenn der Input ein anderer wäre, gibt es noch zwei weitere Ansätze zur Definition von Erklärung: den vereinheitlichenden und den mechanistischen Ansatz. Beim vereinheitlichenden Ansatz geht es darum, möglichst allgemeine Argumentationsmuster zu identifizieren, die es ermöglichen, eine breite Palette von Ergebnissen abzuleiten. Im Idealfall schafft es eine Theorie, ein Regelwerk aufzustellen, welches ein breites Spektrum von Phänomenen zusammenfasst. Die Newtonschen Gesetze gelten als Ideal der Vereinheitlichung, weil sie eine einheitliche Erklärung von Bewegungen auf der Erde und im Weltall ermöglichen.[870] Mechanistische Ansätze hingegen beschreiben das Innenleben eines Systems mit so vielen kausal relevanten Details wie möglich. Die Umsetzung dieses Ansatzes ist bei komplexen KI-Systemen jedoch schwierig, da hier jeder der zahlreichen Parameter eine gewisse kausale Bedeutung für die Ausgabe haben könnte und somit ein vollständiges Modell der Funktionsweise sehr komplex wäre und deshalb nicht viel Aufschluss geben könnte.[871]

Eine Erklärung von algorithmischen Ergebnissen ist laut Goodman und Flaxman nur möglich, wenn der Algorithmus von einem Menschen in Worte gefasst und verstanden werden kann. Sie neh-

868 Buijsman 2022:564
869 Ebd.:573, 580f.; Buijsman 2023:206
870 Buijsman 2023:209
871 Ebd.:212f.

men an, dass dafür die Beziehungen zwischen den Eingabemerkmalen und den Ergebnissen dargestellt werden müssen, sodass deutlich wird, welche Merkmale die Ergebnisse in welcher Weise beeinflussen.[872] Krishnan fügt hinzu, dass durch eine Erklärung in der Lage sein sollte zu wissen, warum der Algorithmus genau diese Ausgabe erzeugt.[873] Sobald die Verhältnisse zwischen Input und Output auf diese Weise offengelegt werden, wären sowohl der vereinheitlichende Ansatz der Erklärung als auch der interventionistische Ansatz im Kontext von Algorithmen umsetzbar. Denn dann können einerseits möglichst allgemeine Argumentationsmuster identifiziert werden, die es ermöglichen, eine breite Palette von Ergebnissen abzuleiten (vereinheitlichender Ansatz).[874] Wenn zum Beispiel das Alter in Kombination mit dem Einkommen einen Einfluss auf die Ergebnisse hat, können mit dieser Regel sowohl Kreditverweigerungen als auch zugeschriebene Kreditwürdigkeit erklärt werden. Andererseits ist es möglich, mit Hilfe der Erkenntnisse über die Beziehung zwischen Input und Output Was-wäre-wenn-es-anders-gewesen-wäre-Fragen zu beantworten (interventionistischer Ansatz).[875] Denn wenn man weiß, welchen Einfluss welcher Input auf das Ergebnis hat, kann man ableiten, wie das Ergebnis aussehen würde, wenn der Input anders wäre.

Manche Algorithmen sind so komplex, dass es schwierig bis unmöglich ist, die Regeln zu verstehen, nach denen der Input mit dem Output verknüpft ist. Für solche Algorithmen haben Informatiker:innen Erklärungsmethoden entwickelt, um die Transparenz zu erhöhen. So gibt es zum Beispiel sogenannte kontrafaktische Erklär-Techniken (counterfactual XAI techniques), die Mittel zur Identifizierung von unterscheidenden Eigenschaften (difference-making properties) eines Algorithmus liefern. Laut Munch et al. kann das ausreichen, um das Verhalten des Algorithmus zu erklären.[876] Die Informatiker:innen Osbert Bastani, Carolyn Kim und Hamsa Bastani nennen ihre Erklärungsmethode „Model Extraction". Hier

872 Goodman, Flaxman 2017:55
873 Krishnan 2020:493
874 Buijsman 2023:209
875 Buijsman 2022:573, 580f.; Buijsman 2023:206
876 Munch et al. 2024:4; Für Hinweise auf Fallstricke der Modellinterpretation siehe Molnar et al. 2022.

wird eine einfachere Version des zu erklärenden Algorithmus mit expliziten Regeln für die Beziehung zwischen Input und Output erstellt. Daraufhin wird geprüft, ob der Output des einfacheren Algorithmus mit dem Output des undurchsichtigen Algorithmus übereinstimmt, wenn beide neuen Input erhalten. Auf diese Weise ist der komplizierte Algorithmus dann anhand von der einfacheren Version interpretierbar.[877]

Laut Buijsman können zum Beispiel Werkzeuge wie Feature Importance Maps oder Bias Detection aufdecken, wenn die algorithmischen Ergebnisse auf irrelevanten oder ungerechten Merkmalen beruhen: „Feature importance maps and counterfactuals can hint that the model looks at features that we think are irrelevant, even if they generally do not enlighten us as to why the model looks at those features."[878] Solche Verfahren und Methoden, die algorithmische Merkmale nachvollziehbar machen sollen, werden als erklärbare KI (explainable AI; XAI) bezeichnet. Andere Werkzeuge aus diesem Bereich können mit Hilfe von statistischer Analyse das Erkennen von Verzerrungen unterstützen und so auf die Gruppen hinweisen, die auf Grund von bestimmten Merkmalen benachteiligt und somit diskriminiert werden.[879] Daraufhin können solche Ungleichheiten manuell von menschlichen Expert:innen geprüft werden und Fehler können womöglich behoben werden. Deshalb können derartige erklärbare KI-Methoden auch dann nützlich sein, wenn sie keine eigentlichen Erklärungen liefern.[880]

Die Informatiker Anupam Datta, Shayak Sen und Yair Zick sind ebenfalls der Meinung, dass die Transparenz von Algorithmen verbessert werden kann, indem erklärt wird, warum eine bestimmte Empfehlung erzeugt wurde. Deshalb führen sie sogenannte „Quantitative Input Influence (QII) measures" ein, die den Grad des Einflusses von Inputs auf die Outputs von Systemen erfassen. Anhand von verschiedenen Variablen können Entscheidungen über Einzelpersonen und Gruppen erklärt werden, indem sie den gemeinsamen Einfluss einer Gruppe von Inputs (z.B. Alter und Einkommen) auf die

877 Bastani et al. 2017; Vredenburgh 2022:221
878 Buijsman 2023:212
879 Für einen Überblick über den Stand der Technik von XAI Methoden siehe Holzinger et al. 2022.
880 Buijsman 2023:212

Ergebnisse sowie den marginalen Einfluss einzelner Inputs innerhalb einer solchen Gruppe (z.B. Einkommen) quantifizieren.[881] Auch diese Methode von Datta et al. gilt als Werkzeug der erklärbaren KI (XAI-Tool). Hierdurch ist es auch bei undurchsichtigen Algorithmen möglich zu beobachten, wie sie bestimmte Inputs zuverlässig mit bestimmten Outputs in Verbindung bringen. Dadurch wissen wir zwar immer noch nicht, warum diese Korrelationen hergestellt werden. Trotzdem ist die Erklärung schon mal spezifischer, wenn nicht nur angegeben wird, welche Merkmale für ein algorithmisches Ergebnis wichtig sind, sondern es genaue Schätzungen darüber gibt, auf welche Weise diese Merkmale wichtig sind.

Der Digitalethiker Lauritz Aastrup Munch, der Technologiephilosoph Jens Christian Bjerring und der Digitalethiker Jakob Thrane Mainz veranschaulichen diesen Erklärungsprozess anhand des Beispiels eines durch ein algorithmisches System verweigerten Kredits. Wenn die Erklärungsmethoden (XAI-Tools) ergeben, dass das niedrige Einkommen einer Person als Input Einfluss hatte auf den Output des Algorithmus, dann wird impliziert, dass das Einkommen als Input zumindest eine Teilursache für die Entscheidung der Kreditverweigerung ist. Die Nennung des Einkommens als zentraler Faktor zur Erklärung einer algorithmischen Schätzung der Kreditwürdigkeit sagt etwas darüber aus, wie das Modell funktioniert. Da das Modell Daten aus der echten Welt reproduziert, spiegeln solche kausalen Beziehungen oft den Einfluss des Einkommens auf die Kreditwürdigkeit in der Vergangenheit wider. Laut Munch et al. besteht die Hoffnung, dass diese XAI-Tools zumindest teilweise ein kausales Verständnis von KI-Systemen ermöglichen und uns somit zumindest einige der Dinge liefern, die uns kausale Erklärungen normalerweise liefern.[882]

Die Technologiephilosophen Timo Freiesleben und Gunnar König, die sich auf Machine Learning spezialisiert haben, machen darauf aufmerksam, dass solche Erklär-Methoden jedoch nicht ohne Kontext hilfreich sind. Sie meinen, oft bleibe unklar, welchem Zweck diese Erklärungen dienen könnten, und unter welchen Bedingungen sie hilfreich sind. Ihrer Meinung nach sollte jede Erklärung mindestens einem präzisierten Zweck dienen: „A purpose is a goal humans

881 Datta et al. 2016:1f.
882 Munch et al. 2024:4f.

have in mind when they ask for explanations."[883] Je nach Ziel der Erklärung können unterschiedliche Methoden angemessen sein. Dabei sollen die Methoden nicht dazu genutzt werden, Datennutzende dazu zu bringen, den Algorithmen zu trauen, sondern sie sollen dazu beitragen, den Mechanismus der Algorithmen transparenter zu machen. Dabei stellen die Autoren in Frage, ob die Datennutzenden die Wirkweise von Algorithmen verstehen werden. Algorithmen sind komplex, Erklärungen hingegen sollen einfach sein und nicht jeden Aspekt eines Algorithmus wiedergeben. Eine mögliche Lösung dafür wäre es laut Freiesleben und König, verschiedene Arten von Erklärmethoden gleichzeitig einzusetzen, damit mehrere Aspekte des Algorithmus beleuchtet werden können.[884]

7.2.1 Erklärungen für algorithmische Ergebnisse

Es gibt jedoch einige Autor:innen, die in Frage stellen, ob wir überhaupt Erklärungen für algorithmische Ergebnisse benötigen. Buijsman hat vier dieser Argumentationen zusammengestellt und widerlegt. Als erstes widmet er sich dem Argument des KI-Ethikers Scott Robbins, dass nicht der Algorithmus erklärt werden müsse, sondern seine Entscheidungen. Wir bräuchten nur dann Erklärungen, wenn die Auswirkungen der Entscheidungen auf andere Menschen hoch sind oder sein können. So behauptet Robbins zum Beispiel, dass die Entscheidungen von Algorithmen, die Muttermale als gutartig oder bösartig einstufen, nicht erklärt werden müssen. Denn wenn ein Muttermal als bösartig eingestuft wird, sei die Folge dieser Entscheidung lediglich eine Biopsie. Ärzt:innen seien auch nicht verpflichtet, ihre Entscheidungen zu erklären, deshalb solle das auch beim Einsatz von Algorithmen nicht Pflicht sein. Robbins ist zudem der Meinung, dass Algorithmen überflüssig wären, wenn man fordert, dass sie erklärbar sein müssen. Denn für eine nützliche Erklärung eines Algorithmus müsse man bewerten können, welche Erwägungen akzeptabel sind und welche nicht – also zum Beispiel woran festgemacht wird, ob ein Muttermal bösartig ist. Wenn man aber schon wüsste, welche Erwägungen akzeptabel sind und an welchen

883 Freiesleben, König 2023:12
884 Ebd.:3–12

Kriterien man dementsprechend feststellen kann, ob ein Muttermal bösartig ist, dann brauche man keinen Algorithmus mehr, der die Kriterien durch Mustererkennung selbst festlegt. Stattdessen könne man einem Wenn-Dann-Algorithmus die genauen Kriterien vorgeben.[885] An dieser Stelle widerspricht Buijsman: „(…) we often do seem to be capable of evaluating decisions even if we can't write down the exact steps needed for making or even evaluating the decision."[886] Wir wissen zum Beispiel, dass wir eine Katze sehen, aber sind trotzdem nicht dazu in der Lage, notwendige und hinreichende Bedingungen dafür anzugeben, wie eine Katze aussieht. Selbst wenn wir Erklärungen bewerten können und feststellen, ob der Algorithmus die richtigen Dinge betrachtet, um zu entscheiden, ob etwas eine Katze ist, bedeutet das nicht, dass wir stattdessen einen Wenn-Dann-Algorithmus spezifizieren können. Laut Buijsman kann deshalb Erklärbarkeit immer noch eine sinnvolle Anforderung an komplizierte Algorithmen sein.[887]

Als nächstes widmet sich Buijsman dem Argument des Technologieethikers Alex John London, der der Meinung ist, wir forderten von Algorithmen eine Erklärbarkeit, die Menschen für ihre Entscheidungen auch nicht bieten könnten. Die meisten menschlichen Entscheidungen seien nicht interpretierbar in dem Sinne, dass die Menschen ein Modell, das sie für die Entscheidungsfindung verwenden, simulieren könnten. Menschliche Entscheidungen seien oft in dem Sinne interpretierbar, dass wir sie im Nachhinein rationalisieren können, und das gebe nicht unbedingt Aufschluss darüber, warum eine Person die Entscheidung getroffen habe. In diesem Sinne seien auch Algorithmen interpretierbar. Aus Londons Sicht ist es dann sogar ein Vorteil von Algorithmen gegenüber menschlichen Entscheidungen, dass Verlässlichkeit und Genauigkeit der ersteren leicht bewertet werden können. Die beste Grundlage für die Entscheidung, welche Behandlung angeboten werden soll, sei die Wirksamkeit der Behandlungen. Daher sollten wir algorithmischen Ergebnissen aufgrund ihrer Wirksamkeit vertrauen und nicht aufgrund der Frage, ob wir sie genau kennen.[888]

885 Robbins 2019:495, 510ff.; Buijsman 2023:198
886 Buijsman 2023:198
887 Ebd.:199
888 London 2019:18ff.; Buijsman 2023:199f.

Buijsman wendet als erstes ein, dass Menschen sehr wohl erklären können, warum sie eine bestimmte Entscheidung getroffen haben. So können zum Beispiel Ärzt:innen, die eine Behandlung verordnen, dies mit früheren Erfahrungen mit Patient:innen mit ähnlichen Symptomen begründen. Algorithmen hingegen schätzen wir für ihre Zuverlässigkeit und nicht für ihre Fähigkeit, Entscheidungen zu begründen. Deshalb sei es wichtig, dass Algorithmen nur als Hilfsmittel verstanden werden, welche bei der Entscheidungsfindung unterstützen können. Buijsman schreibt über London: „(…) the cases he highlights show that making decisions for the wrong reasons (theoretical ones rather than evaluated efficacy) can be damaging, and so that having these reasons can be an important part of figuring out how good a decision is."[889] Aus Buijsmans Sicht liefert London damit unabsichtlich einen Grund, erklärbare Algorithmen zu bevorzugen, damit gezeigt werden kann, dass deren Entscheidungen auf den relevanten Merkmalen, wie zum Beispiel der Wirksamkeit einer empfohlenen Behandlung, beruhen.[890]

Der Einwand gegen die Erklärbarkeit von Algorithmen von Juan Manuel Durán und der Bioethikerin Karin Rolanda Jongsma lautet, dass undurchsichtige Algorithmen nur mit Hilfe von weiteren undurchsichtigen Algorithmen erklärt werden können: „The fundamental problem with transparency is, to our mind, that it is itself based on opaque processes."[891] Dadurch, dass ein weiterer Algorithmus angewendet wird, um den ursprünglichen, undurchsichtigen Algorithmus erklärbar zu machen, wird die Frage nach den Entscheidungskriterien nur auf den neuen, erklärenden Algorithmus verlagert. Dabei können wir wissen, wie dieser zweite Algorithmus Erklärungen generiert, aber wir können weiterhin nicht sicher sein, ob die Erklärungen des zweiten Algorithmus tatsächlich korrekte Beschreibungen dafür sind, warum der ursprüngliche Algorithmus zu seinen Ergebnissen gekommen ist.[892]

Buijsman stimmt zu, dass wir undurchsichtige Algorithmen nicht loswerden könnten, wenn alle Erklärungswerkzeuge selbst undurchsichtige Algorithmen wären, aber er belegt, dass das nicht der Fall

889 Buijsman 2023:200
890 Ebd.:200
891 Durán, Jongsma 2021:331
892 Ebd.:330f.; Buijsman 2023:201

ist. Denn Methoden der Merkmalsbedeutung (feature importance methods), wie zum Beispiel die Datenvisualisierungstechnik der Heatmaps, machen einen Teil des Verhaltens des Algorithmus transparent und kontrafaktische Erklärungen zeigen transparent auf, welche minimalen Änderungen zu einer Änderung der Ergebnisse führen. Auch wenn es noch keine Tools gibt, die Black-Box-Algorithmen vollständig erklären können, gibt es doch Erklärungsansätze, die das Vorgehen von Algorithmen transparenter machen.[893]

Zum Schluss widmet sich Buijsman dem Argument von Krishnan, welches besagt, dass Erklärungen weiteren Zielen dienen, wie dem Schutz vor Diskriminierung, und diese Ziele auch durch andere Mittel erreicht werden können. Sie argumentiert, dass die Interpretierbarkeit oft nicht um ihrer selbst willen wichtig ist, sondern sie dient der Erreichung eines weiteren Ziels, wie z. B. Sicherheit, Gewährleistung der Genauigkeit und Nichtdiskriminierung: „(…) since interpretability is most often proposed as a means to further ends rather than an end in itself, it would be more perspicuous to organize discussion around the fundamental problems rather than one putative solution."[894] Krishnan schlägt vor, dass anstatt über die Interpretierbarkeit zu diskutieren, über die Ziele der Interpretierbarkeit gesprochen werden sollte. Dadurch könnte nicht mehr angenommen werden, dass es nur einen Weg gibt, ein bestimmtes Ziel zu erreichen und Aufmerksamkeit sowie Ressourcen würden nicht mehr ausschließlich auf eine bestimmte Art von Lösung konzentriert.[895]

So werde zum Beispiel oft angenommen, dass die Interpretierbarkeit eine wichtige Rolle bei der Rechtfertigung von algorithmischen Ergebnissen spielt. Es erscheint geradezu als unverantwortlich, den Ergebnissen von solchen Algorithmen zu glauben, die auf unbekannten realen Daten basieren, wenn keine detaillierten Kenntnisse über die Funktionsweise des Algorithmus vorliegen. In manchen Kontexten gibt es jedoch laut Krishnan Möglichkeiten, die gewünschte Sicherheit zu erreichen, auch wenn man das Innenleben der Werkzeuge nicht kennt: „(...) there are ways of being justifiably confident in outputs that circumvent the call for interpre-

893 Buijsman 2023:202
894 Krishnan 2020:500
895 Ebd.:495

tation (...)."[896] Die Position der Reliabilisten liefere eine plausible Erklärung dafür, warum man sich auf die Ergebnisse des visuellen Systems oder des Gehirns verlassen kann, auch wenn es keine detaillierte wissenschaftliche Theorie gibt: weil das System zuverlässig das richtige Ergebnis produziere. In manchen Fällen könne eine Erfolgsbilanz eine genaue Beschreibung der inneren Funktionsweise eines Instruments ersetzen.[897]

Wenn die Ergebnisse diskriminierend sind, weil ein Gesichtserkennungs-Algorithmus zum Beispiel Menschen mit dunkler Hautfarbe als Gorillas identifiziert[898], dann ist das Problem laut Krishnan auch ohne detailliertes Wissen über das Vorgehen des Algorithmus diagnostizierbar. In diesem Fall waren die Ergebnisse eindeutig diskriminierend und als festgestellt wurde, dass in den Trainingsdaten keine Bilder von Menschen mit dunkler Hautfarbe enthalten waren, wurde auch klar, warum. Laut Krishnan ist es nicht notwendig, die Details des algorithmischen Vorgehens zu verstehen, um überprüfen zu können, ob der Algorithmus diskriminierende Ergebnisse erzeugt. Stattdessen könnten bessere Verfahren zur Überprüfung der Konstruktion von Trainingsdatensätzen oder zum Testen von Entscheidungskriterien anhand von Beispieldatensätzen entwickelt werden, um zu überprüfen, wie der Algorithmus Personen unterschiedlicher Identitätskategorien behandelt.[899] Der Wunsch nach Interpretierbarkeit wurde zwar durch die Notwendigkeit motiviert, Diskriminierung zu bekämpfen[900], aber es gibt keine offensichtliche Verbindung zwischen der Bekämpfung von Diskriminierung und dem Wissen über die innere Funktionsweise von Algorithmen. Vielmehr kann die Interpretierbarkeit als ein Werkzeug unter vielen verstanden werden, das sich als nützlich gegen Diskriminierung erweisen kann.[901] Allerdings lenkt Krishnan ein, dass es auch Kontexte gibt, in denen die Interpretierbarkeit des Algorithmus als Selbstzweck wichtig ist. So zum Beispiel, wenn Algorithmen in der

896 Krishnan 2020:496
897 Ebd.:495f.
898 Guynn 2015; Barr 2015
899 Krishnan 2020:496, 498
900 Goodman, Flaxman 2017:55; Krishnan 2020:497
901 Krishnan 2020:497

Wissenschaft eingesetzt werden, um kausal relevante Faktoren und ihre Wechselwirkungen zu identifizieren.[902]

Buijsman hat sich von Krishnan davon überzeugen lassen, dass Erklärbarkeit kein Allheilmittel ist. Er gesteht ein, dass es gerechtfertigt sein kann, algorithmischen Ergebnissen zu glauben, ohne vollständige Erklärung oder Interpretation, weil man stattdessen zum Beispiel die Trainingsdatensätze überprüfen kann. Trotzdem hält Buijsman eine Erklärung der Funktionsweise des Algorithmus für sinnvoll, um mit Unstimmigkeiten umgehen und Beweise aus verschiedenen Quellen miteinander in Einklang bringen zu können. Eine Erklärung könne außerdem bei der Fehlersuche helfen und dazu beitragen, die algorithmischen Ergebnisse anzufechten und Verantwortung für Fehler einzufordern.[903] Die Erklärbarkeit kann also einerseits einen instrumentellen Wert haben, da sie es den Datennutzenden ermöglicht, Algorithmen zu verwalten.[904]

Andererseits ist die Erklärbarkeit von algorithmischen Ergebnissen dem Datenethiker Nathan Colaner zufolge auch intrinsisch wertvoll, weil nicht erklärbare Systeme entmenschlichend sein können. Zum einen geht es Colaner dabei um die Möglichkeit der Partizipation am Entscheidungsprozess: „(…) personal dignity requires the opportunity for meaningful participation in the process when the outcome affects me or my community, and meaningful participation is difficult or impossible if the algorithmic process is fundamentally un-understandable, regardless of whether the outcome is morally and legally acceptable."[905] Bei einem unverständlichen Prozess, dessen Ergebnis eine Person betrifft, könne diese Person nicht oder nur schwer sinnvoll an diesem Prozess beteiligt werden, und das verletze ihre persönliche Würde. Zum anderen erfordere die persönliche Würde eine Möglichkeit zu verstehen, warum ein Ereignis eingetreten ist, und das sei ebenfalls schwer zu erreichen, wenn der algorithmische Prozess grundsätzlich unverständlich ist.[906] Drittens erfordert die vollständige Verwirklichung des Menschen laut Colaner, dass wir in bestimmten Fällen die Verantwortung für

902 Krishnan 2020:499
903 Buijsman 2023:205
904 Colaner 2022:231f.
905 Ebd.:235
906 Ebd.:236

unser eigenes Leben behalten, unabhängig davon, ob ein Algorithmus dies effizienter tun könnte.[907]

Ich möchte dem hinzufügen, dass algorithmische Ergebnisse auch diskriminierend sein können, ohne dass Datennutzende dies bemerken, weil bei einem Bewerbungsprozess zum Beispiel nicht alle Bewerber:innen eingeladen werden können. Wenn die Ergebnisse unbemerkt diskriminierend sind, ist das Problem aus meiner Sicht nicht ohne Wissen über die Trainingsdaten und über die Gewichtung verschiedener Entscheidungskriterien durch den Algorithmus diagnostizierbar. Man braucht vielleicht keine detaillierten Informationen über die Funktionsweise des Algorithmus, aber ich habe argumentiert, dass man eine Antwort auf die Frage braucht, warum diese Ergebnisse erzeugt wurden, um Beweise sammeln zu können für die Verlässlichkeit von algorithmischen Ergebnissen.

7.2.2 Definition von Transparenz im Kontext von Algorithmen

In Bezug auf Algorithmen bedeutet Transparenz, dass zugänglich und verständlich sein muss, wie das Ergebnis zustande gekommen ist. Man könnte also meinen, es müssten technische Details über die Funktionsweise der Algorithmen offenbart werden.[908] Es stellt sich also die Frage, was Transparenz in Bezug auf Algorithmen heißen soll, wenn es nicht möglich ist, alle technischen Details offenzulegen und das zudem auch nicht erstrebenswert erscheint, weil es sehr schwer wäre diese technischen Details zu verstehen. Deshalb muss gefragt werden, was eine solche Erklärung für algorithmische Ergebnisse leisten soll. Sie soll verhindern, dass Datennutzende verzerrten Ergebnissen ausgesetzt werden. Sie soll also Fehler aufdecken. Ich argumentiere, dass Fehler in algorithmischen Ergebnissen aufgedeckt werden können, wenn klar ist, warum dieses Ergebnis erzeugt wurde.

Die Frage, warum ein Algorithmus ein bestimmtes Ergebnis erzeugt hat, kann laut Krishnan auf unterschiedliche Weise verstanden werden: im kausalen oder im rechtfertigenden Sinne. Wir fragen nach dem Warum im kausalen Sinne, wenn wir fragen, wie es dazu

907 Colaner 2022:237
908 Tutt 2017:111; Herzog 2022:4

kam, dass der Algorithmus das Ergebnis erzeugte, das er erzeugte. Wir fragen jedoch nach dem Warum im rechtfertigenden Sinne, wenn wir fragen, welche Gründe für das Ergebnis als Antwort auf eine bestimmte Frage sprechen.[909] Wenn zum Beispiel der von Amazon im Bewerbungsprozess eingesetzte Algorithmus eine Bewerberin ablehnt, kam es zu dem Ergebnis wegen ihres Geschlechts (kausal), und der Grund für das Ergebnis ist, dass in der Vergangenheit vorrangig Männer erfolgreich waren im Bewerbungsprozess und der Algorithmus dies reproduziert (rechtfertigend). Hier zeigt sich, dass sich die kausale Beschreibung eines algorithmischen Ergebnisses nicht ohne Weiteres auf rechtfertigende Überlegungen übertragen lässt. Denn selbst wenn man weiß, was den Algorithmus dazu veranlasst hat, eine Ablehnung zu empfehlen, heißt das noch nicht, dass man auch den Grund für die Ablehnung angeben kann.

Deshalb bezweifelt Krishnan, dass eine Interpretation – im Sinne von Informationen über den Prozess des Algorithmus – hilfreich ist für das Verständnis der algorithmischen Ergebnisse.[910] Ich hingegen ziehe aus dieser Unterscheidung zwischen dem kausalen und dem rechtfertigenden Warum die Erkenntnis, dass man zum Verständnis algorithmischer Ergebnisse nicht nur die Ergebnisse und ihren Entstehungsprozess betrachten muss, sondern immer auch die ihnen zugrundeliegenden Datensätze (Input). Auf diese Weise können Verzerrungen identifiziert werden, die sich bei der Datenverarbeitung in den Ergebnissen reproduzieren würden. So kann geprüft werden, ob die Daten die Wirklichkeit abbilden und ob es sich dabei um eine ungerechte Wirklichkeit handelt, die nicht reproduziert werden sollte. Die Analyse der algorithmischen zugrundeliegenden Datensätzen ist somit auch ein wichtiger Schritt, damit bereits bestehende Vorurteile nicht in Form von einer bestehenden Verzerrung (preexisting bias) in das System einfließen.[911]

In Bezug auf Algorithmen bedeutet Transparenz also weiterhin, dass zugänglich und verständlich sein muss, wie das Ergebnis zustande gekommen ist.[912] Post-hoc Erklärungen von Algorithmen sollen zu diesem Zweck aufzeigen, warum der Algorithmus eine

909 Krishnan 2020:493
910 Ebd.:494
911 Friedman, Kahn 2008:1253; Friedman, Nissenbaum 1996:332–335
912 Mittelstadt et al. 2016:6

bestimmte Ausgabe erzeugt. Sie tun das, indem das Verhältnis zwischen Input und Output offengelegt wird. Dadurch ist es möglich, die Frage nach dem Warum im kausalen Sinne zu beantworten. Um die Frage nach dem Warum im rechtfertigenden Sinne beantworten zu können, müssen die dem zu erklärenden Algorithmus zugrundeliegenden Datensätze betrachtet werden – denn nur diese bieten Aufschluss darüber, ob Verzerrungen vorliegen und welche Aussagen eigentlich gemacht werden können.[913] Durch diese Transparenz der Ergebnisse in Form von Erklärungen der Zusammenhänge zwischen Input und Output, sowie der Analyse der zugrundeliegenden Datensätze, können Verzerrungen und Fehler in den Ergebnissen aufgedeckt werden und somit kann Diskriminierung auf Grund bestimmter Merkmale im Input verhindert werden. Ich schließe, dass man hierdurch vollständig versteht, wie das System seinen Output erreicht, weil man alle epistemisch relevanten Elemente des Prozesses kennt, wenn man den Input kennt und weiß, welchen Output er erzeugt – so reduziert sich die epistemische Undurchsichtigkeit.[914]

7.2.3 Einbettung in den Kontext als Lösung

Trotz aller Widrigkeiten halte ich diese Transparenz im Kontext von Algorithmen für umsetzbar, indem die Ergebnisse im zum Verständnis notwendigen Kontext dargestellt werden. In diesem Teilkapitel wird erörtert, warum das die beste Lösung ist.

Man könnte meinen, wie unter 7.1 angedeutet, dass es zum Schutz der Datennutzenden und Vorbeugen von Verzerrungen sinnvoll wäre, sensible Daten von vorneherein nicht in Trainingsdatensätze oder Datenabfragen zu integrieren. Im Falle einer Berufsempfehlung und auch beim Wohngeldantrag sollten die Ergebnisse aus ethischer Sicht nicht abhängig sein von dem Geschlecht oder der Herkunft der Datenquellen und -nutzenden – hier wäre es durchaus möglich, solche sensiblen Daten gar nicht erst abzufragen und auch aus den Trainingsdatensätzen zu entfernen. Aber in anderen Fällen, zum Beispiel im medizinischen Bereich, können sensible Daten, wie das Geschlecht, ausschlaggebend dafür sein, wie hilfreich das algorith-

913 Mehr dazu im nächsten Teilkapitel unter 7.2.3.
914 Humphreys 2009:618; Buijsman, Veluwenkamp 2023:545

mische Ergebnis ist. In der Vergangenheit wurden hier solche Merkmale sogar zu wenig beachtet, sodass erst neueste Erkenntnisse auf Unterschiede in Krankheitsverläufen sowie effektiven Behandlungsmöglichkeiten zwischen Frauen und Männern hindeuten.[915] Es ist also nicht in allen Fällen ratsam, sensible Daten, wie das Geschlecht, außen vor zu lassen – auch weil man mit wenigen unspezifischen Daten nur unspezifische Ergebnisse erwarten kann und das dem Zweck des Einsatzes der Algorithmen zu widersprechen scheint.

Wenn Datenverarbeitende die Verzerrung aus den Trainingsdaten nicht rausrechnen können und verzerrte Ergebnisse auch anderweitig nicht vermeiden können, haben sie die Pflicht, Datennutzende darüber aufzuklären. So könnte man im Fall einer Berufsempfehlung zum Beispiel die Datennutzenden darüber aufklären, dass in der Vergangenheit vor allem Männer berufstätig waren, und die Trainingsdaten, auf welchen der Algorithmus basiert, deshalb 80 % Männer und 20 % Frauen repräsentieren. Die Datennutzenden, die sich durch diese Trainingsdaten nicht oder nur schlecht repräsentiert fühlen, haben daraufhin eine Erklärung dafür und einen Anlass, die Ergebnisse des Algorithmus zu hinterfragen. Insbesondere sollte durch eine solche Aufklärung für die Datennutzenden deutlich werden, was der Algorithmus leisten kann und was nicht. Denn ein Algorithmus, der auf beruflichen Werdegängen von Menschen in der Vergangenheit beruht, kann nicht vorhersagen, welcher berufliche Weg der Beste sein wird für eine Person heute. Erstens verändern sich Ausbildungswege und ganze Berufsfelder durch den Einsatz von neuen Technologien. Zweitens sagt zum Beispiel die Anzahl von Frauen in bestimmten Berufsgruppen nichts darüber aus, wie viele Frauen in Zukunft in diesen Berufsgruppen arbeiten sollten – das wäre ein naturalistischer Fehlschluss.[916]

Ein solcher Algorithmus kann nur als Anstoß dienen für diejenigen, die noch gar keine Idee haben, in welche Richtung es beruflich gehen könnte und welche Optionen es eigentlich gibt. Selbst dann kann ein Algorithmus, der auf beruflichen Werdegängen von Menschen in der Vergangenheit beruht, nur aussagen „Angesichts der Tatsache, dass in der Vergangenheit die meisten Menschen mit deinen Voraussetzungen folgende Berufe ergriffen haben, ist es emp-

915 Robert Koch-Institut 2020
916 Hume 2000[1739]:Buch III, Teil I, Kapitel I

fehlenswert, dass du auch einen dieser Berufe ergreifst". Es hängt dann von den Trainingsdaten ab, ob man überhaupt eine Aussage darüber machen kann, ob die Menschen in der Vergangenheit zufrieden waren mit ihren beruflichen Entscheidungen. Wenn es dazu keine Informationen gibt, ist es nicht einmal möglich, das Kopieren des Verhaltens anderer Menschen als Empfehlung zu bezeichnen – es handelt sich vielmehr um eine Aussage darüber, was in der Vergangenheit mit den gegenwärtigen Qualifikationen und Voraussetzungen möglich war. Durch solch eine transparente Darstellung der Leistung des Algorithmus und seiner Ergebnisse wird für Datennutzende klar, dass es auch andere Möglichkeiten gibt, als der Empfehlung zu folgen. So kann womöglich die epistemische Abhängigkeit vom Algorithmus verringert werden.[917]

Hier wird noch einmal deutlich, für wen die Ergebnisse transparent sein sollen, nämlich für die Datennutzenden. Aber das heißt nicht, dass jedem Datennutzenden die Struktur des verwendeten Algorithmus offenbart werden soll. Der Durchschnittsnutzende verfügt nicht über fundierte Kenntnisse der Informatik und hat möglicherweise nicht einmal ein starkes Interesse daran, diese zu erlernen.[918] Selbst wenn Datenverarbeitende betriebliche Informationen offenlegen, ist der Nutzen für die Gesellschaft deshalb laut Mittelstadt et al. ungewiss: „Transparency disclosures may prove more impactful if tailored towards trained third parties or regulators representing public interest as opposed to data subjects themselves (…)."[919] Auch die Sozioinformatikerin Katharina Zweig plädiert dafür, dass geschulte Dritte Einblick in das algorithmische Entscheidungssystem bekommen. Sie schlägt vor, dass dies ein ausgewählter Kreis von Expert:innen sein sollte – insbesondere wenn ein Algorithmus schwerwiegende Fehlurteile erzeugt hat. Aber wichtiger noch ist in ihren Augen, dass sowohl Trainingsdaten und -kriterien als auch Wirkungsweise und Entscheidungsregeln transparent dargestellt werden. Laut Zweig ist es immer sinnvoll, über Trainingsdaten und -kriterien solcher Algorithmen aufzuklären, die Menschen in Kategorien einteilen oder Urteile fällen über ihr zukünftiges Verhal-

917 Van den Hoven 1998:104f.; Rooksby 2009:82; Buijsman, Veluwenkamp 2023:546

918 Granka 2010:368

919 Mittelstadt et al. 2016:7

ten. Wenn aber die Wirkungsweise und Entscheidungsregeln des Algorithmus dargestellt werden sollen, schließe das die Methoden des maschinellen Lernens aus, deren Entscheidungsregeln selbst für die Programmierenden nicht mehr nachvollziehbar sind.[920]

Denn in solchen Algorithmen können mehr Merkmale aufgenommen werden als Menschen erfassen können, sodass ein Verständnis der Algorithmen, samt Fehlersuche und Verbesserung, sehr schwierig ist. Laut Burrell kann mit dieser Art von Undurchsichtigkeit umgegangen werden, indem Algorithmen vereinfacht werden. So ist es zum Beispiel möglich zu analysieren, welche Merkmale für das Ergebnis wichtig sind, und alle anderen Merkmale aus dem Modell zu entfernen.[921] Die Vereinfachung von Algorithmen würde auch Zweigs Forderung nach Wirkungsweise und Entscheidungsregeln des Algorithmus ermöglichen. Auch Mittelstadt et al. halten dies für hilfreich zum Verständnis: „Releasing information about the algorithm's decision-making logic in a simplified format can help (…)."[922] Wenn aber das Vorgehen der Algorithmen heruntergebrochen wird auf eine für den Menschen überschaubare Liste von Schlüsselkriterien, wird laut Burrell nur ein unvollständiges Verständnis vermittelt.[923]

Die bereits erwähnte KI-Ethikerin Kate Vredenburgh stimmt Burrell zu und wendet ein, dass es nicht ausreicht, vereinfachte Modelle von komplexen Algorithmen bereitzustellen, da es hierdurch zu irreführenden Erklärungen kommen kann und somit nur eine Illusion des Verstehens erzeugt wird.[924] Der Technologieethiker Christian Herzog merkt an, dass interpretierbare Algorithmen meist nur für ihre Entwickler:innen verständlich sind, bei einem solchen interpretierbaren Modell dann aber die Bemühungen fehlen, die Ergebnisse anderen Akteuren zu erklären.[925] Munch et al. geben zudem zu bedenken, dass es potenziell kognitiv nicht möglich ist, Verständnis von algorithmischen Ergebnissen zu erlangen, indem sie individuell mit Hilfe von XAI-Werkzeugen erklärt werden.[926] Der Soziologe

920 Zweig 2019:9
921 Burrell 2016:9
922 Mittelstadt et al. 2016:7
923 Burrell 2016:9
924 Vredenburgh 2022:225
925 Herzog 2021:222
926 Munch et al. 2024:9

Daan Kolkman bestätigt diese Bedenken, da er bei Fallstudien mit Algorithmusexpert:innen festgestellt hat, dass selbst vergleichsweise einfache Modelle sogar für diejenigen, die täglich mit ihnen arbeiten, schwer zu verstehen sein können.[927] Insbesondere für Menschen, die nicht täglich mit Algorithmen arbeiten, ist deshalb das Verständnis ihres Vorgehens und damit der Ergebnisse eine Herausforderung.[928]

Während Vredenburgh als Lösung vorschlägt, dass Expert:innen kostenlos bereitgestellt werden sollten[929], schlägt Kolkman vor, dass es eine bessere Kommunikation über Algorithmen für die breite Öffentlichkeit geben sollte[930]. Die Ansätze beider Autor:innen ergänzen sich, denn man braucht Expert:innen für die Erklärung der Algorithmen und anschließend eine gute Kommunikationsstrategie, um diese Erklärung der breiten Öffentlichkeit verständlich zu machen. Nimmt man noch die Lösung von Munch et al. hinzu, bräuchte man sogar mehrere Kommunikationsstrategien, da sie vorschlagen, Erklärungen auf verschiedene Personengruppen zuzuschneiden.[931] Die Hoffnung ist, dass es durch eine solche transparente Darstellung für Datennutzende möglich ist, alle epistemisch relevanten Elemente des Prozesses zu kennen.

Buijsman und der Digitalethiker Herman Veluwenkamp behaupten jedoch, dass es auf Grund der epistemischen Abhängigkeit nicht einmal reicht, alle epistemisch relevanten Elemente des Prozesses zu kennen, um epistemisch unabhängig zu sein: „It is possible to know all the epistemically relevant elements of the process (i.e. fully understand how the system reaches its output) while still being dependent on the system, because of a lack of outside information."[932] Wenn alternative Informationsquellen fehlen, sind die Datennutzenden auch dann noch epistemisch abhängig von den algorithmischen Ergebnissen, wenn sie deren Entstehung nachvollziehen können. Nun erscheint es zu viel verlangt, dass Datenverarbeitende neben den algorithmischen Ergebnissen auch noch alternative Informationsquellen bereitstellen.

927 Kolkman 2022:105f.
928 Ebd.:106
929 Vredenburgh 2022:225
930 Kolkman 2022:106
931 Munch et al. 2024:9
932 Buijsman, Veluwenkamp 2023:545

Deshalb schlage ich vor, dass die Ergebnisse im zum Verständnis notwendigen Kontext dargestellt werden sollten. So kann zum Beispiel das Ergebnis eines Berufsempfehlungs-Algorithmus, der auf Trainingsdaten aus der Vergangenheit basiert, durch die Information eingeordnet werden, dass der Algorithmus nur reproduzieren kann, was in der Vergangenheit mit den gegenwärtigen Qualifikationen und Voraussetzungen möglich war. Dafür ist es natürlich notwendig, dass die Datenverarbeitenden nachvollziehen können, wie das Ergebnis zustande gekommen ist, bzw. welcher Input zu welchem Output führt. Es muss für Datennutzende klar sein, dass es sich bei algorithmischen Ergebnissen nicht um objektiv gesetztes Wissen handelt. Diese Tatsache erfordert einen reflektierten Umgang mit algorithmischen Ergebnissen auf Seiten der Datennutzenden. Dieser sollte unterstützt werden durch die Datenverarbeitenden, indem sie den Kontext der Ergebnisse möglichst verständlich einordnen.

7.3 Moralische Pflicht zur Transparenz und daraus abgeleitetes Recht auf Transparenz

Eine Pflicht zur Transparenz sowie ein daraus abgeleitetes Recht auf Transparenz würden vor Handlungen schützen, die das Risiko einer erheblichen Beeinträchtigung der Transparenz mit sich bringen. Eine solche moralische Pflicht und das daraus resultierende Recht auf Transparenz schützen also vor Handlungen, die Verzerrungen und Fehler in den algorithmischen Ergebnissen verdecken und somit Diskriminierung schwer erkennbar machen. Das Ziel dieses Kapitels ist es, eine moralische Pflicht zur transparenten Darstellung algorithmischer Ergebnisse durch Datenverarbeitende zu begründen und zu verstehen, wie daraus für Datennutzende ein Recht auf Transparenz resultieren kann.

Solch eine moralische Pflicht lässt sich aus Kants erster Formulierung des Kategorischen Imperativs ableiten.[933] Denn wenn eine Handlungsmaxime weder gedacht noch gewollt werden kann, ist es eine vollkommene Pflicht, diese Handlung zu vermeiden. Eine Handlungsmaxime, die zwar denkbar ist, aber nicht widerspruchs-

933 Kant 2008:53 [421]

frei gewollt werden kann, etabliert eine unvollkommene Pflicht.[934] Es ist zwar denkbar, dass algorithmische Ergebnisse intransparent zur Verfügung gestellt werden, aber es kann nicht gewollt werden – das etabliert eine unvollkommene Pflicht zum Unterlassen der Bereitstellung intransparenter Ergebnisse. Diese Pflicht sowie eine damit einhergehende Pflicht zur Bereitstellung transparenter Ergebnisse lassen sich begründen, indem aufgezeigt wird, warum die fehlende Transparenz vermieden werden sollte und warum eine transparente Darstellung algorithmischer Ergebnisse erstrebenswert ist.

Vredenburgh argumentiert, dass ohne Erklärungen von Entscheidungen, die ein Individuum betreffen, die Handlungsfähigkeit des Individuums verminderte wäre – insbesondere würde es dem Individuum erschwert, sich für sich selbst einzusetzen: „The right to explanation is grounded in the interest in informed self-advocacy."[935] Diese informierte Selbstvertretung umfasst diverse Fähigkeiten, die man braucht, um die eigenen Interessen und Werte zu vertreten. Eine Person, die sich selbst vertritt, kann Regelsysteme so steuern, dass die eigenen Ziele erreicht und Entscheidungstragende für Fehler oder Ungerechtigkeiten zur Rechenschaft gezogen werden. Um sich selbst vertreten zu können, brauchen Individuen epistemische Unterstützung. Laut Vredenburgh würde durch eine kausale Erklärung einer eine Person betreffende Entscheidung klar, was das Individuum tun müsste, um ein von ihm gewünschtes Ergebnis zu erzielen.[936] Das Konzept der informierten Selbstvertretung betont die Bedeutung von Erklärungen für algorithmische Entscheidungen, die uns direkt betreffen, denn ohne solche Erläuterungen sind wir nicht in der Lage, unser Verhalten als Reaktion auf das algorithmische Ergebnis so anzupassen, dass unsere Interessen und Werte gefördert werden.[937]

---

934 Kant 2008:55-58 [422–424]

935 Vredenburgh 2022:212

936 Ebd.:213, 223f., 217

937 Munch et al. 2024:4f.; Vredenburgh 2022; Vredenburgh begründet mit diesem Argument ein Recht auf Erklärung, aber die Argumente für eine durch eine Erklärung transparente und gegen eine intransparente Darstellung algorithmischer Ergebnisse können ebenfalls zu der Schlussfolgerung führen, dass es eine Pflicht zur Transparenz geben sollte. Da in dieser Arbeit davon ausgegangen wird, dass ein Recht auf Transparenz nur aus einer moralischen Pflicht zur Transparenz resultieren kann (Mohr 2010:77), ist hier die Schlussfolgerung,

Ohne das Verständnis, warum ein Ergebnis zustande gekommen ist, können Datennutzende auch laut London nicht reflektiert Stellung dazu nehmen und das Ergebnis nicht anfechten: „(…) understanding why a decision was made enables stakeholders to evaluate its merits and hold experts accountable for avoidable error."[938] Durch die Bereitstellung von intransparenten Ergebnissen, können die Datennutzenden diese nicht einordnen und deshalb nicht reflektiert Stellung dazu nehmen und ihre eigenen Interessen nicht vertreten. Wenn man Datennutzende mit Ergebnissen konfrontiert, die sie einfach hinnehmen sollen, würdigt man nicht ihre Rationalität, sondern gebraucht sie sogar als bloßes Mittel. Denn ohne das Verständnis der Ergebnisse zu ermöglichen, unterbindet man ihre freiwillige und informierte Zustimmung. Das heißt fehlende Transparenz sollte in Bezug auf algorithmische Ergebnisse vermieden werden, weil Datennutzende dadurch als bloßes Mittel gebraucht werden und das widerspricht Kants dritter Formel des Kategorischen Imperativs.[939]

Der Philosoph Keith Harris bemerkt passend dazu, dass die Beschränkung des Zugangs zu Beweisen bereits als epistemische Dominanz (epistemic domination) gilt.[940] Diese Beziehung epistemischer Dominanz kann den Datennutzenden epistemisch zugutekommen, wenn die Datenverarbeitenden wohlmeinend und gut informiert sind. Wenn aber die Datenverarbeitenden schlechte Absichten haben, ist es wahrscheinlich, dass die Datennutzenden zu falschen Überzeugungen gelangen und womöglich die Orientierung verlieren, da sie nicht mehr wissen, was sie glauben können.[941] Da es für das menschliche Wohlergehen wichtig ist, dass wir uns von richtigen und nicht von falschen Annahmen leiten lassen[942], sollten Datenverarbeitende dafür sorgen, dass möglichst keine falschen Überzeugungen entstehen.

Eine transparente Darstellung algorithmischer Ergebnisse ist hingegen erstrebenswert, weil sich die Datennutzenden dadurch des Risikos bewusst werden, dem sie ausgesetzt sind, und idealerweise

---

dass es eine Pflicht zur transparenten Darstellung algorithmischer Ergebnisse geben sollte.

938 London 2019:16

939 Kant 2008:65 [429]

940 Harris 2022:136

941 Ebd.:137, 139

942 Alston 2005:30

auch die epistemische Asymmetrie ausgeglichen wird, sodass verzerrte Ergebnisse nicht mehr so leicht zu falschen Überzeugungen führen können. Wenn zum Beispiel ein Wohngeldantrag abgelehnt wird und diese Ablehnung für die Antragstellerin in der Rolle der Datennutzenden nicht nachvollziehbar ist, dann sollte es möglich sein herauszufinden, warum der Wohngeldantrag abgelehnt wurde. Denn die Datennutzenden sind auf Informationen über das Vorgehen des Algorithmus angewiesen, um darauf angemessen reagieren zu können.[943] Denn mit Hilfe von der transparenten Darstellung algorithmischer Ergebnisse können Datennutzende eine fundierte Entscheidung darüber treffen, wie sie mit den betreffenden Ergebnissen umgehen[944], und können auf Grundlage dieser Informationen ihre eigenen Interessen vertreten[945].

Auch wenn es bei transparenter Darstellung möglich ist, ist es schwierig für Datennutzende die Ergebnisse selbst überprüfen zu müssen. Es ist leichter für die Datenverarbeitenden, Algorithmen einzusetzen, die zu keinen verzerrten Ergebnissen führen. Hierbei können natürlich auch Fehler unterlaufen, aber wichtig ist, dass Datennutzende nicht bewusst falschen Ergebnissen ausgesetzt werden. Wenn Datenverarbeitende undurchsichtige Algorithmen programmieren, die auswählen, welche Ergebnisse Datennutzenden zugänglich sind, dann haben sie die Möglichkeit, die Datennutzenden einseitig zu informieren oder ihnen nur die Daten zur Verfügung zu stellen, die ein erwünschtes Verhalten generieren. Ein solches strategisches Vorenthalten von Informationen wäre ethisch problematisch, da die Datennutzenden dadurch manipuliert und instrumentalisiert würden. Es wäre aber durchaus möglich, Informationen über die algorithmischen Ergebnisse strategisch vorzuenthalten, da eine Informationsasymmetrie zwischen Datennutzenden und Datenverarbeitenden besteht.[946]

Eine Erklärung, im Sinne der Offenlegung des Verhältnisses zwischen Input und Output, sowie der Analyse des zugrundeliegenden Datensatzes, ermöglicht die transparente Darstellung der algorithmischen Ergebnisse – die Ergebnisse werden so im zum Verständnis notwendigen Kontext dargestellt. Erst diese transparente Darstellung

943 Munch et al. 2024:8
944 Ebd.:9
945 Vredenburgh 2022:212
946 Drew 1991:25

der Erklärung ermöglicht die informierte Selbstvertretung der Datennutzenden. Da die transparente Darstellung der Ergebnisse erstrebenswert ist, und fehlende Transparenz bei der Bereitstellung algorithmischer Ergebnisse nicht gewollt werden kann, haben Datenverarbeitende, welche öffentlich Informationen bereitstellen, eine Pflicht zur Transparenz.

Dabei sind laut Hare aus deontologischer Sicht genau die Handlungsfolgen relevant, welche die moralischen Grundsätze verbieten oder herbeiführen wollen.[947] Somit ist das Bereitstellen von intransparenten algorithmischen Ergebnissen unter anderem deshalb falsch, weil es die unbemerkte Diskriminierung bestimmter Gesellschaftsgruppen zur Folge haben kann. Der Grundsatz des Kategorischen Imperativs verbietet die Handlungsfolge der Diskriminierung, weil die zugrundeliegende Handlungsmaxime nicht gewollt werden kann und somit eine unvollkommene Pflicht entsteht, diese Handlung zu vermeiden. Entsprechend dieses Grundsatzes ist es moralisch relevant, wenn die Folge einer Handlung die Diskriminierung einer Gesellschaftsgruppe ist.[948]

Darum lässt sich die Pflicht zur Transparenz auch dadurch begründen, dass fehlende Transparenz dazu führen kann, dass bestimmte Gruppen auf Grund von Merkmalen benachteiligt werden, die eigentlich keine Rolle spielen sollten.[949] So wurden zum Beispiel von dem Algorithmus bei Amazon im Jahr 2014 Bewerberinnen auf Grund ihres Geschlechts nicht eingeladen[950], und jobsuchende Frauen haben im Jahr 2020 in Österreich auf Grund ihres Geschlechts mit geringerer Wahrscheinlichkeit Fördermaßnahmen finanziert bekommen[951]. Im Fall des Algorithmus zur Bewertung von Bewerbungen bei Amazon wäre bei transparenten Entscheidungskriterien schneller klar geworden, dass das KI-System die Bewerber:innen anhand ihres Geschlechts einordnet, und dass es sich bei dieser Unterscheidung um eine Verzerrung handelt. Denn es wurde weiblichen Bewerberinnen systematisch aus unvernünftigen und unangemessenen Gründen eine Chance vorenthalten, sodass

947 Hare 1993:15
948 Ebd.
949 Stahl et al. 2023:10f.
950 Ebd.:10f.
951 Wimmer 2019; Köver 2019; Pflügl 2024

sich ein ungerechtes Ergebnis eingestellt hat.[952] Das größte Problem dabei ist, dass solche Ergebnisse trotzdem als objektiv dargestellt werden: „(...) biased decisions are presented as the outcome of an objective algorithm."[953] Das schürt jedoch die falsche Erwartung, dass KI-Systeme Wissen im Sinne von wahren und gerechtfertigten Überzeugungen bereitstellen können – dabei handelt es sich nur um die Reproduktion von Mustern in den Trainingsdaten. Es kann nicht gewollt werden, dass algorithmische Ergebnisse verzerrt sein und deshalb zu Diskriminierung führen können, dies jedoch nicht offensichtlich ist, sodass die Datennutzenden algorithmische Ergebnisse nicht auf ihre Plausibilität überprüfen können und ihnen damit willkürlich ausgeliefert sind – insbesondere, wenn Entscheidungen von den oder über die Datennutzenden direkt von den Ergebnissen beeinflusst werden oder sogar abhängen. Deshalb sollten Datenverarbeitende beim Einsatz von Algorithmen und bei der Bereitstellung ihrer Ergebnisse eine Pflicht zur Transparenz haben.[954]

Außerdem ist laut Hare aus deontologischer Sicht das Bereitstellen von intransparenten algorithmischen Ergebnissen deshalb falsch[955], weil sich Datennutzende auf dieser Informationsgrundlage eine falsche Überzeugung bilden können und es schwierig ist, das Ergebnis anzufechten. Ohne die Möglichkeit dem Ergebnis zuzustimmen oder es abzulehnen, in Ermangelung an Informationen, werden die Datennutzenden als bloße Mittel gebraucht, die die Ergebnisse einfach nur akzeptieren sollen. Die dritte Formel des Kategorischen Imperativs verbietet aber, dass Datennutzende als bloßes Mittel gebraucht werden. Diese Konsequenz, Datennutzenden die Möglichkeit zu verwehren zu den Ergebnissen reflektiert Stellung zu nehmen (und sie dadurch bloß als Mittel zu gebrauchen), ist das, was das Bereitstellen intransparenter Ergebnisse aus deontologischer Sicht falsch macht.

Aus dieser vollkommenen moralischen Pflicht zur transparenten Darstellung algorithmischer Ergebnisse auf Seiten der Datenverarbeitenden kann ein Primärrecht auf Transparenz für Datennutzende resultieren.[956] In der Literatur ist die Ansicht weit verbreitet, dass

952 Friedman, Nissenbaum 1996:332
953 Goodman, Flaxman 2017:53
954 Tutt 2017:110, 116f; Mittelstadt et al. 2016:6
955 Hare 1993:15
956 Mohr 2010:77; Mieth, Bambauer 2016:3

Menschen ein Recht auf Erklärung haben, wenn sie algorithmischen Entscheidungen unterworfen sind, bei denen viel auf dem Spiel steht.[957] So argumentieren zum Beispiel die Epistemologen Jeremy Fantl und Matthew McGrath, dass die Tatsache, ob eine Überzeugung epistemisch gerechtfertigt ist, nicht nur von den Beweisen abhängt, die wir für die Wahrheit dieser Überzeugung haben, sondern zudem auch von pragmatischen Faktoren.[958] Der Gedanke findet sich auch bei Dormandy: „Könnte die falsche Handlung z. B. einen Todesfall verursachen, so benötigen wir eine viel bessere Rechtfertigung für die Überzeugungen, die unsere Handlungsbasis bilden, als wenn die schlimmste Konsequenz wäre, dass wir uns mit Vanilleeis statt mit Schokoladeneis zufriedengeben müssten."[959] Wenn also ein Algorithmus im medizinischen Bereich eingesetzt wird, und das Befolgen seiner Empfehlung zum Tod eines Patienten führen könnte, müsse man sich sicherer sein, dass der Glaube an die Empfehlung gerechtfertigt ist, als wenn ein Algorithmus empfiehlt, welche Kleidung man sich kaufen sollte.

Darauf erwidern jedoch Munch et al., dass auch algorithmische Entscheidungen mit geringem Risiko ein Recht auf Erklärung begründen können, wenn sie Teil eines Musters von Entscheidungen sind, deren Ergebnis insgesamt gesehen für den Einzelnen von erheblicher Bedeutung sein kann.[960] Ein Beispiel hierfür wäre die durch einen Algorithmus auf die Nutzenden maßgeschneiderte Werbung auf einer Social-Media-Plattform. Mit Hilfe des Algorithmus und Informationen aus den jeweiligen Profilen wird die Wahrscheinlichkeit optimiert, dass sich Nutzende mit der angezeigten Werbung auseinandersetzen. Die Nutzenden verstehen dabei nicht, nach welchen Kriterien der Algorithmus vorgeht, und durch die wiederholte Anzeige dieser gezielten Werbung über einen gewissen Zeitraum entwickeln die Nutzenden, eine Vorliebe für teuren Spezialitätenkaffee.[961] Einmaliger Kontakt mit der Kaffeewerbung hat wenig Einfluss auf das Verhalten der Nutzenden und birgt deshalb nur ein geringes Risiko. Wenn aber die Nutzenden die Werbung immer wieder wahr-

957 Munch et al. 2024:1; Herzog 2022:50
958 Fantl, McGrath 2002:88
959 Dormandy 2019:184
960 Munch et al. 2024:2
961 Ebd.:7

nehmen, kann das zu einer Veränderung ihres Verhaltens führen, was als erheblicher Effekt und damit größeres Risiko einzuschätzen ist.[962]

Die Nutzenden können nur dann eine fundierte Entscheidung darüber treffen, ob sie mit dem Algorithmus interagieren möchten, wenn sie wissen, wozu der Algorithmus eingesetzt wird. Ein Recht auf Erklärung besteht laut Munch et al. dann, wenn es sich um algorithmische Entscheidungen mit hohem Risiko handelt, oder wenn algorithmische Entscheidungen mit geringem Risiko Teil eines Musters von Entscheidungen sind, deren Ergebnis insgesamt von erheblicher Bedeutung sein kann.[963] Laut Munch et al. brauchen wir also Informationen über die gezielten Einwirkungen der Algorithmen auf Menschen, um bestimmen zu können, ob sie ein Recht auf eine Erklärung für einmalige algorithmische Entscheidungen mit geringem Risiko haben.[964] Im Sinne von Mohr ließe sich jedoch argumentieren, dass allein aus der moralischen Pflicht zur transparenten Darstellung algorithmischer Ergebnisse Primärrechte für Datennutzende resultieren, die in der Rechtsordnung verankert werden sollten.[965]

7.3.1 Rechtliche Lage im Vergleich zu ethischen Anforderungen

Tatsächlich ist zumindest die Transparenz von Informationen in Verwaltungsverfahren und algorithmischen Ergebnissen im deutschen und europäischen Recht institutionalisiert. Um zu überprüfen, inwiefern die ethische Pflicht zur Transparenz bei der Bereitstellung von öffentlichen Informationen schon in juridischen Rechten verankert ist, soll hier ein kleiner Überblick über die rechtliche Lage gegeben werden. Artikel 5, Absatz 1 des deutschen Grundgesetzes besagt: „Jeder hat das Recht, (...) sich aus allgemein zugänglichen Quellen ungehindert zu unterrichten."[966] Das heißt, staatliche Organe, wie die Bundesregierung, müssen aktiv Informationsarbeit leisten

962 Munch et al. 2024:8
963 Ebd.:2
964 Munch et al. 2024:10
965 Mohr 2010:77
966 Art. 5, Abs. 1, Satz 1 GG

und solche allgemein zugänglichen Quellen zur Verfügung stellen, damit Staatsbürger:innen die Politik beurteilen und bei demokratischen Wahlen informierte Entscheidungen treffen können. Zudem besagt das Gesetz zur Regelung des Zugangs zu Informationen des Bundes (Informationsfreiheitsgesetz – IFG) Folgendes: „Jeder hat nach Maßgabe dieses Gesetzes gegenüber den Behörden des Bundes einen Anspruch auf Zugang zu amtlichen Informationen. Für sonstige Bundesorgane und -einrichtungen gilt dieses Gesetz, soweit sie öffentlich-rechtliche Verwaltungsaufgaben wahrnehmen."[967] Wo also Verwaltungsaufgaben übernommen werden, haben Staatsbürger:innen ein Recht auf Zugang zu amtlichen Informationen – damit ist gemeint: „(...) jede amtlichen Zwecken dienende Aufzeichnung (...)."[968] Es gibt jedoch auch Ausnahmen zu diesem Recht, wenn es sich bei den amtlichen Informationen zum Beispiel um personenbezogene Daten von Dritten handelt. Dann darf der Zugang nur bewilligt werden, wenn die Dritten eingewilligt haben oder das Informationsinteresse das schutzwürdige Interesse der Dritten überwiegt.[969] Noch konkreter auf die öffentliche Verwaltung in Deutschland bezogen gibt es Informationsrechte, die im Verwaltungsverfahrensgesetz (VwVfG) festgehalten sind: „Die Behörde hat den Beteiligten Einsicht in die das Verfahren betreffenden Akten zu gestatten, soweit deren Kenntnis zur Geltendmachung oder Verteidigung ihrer rechtlichen Interessen erforderlich ist. Satz 1 gilt bis zum Abschluss des Verwaltungsverfahrens nicht für Entwürfe zu Entscheidungen sowie die Arbeiten zu ihrer unmittelbaren Vorbereitung."[970] Hier wird deutlich, dass zumindest der Zugänglichkeits-Teil[971] der ethischen Pflicht zur Transparenz bei der Bereitstellung von öffentlichen Informationen schon in juridischen Rechten verankert ist.

Es fragt sich nun, inwiefern sich diese rechtlichen Regelungen übertragen lassen, wenn in Verwaltungsverfahren Algorithmen eingesetzt werden. Beim Einsatz von Algorithmen könnte der Anspruch auf Zugang zu jeder „(...) amtlichen Zwecken dienende[n] Aufzeich-

967 §1, Abs. 1, Satz 1 und 2 IFG
968 §2, Abs. 1, Satz 1 IFG
969 §5, Abs. 1, Satz 1 IFG
970 §29, Abs. 1, Satz 1 und 2 VwVfG
971 Mittelstadt et al. 2016:6

nung (...)“[972] eine neue Bedeutung bekommen. Denn während Verwaltungsmitarbeitende vermutlich aufzeichnen, wie es zu gewissen Entscheidungen kommt, erzeugen Algorithmen nur ein Ergebnis. Angenommen, der Wohngeldantrag einer Person wird abgelehnt, und sie fordert Einsicht in die das Verfahren betreffenden Akten, dann hätte sie nach Abschluss des Verwaltungsverfahrens auch Anspruch auf Einsicht in Entwürfe zu Entscheidungen sowie deren Vorbereitung. Hat aber ein Algorithmus die Entscheidung getroffen bzw. wurde die Entscheidung von Verwaltungsmitarbeitenden allein auf der Grundlage der algorithmischen Empfehlung getroffen, gibt diese Information nicht viel Aufschluss darüber, warum es zu dieser Entscheidung kam.

Das deutsche Informationsrecht im Kontext der Verwaltung bezieht sich noch auf von Menschen durchgeführte und dementsprechend anhand von Akten nachvollziehbare Verwaltungsprozesse. Solche Prozesse würden durch den Einsatz von Algorithmen jedoch stark verändert, da das Vorgehen nicht mehr dokumentiert und womöglich nicht mehr nachvollziehbar wäre. An dieser Stelle kann das europäische Recht weiterhelfen, welches sich der digitalen Datenverarbeitung und ihren rechtlichen Umständen in den letzten Jahren ausführlich gewidmet hat. So enthält zum Beispiel die Datenschutzgrundverordnung (DSGVO) eine Informationspflicht bei der Erhebung von personenbezogenen Daten. Das heißt, um eine faire und transparente Verarbeitung zu gewährleisten, müssen bei einer automatisierten Entscheidungsfindung, von der Personen betroffen sind, unter anderem „(...) aussagekräftige Informationen über die involvierte Logik (...)“[973] bereitgestellt werden. Die involvierte Logik algorithmischer Entscheidungsfindung könnte durch Informationen über den Zusammenhang zwischen Input und Output dargestellt werden. Hier stimmen meine ethischen Forderungen mit den rechtlichen Vorschriften der DSGVO überein.

Außerdem haben laut DSGVO alle Personen das Recht, nicht einer Entscheidung unterworfen zu werden, die ausschließlich auf einer automatisierten Verarbeitung beruht.[974] Dies gilt allerdings nur dann, wenn die Entscheidung der Person gegenüber rechtliche Wir-

---

972 §2, Abs. 1, Satz 1 IFG
973 Art. 13, Absatz 2 f) und Artikel 14, Absatz 2 g) DSGVO
974 Art. 22, Abs. 1 DSGVO

kung entfaltet oder sie anderweitig erheblich beeinträchtigt, wie zum Beispiel eine „(...) automatische Ablehnung eines Online-Kreditantrags oder Online-Einstellungsverfahren ohne jegliches menschliche Eingreifen."[975] Bei der Verwendung eines hochkomplexen Algorithmus kann man nie ganz sicher sein, ob der Glaube an den Output gerechtfertigt ist. Deshalb schlägt Burrell vor, die Verwendung von Algorithmen in bestimmten kritischen Anwendungsbereichen zu vermeiden.[976] Der AI Act hingegen schließt nicht gesamte Anwendungsbereiche aus, sondern er verbietet bestimmte Praktiken im KI-Bereich.[977]

Der AI Act schreibt vor, dass KI-Systeme nach Intensität und Umfang ihrer Risiken bewertet und eingeteilt werden. Damit der Einsatz sogenannter Hochrisiko-KI-Systeme erlaubt ist, werden hohe Anforderungen gestellt und Transparenzpflichten festgelegt.[978] In Artikel 6 und den Anhängen I und III wird definiert, was ein Hochrisiko-KI-System ausmacht. Das hohe Risiko bezieht sich hier vor allem auf kritische Anwendungsbereiche und direkte Auswirkungen auf natürliche Personen.[979] Bestimmte KI-Systeme werden als hochriskant eingestuft, „(...) um nachteilige Auswirkungen zu vermeiden, das Vertrauen der Öffentlichkeit zu erhalten und die Rechenschaftspflicht und einen wirksamen Rechtsschutz zu gewährleisten."[980] An dieser Stelle geht die Ethik über das Recht hinaus, da sie ihre Vorsichtsmaßnahmen nicht auf KI-Systeme in hochriskanten Kontexten beschränkt. Denn laut Munch et al. können auch algorithmische Entscheidungen mit geringem Risiko ein Recht auf Erklärung begründen, und zwar wenn sie Teil eines Musters von Entscheidungen sind, deren Ergebnis insgesamt gesehen für den Einzelnen von erheblicher Bedeutung sein kann.[981]

Transparenz bedeutet im AI Act, dass KI-Systeme „(...) angemessen nachvollziehbar und erklärbar (...)"[982] sind, und dass den Nutzenden bewusst gemacht wird, dass sie mit einem KI-System

975 Erwägungsgrund 71 DSGVO
976 Burrell 2016:9
977 Art. 5 EU AI Act
978 Vorbemerkung Nr. 26 und 46 EU AI Act
979 Art. 6; Anhang I und III EU AI Act
980 Vorbemerkung Nr. 59 EU AI Act
981 Munch et al. 2024:2
982 Vorbemerkung Nr. 27 EU AI Act

interagieren.[983] Dafür müssen die Betreiber:innen die Nutzenden „(...) über die Fähigkeiten und Grenzen des KI-Systems informieren und die betroffenen Personen über ihre Rechte in Kenntnis setzen (...).“[984] Hochrisiko-KI-Systeme sollen transparent sein, damit die Ausgaben des Systems angemessen interpretiert und verwendet werden können.[985] Zum Erreichen der Transparenz sollen Hochrisiko-KI-Systeme mit Betriebsanleitungen ausgestattet werden, welche Informationen über ihre Merkmale, Fähigkeiten und Leistungsgrenzen enthalten. So soll deutlich werden, zu welchem Zweck das Hochrisiko-KI-System entwickelt wurde, aber auch, welche Risiken alle vorhersehbaren Fehlanwendungen bergen. Zudem soll die zu erwartende Genauigkeit des Hochrisiko-KI-Systems sowie seine Leistung in Bezug auf bestimmte Personen oder Personengruppen, auf die es angewandt werden soll, offenbart werden. Wenn sie zur Erläuterung der Ausgaben des Hochrisiko-KI-Systems relevant sind, müssen auch technische Merkmale und Informationen über die verwendeten Trainings-, Validierungs- und Testdatensätze bereitgestellt werden. Außerdem soll eine solche Betriebsanleitung Informationen beinhalten, „die es den Betreibern ermöglichen, die Ausgabe des Hochrisiko-KI-Systems zu interpretieren und es angemessen zu nutzen“.[986] Damit hierbei Geschäftsgeheimnisse gewahrt werden können, müssen die Informationen nicht technisch detailliert sein.[987]

Der in deutschen Gesetzen verankerte Zugang zu amtlichen Informationen sowie die Einsicht in die das Verfahren betreffenden Akten ermöglichen nur einen Teilaspekt der Transparenz: die Zugänglichkeit. Die Verständlichkeit der Informationen ist dadurch nicht garantiert.[988] Erst die europäischen Gesetze bzw. Verordnungen verpflichten zusätzlich dazu, die zur Verfügung gestellten Informationen oder algorithmischen Ergebnisse verständlich zu machen. Trotzdem leistet die Ethik hier einen wichtigen Beitrag, weil sie eine ganz genaue Analyse davon ermöglicht, welche Informationen in Bezug auf algorithmische Ergebnisse bereitgestellt werden sollten.

983 Art. 50, Abs. 1, Satz 1 EU AI Act
984 Vorbemerkung Nr. 27 EU AI Act
985 Art. 13, Abs. 1, Satz 1 EU AI Act
986 Art. 13, Abs. 3 (b) EU AI Act
987 Vorbemerkung Nr. 107 EU AI Act
988 Mittelstadt et al. 2016:6

Die rechtliche Pflicht, „(...) aussagekräftige Informationen über die involvierte Logik (...)“[989] von algorithmischer Entscheidungsfindung bereitzustellen, ist noch sehr unkonkret. Auch die Anforderungen an die Betriebsanleitung für Hochrisiko-KI-Systeme aus dem AI Act bleibt vage. Hier sollen die Informationen über die algorithmischen Ergebnisse ihre Interpretation und angemessene Nutzung ermöglichen.[990] Wann Informationen über die Logik eines Algorithmus aussagekräftig sind, beziehungsweise wann Informationen über algorithmische Ergebnisse deren Interpretation und angemessene Nutzung ermöglichen, wurde erst durch meine ethische Erläuterung spezifiziert. Mit Hilfe der Ethik ließ sich herausarbeiten, dass man wissen muss, welcher Input welchen Output erzeugt und welche Datensätze zugrunde liegen, um ein algorithmisches Ergebnis verstehen zu können[991], und dass die Ergebnisse im zum Verständnis notwendigen Kontext dargestellt werden sollten, um transparent zu sein. Wer dafür verantwortlich ist, diese Pflicht zu erfüllen, wird in Kapitel 8 erörtert.

7.4 Aus der Pflicht zur Transparenz abgeleitete Anforderungen

In diesem Kapitel ist klar geworden, dass Datenverarbeitende moralisch verpflichtet sind, Datennutzenden transparente Ergebnisse bereitzustellen, da sie ohne Transparenz nicht reflektiert Stellung nehmen können und somit als Mittel gebraucht werden. Die fehlende Transparenz kann auf Grund ihrer unerwünschten Folgen, von dem Bilden falscher Überzeugungen bis zur Unmöglichkeit die eigenen Interessen zu vertreten, als Risiko bezeichnet werden. Da dieses Risiko den Datennutzenden von den Datenverarbeitenden auferlegt wird, handelt es sich um eine Risikoauferlegung. Mit Hilfe der deontologischen Risikoethik lässt sich deshalb argumentieren, dass diese Risikoauferlegung nur im Rahmen eines gerechten sozialen Systems der Risikobereitschaft geschehen darf, welche allen Beteiligten zum

989 Art. 13, Absatz 2 f) DSGVO und Artikel 14, Absatz 2 g) DSGVO
990 Art. 13, Abs. 3 (b) EU AI Act
991 Humphreys 2009:618; Buijsman, Veluwenkamp 2023:545

Vorteil gereicht.[992] Das bedeutet, durch die Regel der dritten Formel des Kategorischen Imperativs von Kant, dass niemand als bloßes Mittel gebraucht werden darf.[993] Das heißt in der Praxis, das die individuelle Kontrolle der Datennutzenden über das Risiko sollte so groß wie möglich sein. Dafür sollten die Risikoexponierten ohne Täuschung über das Risiko informiert werden und ohne Zwang die Wahl haben, ob sie das Risiko tragen wollen. Auf diese Weise kann die Transparenz gefördert werden.

Aus den theoretischen Überlegungen zur Transparenz bezüglich algorithmischer Ergebnisse lassen sich Anforderungen an die tatsächliche Umsetzung von Datenverarbeitung ableiten. Definieren wir Transparenz, wie bei Mittelstadt et al., als Zugänglichkeit und Verständlichkeit[994], so müssen wir die Bedeutung algorithmischer Ergebnisse zugänglich und verständlich machen, indem wir ihre Entstehung betrachten – nur durch die Erklärung der Entstehung des algorithmischen Ergebnisses sind wir in der Lage, seine Bedeutung zu verstehen. Um den Datennutzenden das Risiko der fehlenden Transparenz im Rahmen eines gerechten sozialen Systems der Risikobereitschaft aufzulegen, sollten Datenverarbeitende ihrer Pflicht zur Transparenz nachkommen, indem sie die algorithmischen Ergebnisse im zum Verständnis notwendigen Kontext darstellen. Nur so können die Datennutzenden nachvollziehen, wie das Ergebnis zu verstehen ist. Um diese Einordnung für die Datennutzenden vornehmen zu können, müssen die Datenverarbeitenden dazu in der Lage sein nachzuvollziehen, wie das Ergebnis zustande gekommen ist bzw. welcher Input zu welchem Output führt und welche Trainingsdaten verwendet wurden (siehe Tabelle 4).

Diese administrativen Anforderungen sorgen dafür, dass die Datennutzenden in Bezug auf den Umgang mit den Daten, denen sie im Datenökosystem ausgeliefert sind, selbstbestimmt handeln können und sie somit nicht als bloßes Mittel gebraucht werden. Da die Auferlegung eines Risikos nur dann zulässig ist, wenn niemand als bloßes Mittel gebraucht wird, ist die Erfüllung der administrativen Anforderungen eine notwendige Bedingung für die Zulässigkeit der Risikoauferlegung.

992 Hansson 2003:305
993 Kant 2008:65 [429]
994 Mittelstadt et al. 2016:6

Tabelle 4 Administrative Anforderungen für die Umsetzung von Transparenz

Transparenz Definition	**Transparenz Ziele**	**Umsetzung des Ziels**	**Administrative Anforderungen**
Transparenz = Zugänglichkeit und Verständlichkeit[995]	Algorithmische Ergebnisse sollen *verständlich* sein.	Nur durch die *Erklärung der Entstehung* des algorithmischen Ergebnisses sind wir in der Lage, seine Bedeutung zu verstehen.	*Entstehung des Ergebnisses verstehen:* Welcher Input führt zu welchem Output, und welche Trainingsdaten wurden verwendet?
	Algorithmische Ergebnisse sollen *zugänglich* sein.	Algorithmische Ergebnisse werden den Datennutzenden zur Verfügung gestellt, *samt leicht verständlicher Erklärung.*	Algorithmische Ergebnisse im *zum Verständnis notwendigen Kontext* darstellen.

Aus der Pflicht zur Transparenz der algorithmischen Ergebnisse und den zur Umsetzung notwendigen administrativen Anforderungen lässt sich ein klarer Vorteil erkennen für den Einsatz von Wenn-Dann-Algorithmen. Durch die Klarheit der vorgegebenen Entscheidungsregeln sind die Ergebnisse transparent und können problemlos zugänglich gemacht und verständlich dargestellt werden.

Es lassen sich aber nicht alle Sachverhalte auf klare Regeln herunterbrechen. Oft gibt es einen Ermessensspielraum, sodass für eine Eingabe mehrere Ausgaben möglich sind. Zum Beispiel hat eine Behörde im Rahmen von vorgegebenen Bauvorschriften und Umweltstandards den Ermessensspielraum, eine Baugenehmigung zu erteilen oder nicht. Oder wenn ein Strafmaß nicht eindeutig festgelegt ist, hat ein Richter, innerhalb des Gesetzesrahmens, den Ermessensspielraum, die Höhe des Strafmaßes zu bestimmen. Eine andere Situation, in der die Ausgabe nicht vorgegeben werden kann, ist die Suche nach von Menschen unentdeckten Mustern in Datensätzen. Eine solche Mustersuche kann zum Beispiel angewendet werden bei der Einstellung von Personal. So können zum Beispiel, wie bei

995 Mittelstadt et al. 2016:6

Amazon im Jahr 2014, Muster identifiziert werden in Bewerbungen, die in der Vergangenheit erfolgreich waren. Auf diese Weise werden Merkmale des Erfolgs abgeleitet und zu Entscheidungskriterien für neue Bewerbungen gemacht. Nun können aus vorliegenden Bewerbungen die Besten herausgefiltert werden, ohne vorzugeben was eine Bewerbung gut macht.[996]

Sobald es keine klare Wenn-Dann Beziehung gibt, sondern ein Ermessensspielraum gilt oder eine Mustersuche das Ziel ist, kann ein regulärer Wenn-Dann-Algorithmus nicht mehr eingesetzt werden. Um in einem solchen Fall trotzdem ein Ergebnis zu erhalten, ist der Einsatz von KI-Systemen möglich. Expert:innen wie Datenverarbeitende kennen die unter 7.2 beschriebenen Praktiken, mit deren Hilfe sie die Glaubwürdigkeit von Algorithmen überprüfen können. Datennutzende, welche die Ergebnisse des Modells nutzen oder von ihnen betroffen sind, sind hierbei jedoch nicht einbezogen, sodass es für die breite Öffentlichkeit schwer ist, Algorithmen zu überprüfen. Dadurch bleiben epistemische Asymmetrien zwischen Datenverarbeitenden und Datennutzenden meist bestehen.[997]

Das ist deshalb problematisch, weil Datennutzende epistemisch abhängig sind von KI-Systemen, wenn sie sich auf ihre Ergebnisse verlassen (müssen). Auch von anderen Technologien, wie zum Beispiel Thermometern, sind wir epistemisch abhängig. Aber die Ergebnisse von Thermometern sind deskriptiv und erfordern offensichtlich Kontextualisierung. Algorithmische Ergebnisse hingegen fordern oft direkt zum Handeln auf – zum Ablehnen eines Wohngeldantrages oder zum Verfolgen eines bestimmten Berufsweges. Jedoch treffen in der Regel nicht die Algorithmen die endgültige Entscheidung, da Menschen beteiligt sind, die ihre Entscheidungen auf der Grundlage der algorithmischen Ergebnisse treffen. Die menschlichen Entscheidungstragenden sind laut Grote et al. nicht nur passive Empfänger von Informationen, sondern Expert:innen in eigener Sache.[998] Im Kontext eines Datenökosystems läge die endgültige Entscheidung über die Ablehnung eines Wohngeldantrages demnach bei den Mitarbeitenden der Wohngeldbehörde.

996 Stahl et al. 2023:10f.
997 Kolkman 2022:104
998 Grote et al. 2024:5

Allerdings ist es, wie unter 7.1 erläutert, laut van den Hoven oft nicht möglich, anders zu entscheiden, als es der Algorithmus empfiehlt: „Users may become epistemically dependent upon the system they are working with, they cannot but cast their accounts of what they did and why they did it wholly in terms of the system. They may me unable to put forward 'system independent reasons'."[999] Aktuell sind Verwaltungsmitarbeitende Expert:innen für die Frage, wann ein Wohngeldantrag abgelehnt wird. Sie kennen die Gesetzeslage und arbeiten sich durch die Antragsunterlagen. Doch wenn ein Algorithmus eingesetzt wird, um diese Arbeit zu erleichtern, dann ist es nicht mehr notwendig, dass die Mitarbeitenden die Antragsunterlagen im Detail betrachten – es kann nicht einmal gewollt sein, denn sonst wäre der Einsatz des Algorithmus ja nicht notwendig oder sinnvoll. Dadurch entsteht zusätzlich die Gefahr des Verlustes dieser Expert:innenkenntnisse (deskilling), da sie nicht mehr regelmäßig gebraucht werden.[1000] So kann es über die Zeit sogar zur Herausforderung werden, die algorithmischen Ergebnisse im Zweifel händisch zu überprüfen.

Beim Einsatz von KI-Systemen muss also bedacht werden, dass es zusätzliche Arbeit erfordert, um die Pflicht zur Transparenz erfüllen zu können. Deshalb empfiehlt es sich, für alle Anwendungsfelder ohne Entscheidungsspielraum ausschließlich Wenn-Dann-Algorithmen zu verwenden. So kann zum Beispiel die Ablehnung eines Wohngeldantrages mit der Gesetzeslage begründet werden, welche dem Algorithmus als Entscheidungskriterium vorgegeben wurde. KI-Systeme hingegen sollten nur dann eingesetzt werden, wenn ihr Vorgehen bei der Beantwortung der Frage hilfreich ist, wenn der zugrundeliegende Datensatz wirklich zu hilfreichen Aussagen führen kann und wenn sich im Nachhinein erklären lässt, welcher Input zu welchem Output geführt hat.

999 Van den Hoven 1998:104
1000 Rafner et al. 2021