

Datenqualität bei algorithmischen Entscheidungen

Jeremy Stevens¹

Abstract: Die Kontrolle der Ergebnisse algorithmischer Entscheidungen auf Basis von Machine Learning-Verfahren stellt das Recht vor eine Vielzahl von Herausforderungen. Entsprechende Algorithmen analysieren umfangreiche Datenbestände und nutzen die in den Daten enthaltenen Muster und Gesetzmäßigkeiten zur Prognose. Die Qualität der verwendeten Daten hat daher einen erheblichen Einfluss auf das Ergebnis algorithmischer Entscheidungen. Der Beitrag soll aufzeigen, dass bestehende datenschutzrechtliche Konzepte zur Überprüfung der Qualität der verwendeten Daten keine hinreichende Kontrolle gewährleisten. Weder das System individueller Betroffenenrechte noch die Befugnisse der Datenschutzbehörden sind hierfür ausreichend. Dementsprechend sollen zuletzt Regulierungsmöglichkeiten aufgezeigt werden.

Keywords: Datenqualität; Algorithmische Entscheidungen; DSGVO

1 Einleitung

Künstliche Intelligenz und deren Anwendung im Rahmen automatisierter Entscheidungen stellen das Recht vor eine Vielzahl von Herausforderungen, insbesondere vor die Frage, inwiefern eine Kontrolle algorithmischer Entscheidungsergebnisse gewährleistet werden kann. Die fortschreitende technische Entwicklung macht dabei zunehmend komplexe Anwendungen und eine Automatisierung weiter Lebensbereiche möglich. Das Versprechen einer objektiven und präzisen Bearbeitung von Massenverfahren macht algorithmische Entscheidungsverfahren für Wirtschaft und Verwaltung interessant.² Berichte über bestehende Anwendungen bieten jedoch einige problematische Beispiele, in denen der Einsatz von Algorithmen zu falschen oder diskriminierenden Ergebnissen geführt hat. So besteht berechtigter Anlass zur Sorge, dass der im US-Justizsystem eingesetzte COMPAS-Algorithmus zur Einschätzung der Rückfallwahrscheinlichkeit verurteilter Straftäter erheblich zulasten afroamerikanischer Personen urteilt.³ Ähnliche Bedenken bestehen im Rahmen des polizeilichen Einsatzes von Algorithmen zur Prognose

¹ Goethe-Universität Frankfurt, Lehrstuhl für Öffentliches Recht, Informationsrecht, Umweltrecht, Verwaltungswissenschaft, Theodor-W.-Adorno-Platz 4, 60323 Frankfurt am Main, stevens@jur.uni-frankfurt.de

² Entsprechende Überlegungen finden sich beispielsweise in der KI-Strategie der Bundesregierung.

³ Siehe hierzu die Recherche von ProPublica, verfügbar unter <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>; <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>

von Straftaten.⁴ Mehrere Projekte zum Einsatz von IBM Watson zur Behandlung von Krebspatienten wurden wieder eingestellt, nachdem der Algorithmus häufig fehlerhafte Diagnosen und Behandlungsempfehlungen erstellt hatte.⁵ Ziel einer rechtlichen Kontrolle algorithmischer Entscheidungen muss sein, solche Fehleinschätzungen zu vermeiden.

Moderne Algorithmen basieren auf Machine Learning-Methoden. Diese erstellen auf Grundlage einer Analyse umfangreicher Datenbestände eigenständige Entscheidungsmodelle. Eine effektive Regulierung algorithmischer Entscheidungsprozesse setzt ein Verständnis dieser Methoden voraus. Der Beitrag beginnt daher mit einer kurzen Darstellung der Grundlagen des Machine Learning mit einem besonderen Fokus auf die Bedeutung der verwendeten Daten. Im Anschluss werden bestehende Datenqualitätsstandards im Datenschutzrecht analysiert und Probleme mit deren Anwendung auf algorithmischen Entscheidungssystemen aufgezeigt. Zuletzt sollen alternative Regulierungsmöglichkeiten skizziert werden.

2 Grundlagen des Machine Learning und Bedeutung der Datenqualität

Der Begriff des Machine Learning umfasst verschiedene statistische Methoden, durch die Algorithmen ohne umfassende menschliche Programmierung Lösungsmodelle für spezifische Aufgaben entwickeln können. Machine Learning Algorithmen nutzen umfangreiche Datenbestände als Input, um mittels statistischer Analyse Muster und Gesetzmäßigkeiten in den Daten zu erkennen und auf dieser Grundlage einen entsprechenden Output zu prognostizieren. Die Funktionsweise der einzelnen Methoden weicht zum Teil erheblich voneinander ab. Grundlegende Strukturen des Entwicklungsprozesses lassen sich aber für die rechtliche Diskussion verallgemeinern.

Ausgangspunkt der Entwicklung eines Machine Learning Algorithmus sind zumeist umfangreiche Datenbestände. Hierbei handelt es sich in der Regel um Big Data, daher große Mengen von unstrukturierten und semi-strukturierten Daten.⁶ Diese Daten können unmittelbar bei den Betroffenen erhoben werden, auf den Angaben und Einschätzungen von Mitarbeitern und Einsatzkräften, wie medizinischem Personal und Polizeibeamten, beruhen oder aus bestehenden Datenbanken bezogen werden.

Algorithmen prüfen diese Datenbestände im Rahmen einer statistischen Analyse auf Korrelationen zwischen einzelnen Inputvariablen und dem zu prognostizierenden Output. Die hierbei gefundenen Muster dienen als Grundlage des Prognosemodells. Aufgrund der häufig hohen Anzahl an Variablen und der Funktionsweise vieler Methoden ist

⁴Siehe hierzu Ki17.

⁵<https://www.faz.net/aktuell/wirtschaft/kuenstliche-intelligenz/computer-watson-scheitert-zu-oft-bei-datenanalyse-15619989.html>

⁶ Daneben beschreibt der Begriff auch Technologien, welche zur Verarbeitung der umfangreichen Datenmengen eingesetzt werden.

die Herleitung der Prognose für Menschen häufig gar nicht oder nur mit erheblichem Aufwand nachvollziehbar. Aus diesem Grund wird im Zusammenhang mit Machine Learning häufig von einer Black Box gesprochen.⁷ Die fehlende Interpretierbarkeit kann durch aktuelle Methoden zur Erklärbarkeit algorithmischer Ergebnisse auch nicht hinreichend ausgeglichen werden.⁸

Nach Abschluss der Entwicklung soll der Algorithmus im Realeinsatz zutreffende Prognosen erstellen können. Eine fehlerhafte Datengrundlage oder die Anwendung in einem abweichenden Kontext können zu fehlerhaften Entscheidungen führen, ohne dass dies für Betroffene erkennbar wäre. Ein Bericht der Investigativplattform ProPublica legt eine solche Situation im Fall des COMPAS-Algorithmus nahe.⁹ Die Veröffentlichung hat eine umfassende Diskussion über den Einsatz und die Beurteilung von Prognosesoftware im Justizsystem ausgelöst.¹⁰ Die fehlerhaften Prognosen von IBM Watson lassen sich wiederum zum Teil auch auf eine Diskrepanz zwischen den in der Entwicklung genutzten Datenbeständen und der Anwendung in den einzelnen Krankenhäusern zurückführen. Die Entwicklung erfolgte in Kooperation mit einem US-Krankenhaus mit überwiegend wohlhabenden Patienten. Muster in den Krankheitsbildern bilden diese spezifischen Umstände ab und sind als Grundlage einer Prognose lediglich für diese spezielle Situation geeignet. Eine Übertragung des Algorithmus auf andere Krankenhäuser führt zu fehlerhaften Ergebnissen.¹¹

Die verwendeten Daten haben einen erheblichen Einfluss auf algorithmische Entscheidungen. Die Gewährleistung einer qualitativen Datengrundlage stellt einen wichtigen Bestandteil der Kontrolle algorithmischer Entscheidungen dar. Wie die Beispiele zeigen, handelt es sich bei der Beurteilung der Datenqualität jedoch um eine komplexe Aufgabe, bei der unter anderem der Anwendungskontext, mögliche Fehler oder Verzerrungen bei der Datenerhebung oder strukturelle Benachteiligung einzelner Bevölkerungsgruppen relevant sein können.

Fraglich ist vor diesem Hintergrund, wie das Recht mit der Problematik der Datenqualität umgehen soll.

3 Rechtliche Anforderungen an die Datenqualität

Im Folgenden sollen kurz bestehende rechtliche Anforderungen an die Datenqualität, insbesondere im Zusammenhang mit algorithmischen Entscheidungen, untersucht wer-

⁷ So beispielsweise Sh18, Kr17, Wi18, Pa15.

⁸ Entsprechende Methoden, wie LIME (Local Interpretable Model-Agnostic Explanations, siehe hierzu RSG16) bieten zwar Ansätze zur Erklärbarkeit, sind aber nicht in jeder Situation wirksam einsetzbar.

⁹ Eine abschließende Beurteilung ist bei kommerziellen Algorithmen häufig schwierig, da die Anbieter deren Funktionsweise als Geschäftsgeheimnis behandeln.

¹⁰ Zum Bericht von ProPublica siehe Fn. 3; eine Übersicht über die Gesamtdebatte bietet Wa.

¹¹ Siehe Fn. 5

den. Als Untersuchungsgegenstand bietet sich insoweit die Datenschutzgrundverordnung an.¹² Diese befasst sich zum einen mit der Regulierung algorithmischer Entscheidungen¹³ und enthält zum anderen Vorgaben zur Datenqualität, die bisher noch nicht so viel Beachtung gefunden haben.¹⁴ Daneben soll der Umgang mit Datenqualität im US-Recht beleuchtet und der rechtlichen Sichtweise ein Überblick über technische Konzeptionen der Datenqualität entgegengestellt werden.

3.1. Anforderungen der Datenschutzgrundverordnung

Vorrangiger Anknüpfungspunkt im Rahmen der Datenschutzgrundverordnung ist Art. 5 Abs. 1 lit. d DSGVO, der die sachliche Richtigkeit der Daten zu einem Grundsatz der Datenverarbeitung erhebt¹⁵. Daneben könnten auch die Anforderungen an den Einsatz algorithmischer Entscheidungssysteme nach Art. 13 ff. und 22 DSGVO möglicherweise Auswirkungen auf Datenqualitätsstandards haben.

Art. 5 Abs. 1 lit. d DSGVO schreibt vor, dass personenbezogenen Daten „sachlich richtig und erforderlichenfalls auf dem neuesten Stand sein“ müssen und darüber hinaus „alle angemessenen Maßnahmen zu treffen [sind], damit personenbezogene Daten, die im Hinblick auf die Zwecke ihrer Verarbeitung unrichtig sind, unverzüglich gelöscht oder berichtigt werden („Richtigkeit“)“.

Die Auslegung des Begriffs der „Richtigkeit“ im Sinne von Art. 5 Abs. 1 lit. d DSGVO ist noch nicht eindeutig geklärt. Basierend auf dem deutschen Wortlautverständnis wird die Anwendung zum Teil auf Fakten im Sinne einer klaren Trennung zwischen „richtig“ und „unrichtig“ beschränkt.¹⁶ Zwar kann hiernach auch eine unvollständige Darstellung der Wirklichkeit zur Unrichtigkeit der Daten führen. Insgesamt beschränkt sich dieser Ansatz jedoch auf eindeutig nachweisbare Daten. Ein Kontextbezug besteht insoweit lediglich hinsichtlich der Aktualität der Daten. Persönliche Daten sind grundsätzlich auf dem neusten Stand zu halten, es sei denn, sie beziehen sich explizit auf einen früheren Zeitpunkt.¹⁷

Die Gegenansicht beruft sich auf die komplexere Bedeutung des „accuracy“-Begriffs der englischen Sprachfassung¹⁸ und befürwortet eine Ausweitung des Begriffs der Richtig-

¹² Wie Ho16a in Bezug auf Big Data zutreffend hinweist, betrifft die Frage der Datenqualität nicht nur personenbezogene Daten, sondern ebenso anonyme oder sachbezogene Daten.

¹³ Dies betrifft insbesondere Art. 22 DSGVO, welcher „automatisierte Entscheidungen im Einzelfall“ regeln soll.

¹⁴ Siehe insoweit Ho16a, Ho16b.

¹⁵ Dieser ist gemäß Art. 83 Abs. 5 lit. a DSGVO mit einer erhebliche Bußgeldandrohung verbunden, weswegen die Norm auch in der Praxis von hoher Bedeutung sein wird.

¹⁶ So Ro19.

¹⁷ So Ro19.

¹⁸ Der französische Wortlaut „exactitude“ ermöglicht eine vergleichbare Auslegung des Begriffs.

keit, im Sinne von Zielgerichtetheit oder Genauigkeit, auf Werturteile und Prognosen.¹⁹ Eine unrichtige Prognose wäre nach dieser Ansicht beispielsweise bei Zugrundelegung falscher Tatsachen gegeben.²⁰

Die Diskussion um die sachliche Richtigkeit von Daten wurde bisher vorwiegend im Rahmen der Betroffenenrechte nach Art. 12 ff. DSGVO geführt.²¹ Bei der Beurteilung des Anspruchs auf Berichtigung unrichtiger Daten nach Art. 16 DSGVO und vergleichbaren Sachverhalten²² sind insbesondere durch den Betroffenen beweisbare Tatsachen relevant. Bei der Kontrolle nichtrechtlicher Wertungen halten sich Gerichte mangels entsprechender Fachkompetenz in der Regel zurück.²³ Angesichts einfacher Sachverhalte ist eine enge Konzeption der Datenqualität möglicherweise ausreichend. In Bezug auf Big Data und algorithmische Entscheidungen sind einem solchen Ansatz jedoch Grenzen gesetzt. Big Data ist per Definition zu umfangreich und zu komplex, um diese mit manuellen Methoden zu bewältigen. Eine Kontrolle der Datenqualität muss daher gegebenenfalls an anderen Kriterien anknüpfen können. Dies gilt insbesondere bei Daten, deren Richtigkeit nicht ohne Weiteres ersichtlich ist sowie bei kontextbezogenen Daten als Grundlage einer algorithmischen Entscheidung.

Ein Beispiel hierfür sind Daten zur Social Media-Nutzung oder dem Einkaufsverhalten bestimmter Personen. Die einzelnen Likes bzw. Einkaufsartikel tragen für den Einzelnen häufig keine große Bedeutung. Der Betroffene wird daher häufig nicht in der Lage sein, die Daten vollständig zu verifizieren. Das gleiche gilt für auf dieser Grundlage getroffenen Prognosen über zukünftiges Verhalten. Die Probleme bei der Nutzung der medizinischen Prognosen des IBM Watson zeigen deutlich, dass Daten in einer spezifischen Situation inhaltlich zutreffend erhoben worden sein und trotzdem in einer anderen Situation zu unbrauchbaren Ergebnissen führen können.

Ein differenzierterer rechtlicher Datenqualitätsstandard im Sinne des „accuracy“-Ansatzes wäre eher geeignet, technische Standards zu integrieren und einen nutzbaren Maßstab für die Kontrolle von Big Data Anwendungen zu liefern.²⁴ Mangels einer eindeutigen Begriffsbedeutung verbleibt bis zur konkreten Klärung eine gewisse Rechtsunsicherheit bei der Anwendung des Grundsatzes.²⁵

Weitere Anforderungen zur Datenqualität könnten gegebenenfalls aus den spezifischen Anforderungen an den Einsatz automatisierter Verfahren nach Art. 13 Abs. 2 lit. f, Art. 14 Abs. 2 lit. g, Art. 15 Abs. 1 lit. h DSGVO und Art. 22 DSGVO folgen. Demnach sind

¹⁹ Ho16a; Scb Eine weite Anwendung liegt auch den Überlegungen der Art. 29 Datenschutzgruppe zugrunde, auch wenn dies nicht explizit benannt wird, siehe Ar18.

²⁰ Scb.

²¹ Vergleiche die Diskussion in diesem Rahmen mit teils anderen Positionen: Ho16a, Wo.

²² Vergleichbare Entscheidungen über die Richtigkeit von Aussagen sind beispielsweise in Unterlassungsklageverfahren zu treffen. In diesem Kontext schränkt die Meinungsfreiheit die Kontrolle wertender Äußerungen ein.

²³ Ein Beispiel hierfür bietet die umfassende Rechtsprechung zur Beurteilung von Prüfungsleistungen.

²⁴ So auch Ho16a.

²⁵ Deswegen für eine nachsichtige Anwendung plädierend Ho16a.

Betroffenen aussagekräftige Informationen über die involvierte Logik des automatisierten Verfahrens zur Verfügung zu stellen. Dies wird durch Erwägungsgrund 71 ergänzt, der die Verwendung geeigneter mathematischer oder statistischer Verfahren sowie technische und organisatorische Maßnahmen zur Sicherstellung richtiger Daten und Fehlervermeidung vorsieht. Der Begriff der „involvierten Logik“ sowie die in Erwägungsgrund 71 angesprochenen Vorkehrung könnten bei entsprechender Auslegung zur Bestimmung notwendiger Datenqualitätsstandards herangezogen werden. Soweit ersichtlich werden aber bisher keine entsprechenden Konzepte in diesem Rahmen diskutiert. Vielmehr beschränkt die Literatur die Information zur involvierten Logik auf die Darstellung der entscheidungsrelevanten Kriterien.²⁶

3.2. Datenqualität im US-Recht

Die Datenqualität ist auch Gegenstand des US-amerikanischen Rechts. Interessant sind insoweit insbesondere der Privacy Act 1974, der Data Quality Act 2001 und die hierzu ergangenen Richtlinien sowie der Fall *Loomis v. Wisconsin*.

§ 552a (e) (5) des Privacy Act 1974 verpflichtet staatliche Behörden bei Entscheidungen gegenüber Einzelpersonen „accuracy, relevance, timeliness, and completeness“ der zugrunde gelegten Daten sicherzustellen. Anstelle eines allgemeinen Begriffs werden direkt mehrere Qualitätsmerkmale formuliert. Die Merkmale der Genauigkeit, Relevanz, Aktualität und Vollständigkeit der Daten werden in der Datenschutzgrundverordnung zwar ebenfalls erfasst. Die Formulierung des Privacy Act ist jedoch klarer gefasst und erlaubt eine differenzierte Anwendung einzelner Qualitätsmerkmale. Der Begriff der „accuracy“ im Kontext des Privacy Act entspricht einem weiten Verständnis der Richtigkeit gemäß Art. 5 Abs. 1 lit. d DSGVO im Sinne von Genauigkeit und Zielgerichtetheit.²⁷ Anders als die Datenschutzgrundverordnung ist der Privacy Act jedoch auf die Datenverarbeitung von staatlichen Behörden beschränkt. Für andere Rechtsbereiche existieren vergleichbare Spezialgesetze.²⁸

Im Rahmen des Data Quality Act²⁹ wurde das Office of Management and Budget zum Erlass v on Richtlinien ermächtigt, um „quality, objectivity, utility, and integrity of information“ zu gewährleisten. Die auf dieser Grundlage erlassenen Vorgaben zur Daten- bzw. Informationsqualität sind vorwiegend verfahrensbezogen, wie beispielsweise die Gewährleistung der Objektivität durch Peer-Review-Verfahren.³⁰ Auch wenn diese Standards vorwiegend mit Blick auf die Veröffentlichung von Studien geschaffen wurden, lassen sich hieraus Schlussfolgerungen für rechtliche Anforderungen an die Daten-

²⁶ So beispielsweise *Sca, Di19* und die Überlegungen der Art. 29 Datenschutzgruppe, siehe Ar18.

²⁷ Siehe hierzu beispielsweise *Doe v. U.S.*, 821 F.2d 694 (C.A.D.C., 1987)

²⁸ Weitere Beispiele und Verweise bei Ho16a.

²⁹ Der Data Quality Act bzw. Information Quality Act wurde als sogenannter „rider“ als Section 515 of the Consolidated Appropriations Act, 2001 durch den Kongress beschlossen und hat daher keinen offiziellen Namen.

³⁰ Zu einer Analyse dieses Ansatzes im Rahmen des Data Quality Acts, siehe Ga.

qualität algorithmischer Entscheidungen ziehen. Der Data Quality Act zeigt insbesondere auf, dass eine Regulierung der Daten- oder Informationsqualität sich auch auf deren Inhalt auswirkt.³¹ Denn verschiedene Lobbyorganisationen haben versucht mit dem Vorwurf einer fehlenden Informationsqualität kritische Studien zu unterdrücken, beispielsweise zur gesundheitsschädlichen Wirkung übermäßigen Salzkonsums.³²

Der Supreme Court of Wisconsin hatte im Verfahren *Loomis v. Wisconsin* über die Zulässigkeit der Verwendung des COMPAS-Algorithmus zu entscheiden.³³ Die Nutzung der Prognosesoftware wurde für zulässig befunden unter Berufung auf die Möglichkeit des Beschwerdeführers unrichtige Daten korrigieren zu können. Die der engen Sichtweise zu Art. 5 Abs. 1 lit. d DSGVO vergleichbare Ansicht des Gerichts wurde in der Literatur als den Anforderungen des Problems nicht entsprechend kritisiert.³⁴

Das US-Recht bietet daher verschiedene Anknüpfungspunkte für eine Diskussion zur Datenqualität, die im Rahmen einer europäischen Debatte berücksichtigt werden können.

3.3. Technische Datenqualitätsstandards

Bei der Ausarbeitung rechtlicher Anforderungen an die Datenqualität kann auf umfassende technische Standards und entsprechende Forschungsarbeiten zurückgegriffen werden. Der Umfang und die Komplexität vorhandener Ansätze stellt die Rezeption durch die Rechtswissenschaften jedoch auch vor Schwierigkeiten.³⁵ Daher soll nur kurz ein Überblick zu technischen Konzeptionen der Datenqualität gegeben werden.

Die Datenqualität kann aus technischer Sicht beispielsweise als „fitness for use“ definiert werden.³⁶ Dieser nutzenorientierte Ansatz macht die Kontextabhängigkeit der Datenqualität deutlich. Eine andere Möglichkeit ist die Definition der Datenqualität als Maß der Übereinstimmung zwischen dem durch die Daten des Informationssystems präsentierten Bild und den tatsächlichen Daten.³⁷ Technische Standards sind stark ausdifferenziert mit einer Vielzahl von konkreteren Indikatoren, wie die Verlässlichkeit der Daten entsprechend der Kompetenz und Glaubwürdigkeit der datenerhebenden Stelle. Entsprechende Standards finden sich auch in technischen Normierungen, wie der ISO 8000 Norm. Aufgrund der fortschreitenden Entwicklung und der Vielzahl von Anwendungs-

³¹ Zu den weiteren Implikationen, siehe Ga.

³² *Salt Institute v. Leavitt*, 440 F.3d 156 (4th Cir. 2006), daneben haben auch andere Gruppen versucht den Data Quality Act für Deregulierungsinteressen zu instrumentalisieren: https://www.washingtonpost.com/archive/politics/2004/08/16/data-quality-law-is-nemesis-of-regulation/3c35ae56-7935-4694-b125-348a374b657b/?utm_term=.79ca035c7e08

³³ *Loomis*, 881 N.W.2d at 756

³⁴ Zu einer umfassenden Darstellung der US-amerikanischen Diskussion hierzu, siehe Wa.

³⁵ Für eine Übersicht zur Forschung zur Datenqualität bei Big Data, siehe beispielsweise MBP.

³⁶ Siehe beispielsweise TB98.

³⁷ Mit weiteren Nachweisen He.

bereichen stehen allgemeingültige Standards zur Kontrolle der Datenqualität derzeit noch aus.³⁸

Aufgrund der unterschiedlichen Zielsetzungen lassen sich technische Ansätze nicht ohne weiteres auf rechtliche Standards übertragen. Sie bieten jedoch einen Anknüpfungspunkt für eine fundierte und tiefergehende rechtliche Diskussion und eine Ausarbeitung differenzierter rechtlicher Standards.

4 Anwendung der Regulierung auf algorithmische Entscheidungssysteme

Fraglich ist, ob die dargestellten rechtlichen Ansätze zur Überprüfung der Datenqualität zur Regulierung algorithmischer Entscheidungssysteme geeignet sind. Dies setzt voraus, dass durch rechtliche Anforderungen mögliche Fehler vermieden werden können und dass die entsprechenden Normen praktisch anwendbar sind.

Die Rechtsdurchsetzung im Rahmen der Datenschutzgrundverordnung erfolgt zum einen durch Tätigwerden der Aufsichtsbehörden und zum anderen durch die Betroffenen selbst. Fraglich ist, inwiefern diese jeweils die Datenqualität algorithmischer Entscheidungssysteme überprüfen können.

4.1. Geltendmachung von Betroffenenrechten

Betroffene einer algorithmischen Entscheidung sind in der Regel auf Informationen des Verantwortlichen angewiesen, um deren Rechtmäßigkeit überprüfen und gegebenenfalls weitere Rechte geltend machen zu können. Angesichts der Schwierigkeiten bei der Erklärung der maßgeblichen Gründe einer algorithmischen Entscheidung³⁹ stellt sich die Frage, ob die Datenschutzgrundverordnung Betroffenen eine Überprüfung der Datenqualität ermöglicht.

Betroffene können gegenüber Verantwortlichen ein Auskunftsrecht nach Art. 15 Abs. 1 DSGVO geltend machen. Dies umfasst einerseits eine Auskunft über die sie betreffenden beim Verantwortlichen gespeicherten Daten und zum anderen nach Art. 15 Abs. 1 lit. h DSGVO aussagekräftige Informationen über die involvierte Logik der automatisierten Entscheidungsfindung.⁴⁰ Der Bedeutungsumfang des Begriffs der involvierten Logik ist noch nicht abschließend geklärt. Überwiegend wird hierunter eine Offenlegung der wesentlichen Entscheidungskriterien gefasst und das Verhältnis zum Ge-

³⁸ Siehe zu technischen Ansätzen und Standards: TSD18, CZ15.

³⁹ Vergleiche zu möglichen Erklärungsansätzen und deren Akzeptanz durch Betroffene: Bi18.

⁴⁰ Art. 13, 14 DSGVO normiert vergleichbare Informationspflichten des Verantwortlichen.

schäfts- und Betriebsgeheimnisschutz diskutiert.⁴¹ Die Datenqualität wird bisher nicht erfasst. Dementsprechend umfasst der Auskunftsanspruch nach Art. 15 Abs. 1 DSGVO keine Informationen zur Datenqualität.

Betroffenen fehlt damit die Möglichkeit die Datenqualität algorithmischer Entscheidungen zu überprüfen.

4.2. Kontrolle durch Aufsichtsbehörden

Die datenschutzrechtlichen Aufsichtsbehörden können grundsätzlich die Einhaltung der Grundsätze der Datenverarbeitung nach Art. 5 DSGVO überprüfen und im Zweifel nach Art. 83 DSGVO ein Bußgeld verhängen. Es besteht daher theoretisch die Möglichkeit auf Grundlage des Grundsatzes der sachlichen Richtigkeit von Daten nach Art. 5 Abs. 1 lit. d DSGVO die Datenqualität algorithmischer Entscheidungssysteme überprüfen.

Fraglich ist jedoch die praktische Umsetzbarkeit der Kontrolle der Datenqualität. Die Aufsichtsbehörden können umfangreiche Informationen von den Verantwortlichen einfordern. Eine umfassende Überprüfung der Qualität der verwendeten Datenbestände mittels mathematischer und statistischer Verfahren würde ihre personellen und fachlichen Kapazitäten jedoch überfordern.⁴² Es fehlt zum einen an der notwendigen Ausstattung der Behörden und zum anderen an dem für eine effektive Kontrolle notwendigen bereichsspezifischen Knowhow. Das entsprechende Wissen, beispielsweise zum Gesundheitssektor oder der Polizeiarbeit, findet sich bei den Fachbehörden.⁴³ Eine technische Überprüfung der Datenqualität durch die Datenschutzbehörden wäre daher nicht zweckmäßig.⁴⁴

In Betracht kommt daher lediglich eine Integration von Datenqualitätsstandards in den prozeduralen Ansatz der Datenschutzgrundverordnung. Maßgeblicher Anknüpfungspunkt ist hierbei die Rechenschaftspflicht gemäß Art. 5 Abs. 2 DSGVO. Der Verantwortliche ist verpflichtet, die Einhaltung datenschutzrechtlicher Anforderungen und damit auch der Grundsätze der Datenqualität im Sinne von Art. 5 Abs. 1 lit. d DSGVO nachzuweisen. Dies wird maßgeblich durch technische und organisatorische Maßnahmen, wie beispielsweise die Führung eines Verarbeitungsverzeichnisses nach Art. 30 DSGVO oder den Datenschutzbeauftragten gemäß Art. 37 DSGVO sichergestellt. Beim Einsatz algorithmischer Entscheidungen ist außerdem gemäß Art. 35 Abs. 3 lit. a DSGVO regelmäßig eine Datenschutzfolgeabschätzung nach Art. 35 DSGVO durchzuführen. Im Rahmen einer Folgenabschätzung sind mögliche Risiken der Datenverarbeitung und entsprechende Gegenmaßnahmen darzustellen. In diesem Rahmen könnte auch der Nachweis eines adäquaten Datenqualitätsmanagements als Ge-

⁴¹ Sca, Di19.

⁴² So Ho16a.

⁴³ Diese Überlegung gilt für den gesamten Bereich der Algorithmenregulierung, weswegen zum Teil eine Stärkung der sektorspezifischen Behörden gefordert wird.

⁴⁴ Auf den Konflikt zur eigentlichen Zwecksetzung der Datenschutzaufsicht hinweisend Fn. 42.

samtheit der technischen und organisatorischen Maßnahmen zur Gewährleistung der Datenqualität im Sinne von Erwägungsgrund 71 erforderlich sein. Eine Integration der Anforderungen an die Datenqualität in die Verfahrensvorgaben der Datenschutzgrundverordnung hätte den Vorteil, dass mögliche Verstöße im regulären System der Bußgeldtatbestände einfacher berücksichtigt werden können.

Für eine entsprechende Berücksichtigung von Datenqualitätsanforderungen sind verlässliche Maßstäbe erforderlich. Der Grundsatz der sachlichen Richtigkeit der Daten nach Art. 5 Abs. 1 lit. d DSGVO ist als Grundlage hierfür derzeit noch nicht hinreichend klar definiert. Eine allgemeine Pflicht zur Gewährleistung der Richtigkeit der Daten ist für die konkrete Umsetzung zu unbestimmt.⁴⁵ Zwar bieten das US-Recht sowie technische Standards Anknüpfungspunkte für eine Konkretisierung. Eine rechtssichere Festlegung von Datenqualitätsmerkmalen könnte jedoch noch einige Zeit auf sich warten lassen.

Die Datenschutzgrundverordnung bietet Ansätze zur Kontrolle algorithmischer Entscheidungen. Mangels hinreichender Begriffsklärung sind diese jedoch nicht geeignet, um die Qualität der Datengrundlage algorithmischer Entscheidungen zu gewährleisten.

5 Möglichkeiten zur Regulierung der Datenqualität

Nachdem festgestellt wurde, dass die bestehenden Möglichkeiten der Datenschutzgrundverordnung zur Kontrolle der Datenqualität algorithmischer Entscheidungen nicht ausreichen, stellt sich die Frage, wie eine Überprüfung der Datenqualität durch Betroffene und Aufsichtsbehörden anderweitig gewährleistet werden könnte.

Dies betrifft zum einen die rechtliche Definition des Begriffs der Datenqualität und zum anderen die konkrete Überprüfung. Wie bereits dargelegt, ist eine Auslegung des Art. 5 Abs. 1 lit. d DSGVO im Sinne einer klaren Trennung zwischen richtig und unrichtig für die Anwendung auf Big Data und algorithmische Entscheidungssysteme ungeeignet. Ein praktikabler Ansatz müsste sich stärker an technischen Konzeptionen von Datenqualität orientieren und verschiedene Einzelmerkmale integrieren. Dabei stellt sich die Frage, ob Art. 5 Abs. 1 lit. d DSGVO der richtige Anknüpfungspunkt ist oder ob das Wortlautverständnis durch eine entsprechende Auslegung überstrapaziert werden könnte. Alternativ könnte beispielsweise an Erwägungsgrund 71 angeknüpft werden, der sich explizit auf automatisierte Entscheidungen bezieht. Gleichzeitig würde die Etablierung divergierender Datenqualitätsstandards in der DSGVO aber nicht zur Rechtssicherheit beitragen. Daher empfiehlt sich vermutlich eine Auslegung des Art. 5 Abs. 1 lit. d DSGVO im Sinne des „accuracy“-Verständnisses.

Für die Durchsetzung mittels aufsichtsrechtlicher Kontrolle bietet sich wie bereits angedeutet eine stärkere Prozeduralisierung der Datenqualitätsanforderungen an. Dabei braucht es einerseits allgemeingültige Rahmenanforderungen sowie andererseits konkre-

⁴⁵ Siehe Fn. 25

te anwendungsspezifische Vorgaben. Insbesondere bei Letzteren wird eine Kooperation mit den zuständigen Fachbehörden notwendig sein.

Im Rahmen einer allgemeinen Richtlinie können generelle Grundsätze des Datenqualitätsmanagements rechtlich verbindlich festgeschrieben und grundsätzliche Maßstäbe für den Umfang der notwendigen Maßnahmen etabliert werden. Dabei sind insbesondere Risiken für die Betroffenen einer algorithmischen Entscheidung zu beachten. Als Grundlage solcher allgemeinen Anforderungen bietet sich die Konzeption der Datenqualität als Maß der Übereinstimmungen zwischen dem durch die Daten vermittelten Bild und der realen Situation an.⁴⁶ Dabei handelt es sich erstmal um eine bloße Zielvorgabe. Es bedürfte daher daneben einer Ausdifferenzierung durch verschiedene rechtliche Merkmale der Datenqualität. Hierbei kann zum Teil auf technische Standards oder auf Ansätze aus anderen Rechtsordnungen zurückgegriffen werden. In Betracht kommen Kategorien wie Objektivität, Vollständigkeit, Eignung für den spezifischen Kontext oder Verlässlichkeit bzw. Belastbarkeit der Daten.

Die Ausgestaltung sollte idealerweise derart erfolgen, dass der Bezug zur Gewährleistung der Datenqualität ohne weiteres nachvollziehbar ist und die Kontrolle einheitlich und zügig erfolgen kann. So könnte dem Risiko einer übermäßigen Anpassung des Algorithmus an lokale Besonderheiten durch eine Verpflichtung zum Test des Systems anhand der Daten verschiedener Standorte. So könnte beispielsweise im Gesundheitsbereich die Einbeziehung verschiedener Krankenhäuser vorgeschrieben werden. Bei einer solchen Regelung wäre zum einen der Effekt zugunsten der Datenqualität leicht ersichtlich und zum anderen die jeweiligen Testergebnisse einfach zu überprüfen. Bei der Ausgestaltung entsprechender Standards muss jedoch auch darauf geachtet werden, dass die Anforderungen an den Entwicklungsprozess nicht überfrachtet werden. Angesichts der fortschreitenden technischen Entwicklung sollten mögliche Standards flexibel bleiben, um falls notwendig angepasst werden zu können.

Für konkrete Anwendungsbereiche mit spezifischen Herausforderungen, wie beispielsweise den Gesundheitsbereich oder polizeiliche Prognosen, sind darüber hinaus explizite, den Gegebenheiten angepasste Vorgaben sinnvoll. Gegebenenfalls empfiehlt sich hier auch eine spezialgesetzliche Regelung, insbesondere beim staatlichen Einsatz algorithmischer Entscheidungssysteme wie bei der Polizei.

Entwickler und Verantwortliche algorithmischer Entscheidungssysteme müssten für den gesamten Entwicklungsprozess und während der Dauer des Einsatzes die Beachtung der Datenqualitätsanforderungen nachweisen können. Bereits als Teil der Entwicklung des Programms sind entsprechende Maßnahmen zu treffen. Der Verantwortliche ist darüber hinaus für die Lebensdauer eines algorithmischen Entscheidungssystems zur kontinuierlichen Überprüfung der Datenqualität verpflichtet. Dies wird nur in Zusammenarbeit mit dem Entwickler durchführbar sein und erfordert angemessene prozedurale Vorkehrun-

⁴⁶ Siehe hierzu Fn. 37

gen, wie beispielsweise die Implementierung von Monitoring-Schleifen.⁴⁷ Solange Unsicherheit hinsichtlich der algorithmischen Entscheidungsfindung besteht, ist es angezeigt die Entscheidungsgrundlage regelmäßig zu überprüfen und mit neuen Erkenntnissen und Entwicklungen abzugleichen.

Eine Kontrolle der Dokumentation eines entsprechenden Datenqualitätsmanagements könnte sowohl durch die Datenschutzbehörde als auch durch die jeweiligen Fachbehörden erfolgen. Jedenfalls im Rahmen einer vertieften Prüfung wird häufig eine Kooperation erforderlich sein, um auf die jeweiligen Besonderheiten angemessen eingehen zu können.

Fraglich ist, inwiefern Betroffenen eine Kontrolle der Datenqualität ermöglicht werden kann. Unabhängig von der konkreten Ausgestaltung werden hier praktische Grenzen gesetzt sein. Bereits die Erklärung des Ergebnisses gestaltet sich bei algorithmischen Entscheidungen schwierig.⁴⁸ Gleichzeitig können Überlegungen zur Vermittlung von Informationen über die Entscheidung an sich auch auf die Darstellung der Datenqualität übertragen werden. In vielen Fällen wird ein Laie nicht in der Lage sein, Fehler des Algorithmus erkennen zu können. Soweit im Rahmen der Erklärung beispielsweise Informationen über die Vergleichsdaten einbezogen werden, kann der Betroffene aber gegebenenfalls einschätzen, ob diese zu seiner Situation passen.⁴⁹ Es müsste dabei Sorge getragen werden, dass nicht lediglich die Illusion einer wirksamen Kontrolle entsteht. Grundlegende Informationen könnten jedoch Ausgangspunkt eines aufgeklärten Umgangs mit dem Problem der Datenqualität sein.

Ein solcher Informationsanspruch könnte beispielsweise durch eine weite Auslegung des Auskunftsanspruchs gemäß Art. 15 Abs. 1 lit. h DSGVO erfolgen. Als Teil der Informationen zur involvierten Logik des Entscheidungsverfahrens könnten Angaben zu den zugrundeliegenden Daten erforderlich sein. Informationen zur Datenqualität sind zum Verständnis der Entscheidungsfindung erforderlich. Eine Einbeziehung unter den Begriff der involvierten Logik wäre bei der begrifflichen Fassung der Datenbestände als Grundlage dieser Logik auch mit dem Wortlaut vereinbar.

Die bestehenden Ansätze der Datenschutzgrundverordnung können fortgeführt und weiterentwickelt werden. Die rechtliche Diskussion sollte sich dabei an der technischen Entwicklung orientieren und möglichst flexible Lösungen wählen. Für bestimmte Bereiche bietet sich eine Normierung konkreterer Vorgaben an.

⁴⁷ Siehe hierzu La95.

⁴⁸ Siehe zu verschiedenen Möglichkeiten hierzu Bi18.

⁴⁹ In Betracht kommt beispielsweise die Darstellung der Gesamtverteilung, wie dies beispielsweise bei Bi18 untersucht wurde. Soweit die entsprechenden lokalen Daten frei verfügbar sind, wie bei amtlichen Statistiken, könnte ein Vergleich erfolgen.

6 Zusammenfassung

Für die Kontrolle algorithmischer Entscheidungen ist es entscheidend eine angemessene Qualität der verwendeten Datenbestände sicherzustellen. Eine verzerrte oder fehlerhafte Datengrundlage führt zu fehlerhaften Entscheidungen. Die Datenschutzgrundverordnung bietet mit Art. 5 Abs. 1 lit. d DSGVO einen Anknüpfungspunkt für eine Debatte, ist jedoch für eine praktische Anwendung derzeit noch nicht hinreichend bestimmt. Ansätze aus anderen Rechtsordnungen oder technischen Standards können jedoch zur Konkretisierung herangezogen werden. Änderungsbedarf besteht auch bei den Möglichkeiten der Betroffenen sowie der Aufsichtsbehörden zur Überprüfung der Datenqualität algorithmischer Entscheidungen. Es empfiehlt sich eine Fortführung des prozeduralen Ansatzes der Datenschutzgrundverordnung. Die Information der Betroffenen kann sich an der allgemeinen Diskussion zur Erklärung von Algorithmen orientieren. Eine konsequente Überprüfung der Datenqualität stellt einen relevanten Baustein bei der Kontrolle algorithmischer Entscheidungen dar.

Literaturverzeichnis

- [Ar18] Article 29 Datenschutzgruppe: Working Paper 251. Leitlinien zu automatisierten Entscheidungen im Einzelfall einschließlich Profiling für die Zwecke der Verordnung 2016/679, 2018.
- [Bi18] Binns, R. et al.: 'It's Reducing a Human Being to a Percentage'. In (CHI Hrsg.): CHI 2018. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, April 21-26, 2018, Montreal, QC, Canada. ACM, New York, NY, 2018; S. 1–14.
- [CZ15] Cai, L.; Zhu, Y.: The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. In *Data Science Journal*, 2015, 14; S. 2.
- [Di19] Dix, A.: DSGVO Art. 15: Simitis/Hornung/Spiecker, *Datenschutzrecht*, 2019; 25, 26.
- [Ga] Gasser, U.: Information Quality and the Law. or How to Catch A Difficult Horse. In *Berkman Center Research Publication*, No. 2003-08.
- [He] Heinrich, B. et al.: Requirements for Data Quality Metrics. In *Journal of Data and Information Quality*, 2018, 9; S. 1–32.
- [Ho16a] Hoeren, T.: Big Data und Datenqualität - ein Blick auf die DS-GVO. Annäherungen an Qualitätsstandards und deren Harmonisierung. In *ZD*, 2016; S. 459–463.
- [Ho16b] Hoeren, T.: Thesen zum Verhältnis von Big Data und Datenqualität. Erstes Raster zum Erstellen juristischer Standards. In *MMR*, 2016; S. 8–11.
- [Ki17] Kirkpatrick, K.: It's not the algorithm, it's the data. In *Communications of the ACM*, 2017, 60; S. 21–23.
- [Kr17] Kroll, J. A. et al.: Accountable Algorithms. In *University of Pennsylvania Law Review*, 2017, 165; S. 633.

- [La95] Ladeur, K.-H.: Das Umweltrecht der Wissensgesellschaft von der Gefahrenabwehr zum Risikomanagement, Berlin, 1995.
- [MBP] Mirzaie, M.; Behkamal, B.; Paydar, S.: Big Data Quality: A systematic literature review and future research directions.
- [Pa15] Pasquale, F.: The Black Box Society. The Secret Algorithms That Control Money and Information. Harvard University Press, 2015.
- [Ro19] Roßnagel, A.: DSGVO Art. 5: Simitis/Hornung/Spiecker, Datenschutzrecht, 2019; 139–141.
- [RSG16] Ribeiro, M. T.; Singh, S.; Guestrin, C.: "Why Should I Trust You?". In (Krishnapuram, B. Hrsg.): Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, New York, NY, 2016; S. 1135–1144.
- [Sca] Schmidt-Wudy, F.: DSGVO Art. 15: BeckOK Datenschutzrecht, Wolff/Brink; 76–80.
- [Scb] Schantz, P.: DSGVO Art. 5: BeckOK Datenschutzrecht, Wolff/Brink; 27–31.
- [Sh18] Sheppard, B.: Warming up to Inscrutability: How technology could challenge our concept of law. In University of Toronto Law Journal, 2018, 68.
- [TB98] Tayi, G. K.; Ballou, D. P.: Examining data quality. In Communications of the ACM, 1998, 41; S. 54–57.
- [TSD18] Taleb, I.; Serhani, M. A.; Dssouli, R.: Big Data Quality: A Survey: 2018 IEEE International Congress on Big Data, BigData Congress 2018, San Francisco, CA, USA, July 2-7, 2018, 2018; S. 166–173.
- [Wa] Washington, A. L.: How to argue with an algorithm. Lessons from the COMPAS ProPublica debate. In Colorado Technology Law Journal, 17; Issue 1.
- [Wi18] Wischmeyer, T.: Regulierung intelligenter Systeme. In Archiv des öffentlichen Rechts, 2018, 143; S. 1–66.
- [Wo] Worms, C.: DS-GVO Art. 16: BeckOK Datenschutzrecht, Wolff/Brink; 47–56.