DISKRIMINIERUNG UND FRAUENFEINDLICHKEIT – KI ALS SPIEGEL UNSERER GESELLSCHAFT

Maximilian Leicht / Julia Karst / Jasmin Zimmer

Maximilian Leicht, Wissenschaftlicher Mitarbeiter, Lehrstuhl für Rechtsinformatik, Universität des Saarlandes, 66123 Saarbrücken, DE, maximilian.leicht@uni-saarland.de, https://www.uni-saarland.de/lehrstuhl/sorge

Julia Karst, Studentische Mitarbeiterin, Lehrstuhl für Rechtsinformatik, Universität des Saarlandes, 66123 Saarbrücken, DE julia.karst@uni-saarland.de, https://www.uni-saarland.de/lehrstuhl/sorge

Jasmin Zimmer, Studentische Mitarbeiterin, Lehrstuhl für Rechtsinformatik, Universität des Saarlandes, 66123 Saarbrücken, DE, jasmin.zimmer@uni-saarland.de, https://www.uni-saarland.de/lehrstuhl/sorge

Schlagworte: ***Künstliche Intelligenz, Maschinelles Lernen, Diskriminierung, Regulierungsvorschläge***

Abstract: *KI-Systeme werden vermehrt bei zentralen Geschäftsprozessen in Unternehmen eingesetzt. Im Rahmen der interdisziplinären Diskussion wurden in der Literatur mögliche Diskriminierungen durch den Einsatz von KI-Systemen bereits als eine beachtliche Problemstellung identifiziert. Um eine fokussiertere Darstellung der Gefahren und Ursachen, sowie von potenziellen Regulierungsansätzen zu ermöglichen, konzentriert sich dieser Beitrag auf mögliche geschlechterdiskriminierende Folgen des Einsatzes von KI in der Privatwirtschaft. Zudem werden weitere Herausforderungen dargelegt, die sich im Zusammenhang mit verantwortungsbewusster Digitalisierung ergeben können.*

1. Ausgangspunkt

Im Zeitalter der digitalen Revolution erlangt Künstliche Intelligenz (KI) im Privat- und Berufsleben eine immer größere Bedeutung. Der wachsende Einfluss von KI auf unsere Gesellschaft stellt dabei sowohl für den Gesetzgeber als auch für Unternehmen eine Herausforderung dar. Neben den diversen Chancen neuer Technologien rücken aktuell auch mögliche Problemstellungen vermehrt in den Fokus.

Insbesondere der Themenbereich KI und Diskriminierung wurde zuletzt interdisziplinär sowie besonders in der juristischen Literatur mehrfach diskutiert.[1] In der Regel werden dabei die Gefahren und Ursachen von Diskriminierungen durch den Einsatz von KI-Systemen allgemein behandelt. Dieser Beitrag soll sich daher speziell dem Bereich KI und Frauendiskriminierung widmen. Hierfür werden sowohl relevante technische Ursachen analysiert als auch ausgewählte aktuelle Regulierungsvorschläge dargestellt.

Es erfolgt insofern eine Fokussierung auf den Einsatz von KI in der Privatwirtschaft. Wie die im Folgenden exemplarisch genannten Sachverhalte demonstrieren, zeigten sich gerade beim Einsatz von KI-Systemen durch Unternehmen in der Vergangenheit vermehrt Auffälligkeiten, die auf diskriminierende Elemente hinweisen können.

Bereits 2018 wurde die vierjährige Verwendung eines KI-Systems bei Amazon bekannt, welches im Rahmen von Bewerbungsverfahren im Einsatz war und bei dem eine diskriminierende Einstellungspraxis gegenüber Frauen festgestellt wurde. Als zentrale Ursache konnte insoweit eine mangelnde Qualität der Trainingsda-

[1] Vgl. etwa Steege, Algorithmenbasierte Diskriminierung durch Einsatz von Künstlicher Intelligenz, MMR 2019, S. 715, (S. 715–721); Hartmann, Diskriminierung aus der Black Box – Neue Herausforderungen durch KI-gestützte Personalentscheidungen, EuZA 2019, S. 421 (S. 421–422); Dzida/Groh, Diskriminierung nach dem AGG beim Einsatz von Algorithmen im Bewerbungsverfahren, NJW 2018, S. 1917 (S. 1917–1922); Beck, Diskriminierung durch Künstliche Intelligenz?, ZRP 2019, S. 185; Wischmeyer, Regulierung intelligenter Systeme, AöR 2018, S. 1 (S. 26ff.); Hacker, Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies against Algorithmic Discrimination under EU law, Common Market Law Review, 55, 2018, S. 1143 (S. 1144ff.).

ten der Software identifiziert werden. In diesen Daten waren Frauen unterrepräsentiert. Deshalb nahm das KI-System eine geschlechterdiskriminierende Vorsortierung der Bewerbungen vor. Trotz entsprechenden Nachbesserungsversuchen konnten diese in der Software angelegten Tendenzen nicht vollständig behoben werden, da das System weiterhin Zusammenhänge zwischen dem Geschlecht der Bewerber sowie der Bewerbung selbst fand. Letztendlich stellte Amazon den Betrieb der Software ein.[2]

Ein weiteres aktuelles Beispiel, in welchem Diskriminierungen nach dem Geschlecht vermutet werden, lieferte das Technologieunternehmen Apple. Bei der Verwendung einer neu eingeführten Kreditkarte des Konzerns («Apple Card»), die von Goldman Sachs betrieben wird, wurden von Kunden Auffälligkeiten bezüglich des Kreditrahmens festgestellt. Trotz vergleichbarer Inputdaten von Ehegatten wurden weiblichen Nutzern ein wesentlich geringerer Verfügungsrahmen angeboten als männlichen Nutzern. Nach mehreren Kundenberichten dieser Art hat das New York Department of Financial Services Ermittlungen eingeleitet.[3] Goldman Sachs bestreitet, dass durch den eingesetzten Algorithmus Ungleichbehandlungen nach dem Geschlecht möglich sind.[4]

Das Verständnis davon, was unter den Begriff der Diskriminierung zu fassen ist, unterscheidet sich allerdings nicht nur aufgrund der international verschiedenen rechtlichen Regelungen.[5] Bereits im deutschen Sprachraum bestehen unterschiedliche Auffassungen zwischen der Bedeutung im allgemeinen Sprachgebrauch und der Bedeutung im rechtlichen Kontext.[6] Dabei ist zu beachten, dass nicht jede Differenzierung per se eine unzulässige Diskriminierung darstellt.[7] Im Hinblick auf die im Folgenden dargestellten Regulierungsvorschläge wird in diesem Beitrag der Begriff Diskriminierung als Bezeichnung für eine unerwünschte Ungleichbehandlung verwendet.[8]

2. Technische Hintergründe

2.1. Künstliche Intelligenz und Maschinelles Lernen

Die Diskussion um Künstliche Intelligenz wird durch eine gewisse begriffliche Unschärfe erschwert, die besonders in der juristischen,[9] aber auch in der technischen[10] Literatur thematisiert wird. Die Ursachen hierfür sind vielfältig und werden teilweise bereits in der wörtlichen Übersetzung von «intelligence» gesehen.[11] Im Hinblick auf die Themensetzung dieses Beitrags werden lediglich KI-Systeme betrachtet, die eine bestimmte Eigenschaft aufweisen. Diese besteht darin, aus einer vorgegebenen Datenmenge eigenständig zu lernen,

---

2 Dastin, Amazon scraps secret AI recruiting tool that showed bias against women, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G (aufgerufen am 23. September 2019), Reuters, 2018.

3 Nedlund, Apple Card is accused of gender bias. Here's how that can happen. https://edition.cnn.com/2019/11/12/business/apple-card-gender-bias/index.html (aufgerufen am 14. November 2019), CNN Business 2019.

4 Vgl. die lediglich über Twitter abgegebene Stellungnahme: GS Bank Support, https://twitter.com/gsbanksupport/status/1194022629419704320 (aufgerufen am 19.11.2019), 2019.

5 Zu den Antidiskriminierungsgesetzen und dem Verständnis von Diskriminierung in den USA vgl. Steege, Algorithmenbasierte Diskriminierung durch Einsatz von Künstlicher Intelligenz, MMR 2019, S. 715 (S. 717).

6 Vgl. Gesellschaft für Informatik, Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren. Studien und Gutachten im Auftrag des Sachverständigenrats für Verbraucherfragen. Berlin: Sachverständigenrat für Verbraucherfragen, 2018, S. 84 (im Folgenden zitiert als GI-Gutachten «Algorithmische Entscheidungsverfahren»).

7 GI-Gutachten «Algorithmische Entscheidungsverfahren», S. 84; Beck, Diskriminierung durch Künstliche Intelligenz?, ZRP 2019, S. 185.

8 Vgl. für eine weitergehende Darstellung GI-Gutachten «Algorithmische Entscheidungsverfahren», S. 84–93.

9 Vgl. nur Wischmeyer, Regulierung intelligenter Systeme, AöR 2018, S. 1 (S. 3); Scherer, Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies, Harvard Journal of Law & Technology, Vol. 29, N. 2, Spring 2016, S. 354 (S. 359ff.); Ernst, Algorithmische Entscheidungsfindung und personenbezogene Daten, JZ 2017, S. 1026 (S. 1027) m. w. N. in Fn. 9.

10 Vgl. Ertel, Grundkurs Künstliche Intelligenz. Eine praxisorientierte Einführung[4. Auflage], Springer, Wiesbaden 2016, S. 1–3; Russel/Norvig, Artificial Intelligence. A Modern Approach[3. Ed], S. 1ff.

11 Vgl. dazu etwa Klindt, Editorial, K&R Beilage 1 zu Heft 7/8/2019, der für viele Anwendungen die Verwendung des Wortes «Hochleistungsalgorithmus» als zutreffender ansieht; kritisch zur wörtlichen Übersetzung aber auch: Herberger, «Künstliche Intelligenz» und Recht. Ein Orientierungsversuch, NJW 2018, S. 2825 (S. 2826).

sowie aus dieser Rückschlüsse ziehen zu können.[12] Das hat zur Folge, dass KI-Systeme (in Abgrenzung zu anderen, nicht lernfähigen Algorithmen, die festen Wenn-Dann-Schemata folgen) einen bestimmten Grad an möglicher Entscheidungsfindung aufweisen, der auf den bisherigen Lernerfahrungen basiert. Gerade daraus können sich vor Einsatz eines KI-Systems unbeabsichtigte Effekte ergeben, weshalb die eigenständige Lernfähigkeit aus großen Datenmengen für die folgenden Ausführungen das entscheidende Definitionselement darstellt.

Diese Lernfähigkeit zeigt sich in der Praxis insbesondere in der Unterkategorie des Maschinellen Lernens (*Machine Learning/ML*),[13] welche die Erkennung von Mustern in umfangreichen Datenmengen ermöglicht. Basierend auf dieser Analyse der bekannten Datensätze (*Trainings*daten) ist das KI-System zur Generalisierung fähig und damit in der Lage, neue Vorhersagen treffen zu können. Die Qualität dieser Prognosefähigkeit wird anschließend mittels neuer, zuvor unbekannter Eingaben (*Test*daten) überprüft.[14]

Bezüglich der konkreten Implementierung der Mustererkennung ist für die Zwecke dieses Beitrags eine Darstellung der grundlegenden Prinzipien ausreichend. Beruhend auf der Verarbeitung der Trainingsdaten werden die Parameter des Programms solange adaptiert, bis sich die Genauigkeit der Vorhersagen nicht weiter verbessern lässt.[15] Gerade diese Lernfähigkeit und die damit verbundene Möglichkeit, sehr große Datenmengen zu verarbeiten und zu interpretieren, stellt eine Stärke von Maschinellem Lernen dar.

2.2. Kategorien des Maschinellen Lernens

Im Folgenden werden verschiedene Kategorien des Maschinellen Lernens genauer beschrieben. Grundsätzlich wird zwischen den zwei wesentlichen Kategorien des Supervised und Unsupervised Learning unterschieden.

Bei Supervised Learning sind sowohl Input als auch Output bereits bekannt. Das Ziel dabei besteht darin, die Daten gemäß der vorgesehenen Kategorien einzuteilen.[16] Wird der Algorithmus beispielsweise dazu verwendet Bilder von Hunden und Katzen zu unterscheiden, so werden als Trainingsdaten möglichst viele Fotos dieser Tiere eingesetzt. Dies ermöglicht es dem System durch nachträglichen Abgleich mit den korrekten Ergebnissen seine Entscheidung zu validieren.

Im Gegensatz dazu werden im Falle des Unsupervised Learnings große Mengen an Daten ohne feste Zielvorgabe übermittelt. Anschließend soll der Algorithmus eigenständig Muster in den Trainingsdaten erkennen.[17]

In besonderem Fokus soll im Folgenden die Unterkategorie des Deep Learnings stehen, welche auf künstlichen neuronalen Netzen basiert. Als eine der leistungsfähigsten Varianten des Maschinellen Lernens wird Deep Learning vermehrt eingesetzt. Das Verfahren weist aufgrund seiner Komplexität jedoch auch eine entsprechende Intransparenz auf.[18]

---

[12] Wischmeyer, Regulierung intelligenter Systeme, AöR 2018, S. 1 (S. 3); Martini, Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz, Springer, Berlin, 2019, S. 20f.

[13] Vgl. GI-Gutachten «Algorithmische Entscheidungsverfahren», S. 30; Mata et. al., Artificial Intelligence (AI) methods in optical networks: A comprehensive survey, Optical Switching and Networking 28, 2018, S. 43 (S. 45, Fig. 1).

[14] Vgl. Daumé, A Course in Machine Learning, 2012, S. 8–10; Hacker, Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies against Algorithmic Discrimination under EU law, Common Market Law Review, 55, 2018, S. 1143 (S. 1146f.); Fraunhofer-Gesellschaft, Maschinelles Lernen. Eine Analyse zu Kompetenzen, Forschung und Anwendung, 2018, S. 9.

[15] Vgl. Gollapudi/Laxmikanth, Practical machine learning: tackle the real-world complexities of modern machine learning with innovative and cutting-edge techniques; Packt Publishing, Birmingham, 2016, S. 5.

[16] GI-Gutachten «Algorithmische Entscheidungsverfahren», S. 30; Goodfellow/Bengio/Courville, Deep Learning, http://www.deeplearningbook.org (aufgerufen am 16.11.2019), MIT Press, 2016, Kap. 5, S. 103.

[17] Vgl. Goodfellow/Bengio/Courville, Deep Learning, http://www.deeplearningbook.org (aufgerufen am 16.11.2019), MIT Press, 2016, Kap. 5, S. 103.

[18] Vgl. Fraunhofer-Gesellschaft, Maschinelles Lernen. Eine Analyse zu Kompetenzen, Forschung und Anwendung, 2018, S. 11, 12; Martini, Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz, Springer, Berlin, 2019, S. 43.

2.3. Deep Learning und Neuronale Netze

Die Komplexität von Deep Learning ergibt sich durch den Aufbau aus mehrschichtigen künstlichen neuronalen Netzwerken (KNN), welche sich am biologischen Vorbild der Funktionsweise des menschlichen Gehirns orientieren. KNN bestehen aus Schichten von Recheneinheiten, die Neuronen genannt werden. Die Schichten gliedern sich in die Eingabeschicht, eine gewisse Anzahl an von außen nicht sichtbaren Verarbeitungsschichten (Hidden Layers), sowie die Ausgabeschicht.[19]

Im Rahmen des überwachten Lernens mit sog. FeedForward-Netzen sind die Verbindungen der einzelnen Neuronen verschiedener Schichten zur Ausgabeschicht gerichtet. Dabei ist nur eine Verbindung zu Neuronen der nächsten Schicht erlaubt.[20] Abhängig von der Eingabe leiten so bestimmte Neuronen der ersten Schicht Signale zur nächsten Schicht weiter. Dadurch entsteht eine Abfolge von Verarbeitungsschritten, die letztendlich den Output generiert (vgl. Abbildung 1).

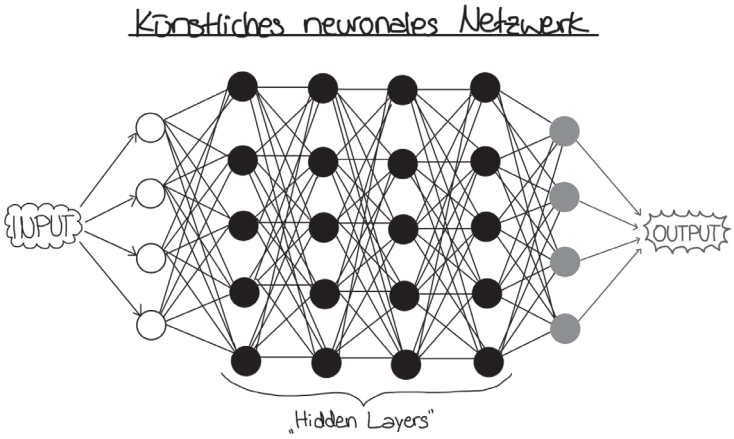

Abbildung 1: Vereinfachte Darstellung eines KNN (eigene Darstellung)

Diese Verarbeitung ist abhängig von den Gewichten der einzelnen Verbindungen, welche wiederum beeinflussen, wie gut Daten zwischen den jeweiligen Neuronen übertragen werden. Die jeweils eintreffenden Daten werden so verarbeitet, dass sich mit ihnen die neue Aktivierung bestimmen lässt. Von dem Aktivierungszustand hängt ab, welchen Einfluss das Neuron auf den weiteren Ablauf hat. Dazu existiert ein Schwellwert, welcher den Punkt markiert, ab dem das Neuron «feuert». In der Regel wird anschließend die Aktivierung als Ausgabe des jeweiligen Neurons übergeben.[21]

Üblicherweise lernt ein solches KNN durch Veränderung seiner Gewichte und Schwellwerte. Nach der Erzeugung des Outputs wird dieser mit dem korrekten Ergebnis verglichen, wodurch sich ein Fehlervektor ergibt, mithilfe dessen die Qualität des Netzes verbessert werden kann.[22] Nachdem das KNN auf diese Weise auf den Trainingsdaten trainiert wurde, wird es mit den Testdaten überprüft.

---

[19] Vgl. Faris/Aljarah/Mirjalili, Training feedforward neural networks using mulit-verse optimizer for binary classification problems, Appl Intell 45:322-332, Springer US, 2016 (im Folgenden zitiert als Faris/Aljarah/Mirjalili, Training feedforward neural networks), S. 322 (S. 323); Russel/Norvig, Artificial Intelligence. A Modern Approach[3. Ed], S. 729; Flasinski, Introduction to Artificial Intelligence, Springer, 2016, S. 168.

[20] Vgl. Faris/Aljarah/Mirjalili, Training feedforward neural networks, S. 322 (S. 323); Russel/Norvig, Artificial Intelligence. A Modern Approach[3. Ed], S. 729.

[21] Vgl. Faris/Aljarah/Mirjalili, Training feedforward neural networks, S. 322 (S. 323); Flasinski, Introduction to Artificial Intelligence, Springer, 2016, S. 159–161; Pohlmann, Künstliche Intelligenz und Cyber-Sicherheit, S. 537.

[22] Vgl. Kubat, An Introduction to Machine Learning[2nd Edition], Springer, 2017, S. 96. Styczynski/Rudion/Naumann, Einführung in Expertensysteme, Springer, Berlin/Heidelberg 2017, S. 158f.

Eine häufig angewendete Methode zu einer solchen Verbesserung der Funktionsweise des Netzes ist die sog. Backpropagation, welche auf dem Verfahren des Gradientenabstiegs basiert. Der Gradient ist ein Vektor, der von einem Punkt aus in die Richtung des steilsten Anstiegs zeigt. Der Gradientenabstieg nutzt nun den Gradienten, um von einem beliebigen Startpunkt aus bergab zu gehen.[23] Um die Backpropagation durchzuführen, wird nun begonnen vom Ende des Netzes aus den Gradientenabstieg rückwärts durch das Netz gehend auszuführen.[24]

Es lässt sich somit feststellen, dass ein KNN stets seine inneren Parameter modifiziert, um ein besseres Ergebnis zu erhalten. Welches der zahlreichen Neuronen in einem der vielen Schichten aufgrund seiner spezifischen Parameter nun aber für eine bestimmte Reaktion zuständig ist, lässt sich in der Regel nicht mehr nachvollziehen.

2.4. Intransparenz algorithmenbasierter Entscheidungen

Eine der wesentlichen Ursachen der Intransparenz besteht somit im technischen Aufbau der verwendeten KI-Systeme. Die komplex miteinander vernetzten Hidden Layers erschweren den Einblick in die einzelnen inneren Entscheidungsvorgänge. Zudem besteht die Möglichkeit, dass nicht nur während des Trainingsprozesses, sondern auch beim späteren Einsatz der KI-Systeme nochmalige Anpassungen der internen Gewichtung erfolgen können, um die Entscheidungsfindung weiter zu optimieren.[25] Durch diese Intransparenz wird die Vertrauensbildung in KI erschwert[26] und Betroffene, die eine algorithmenbasierte Diskriminierung vermuten, können diesen Verdacht selten unmittelbar bestätigen.

3. Kategorisierung von Bias

Eine grundlegende Maßnahme zur Vermeidung von diskriminierenden Entscheidungen liegt in der Kontrolle der Qualität der Trainings- und Testdaten. Eine ggf. vorliegende verzerrte Wiedergabe der Realität innerhalb des Datenmaterials wird als Bias bezeichnet. Die Ursachen für eine solche Verzerrung sind vielfältig und werden innerhalb der Literatur in verschiedene Begriffe untergliedert.[27] Im Folgenden wird die Einteilung in fünf verschiedene Bias-Arten nach Suresh/Guttag dargestellt.[28] Der Kategorisierung ist zu entnehmen, an welchen Stellen des Entwicklungsprozesses die Gefahr von Bias bei falsch ausgewählten Trainings- bzw. Testdaten besteht oder nicht erkannt werden könnte.[29] Ist eine solche Stelle bekannt, so fällt es wesentlich leichter, dem Bias entgegenzuwirken, da je nach dessen Herkunft eine andere Strategie zur Minimierung des Bias erfolgsversprechend ist.[30]

Der **Historical Bias** beschreibt das Phänomen, dass historische Faktoren, etwa in Form von in der Realität bestehenden Ungleichbehandlungen, eine verzerrte Datenbasis verursachen.[31] Beispielsweise liegt der Frauenanteil

---

23 Vgl. Kubat, An Introduction to Machine Learning[2nd Edition], Springer, 2017, S. 97f; Ertel, Grundkurs Künstliche Intelligenz. Eine praxisorientierte Einführung[4. Auflage], Springer, Wiesbaden 2016, S. 289.

24 Vgl. Daumé, A Course in Machine Learning, 2012, S. 118; Goodfellow/Bengio/Courville, Deep Learning, http://www.deeplearningbook.org (aufgerufen am 16.11.2019), MIT Press, 2016, Kap. 6.5.

25 Wischmeyer, Regulierung intelligenter Systeme, AöR 2018, S. 1 (S. 47).

26 Martini, Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz, Springer, Berlin, 2019, S. 324, 362.

27 Vgl. insoweit die anschauliche Darstellung bei Balkow/Eckardt, Denkimpuls Digitale Ethik: Bias in algorithmischen Systemen – Erläuterungen, Beispiele und Thesen, Initiative D21, 2019; eine weitere Einordnung findet sich etwa bei Beck et al., Künstliche Intelligenz und Diskriminierung. Herausforderung und Lösungsansätze, AG IT-Sicherheit, Privacy, Recht und Ethik, 2019, S. 8f.

28 Suresh/Guttag, A Framework for Understanding Unintended Consequences of Machine Learning, https://arxiv.org/pdf/1901.10002.pdf (aufgerufen am 20.11.2019, im Folgenden zitiert als Suresh/Guttag, Framework), 2019, S. 2, 3ff.

29 Suresh/Guttag, Framework, S. 2, Fig. 1.

30 Vgl. insoweit die beispielhafte Darstellung bei Suresh/Guttag, Framework, S. 1.

31 Suresh/Guttag, Framework, S. 4.

in Vorständen von DAX-Unternehmen im Jahr 2018 nur bei ca. 14%.[32] Es entspräche daher der Realität, diese Zahlen bei einer Bildersuche zu spiegeln, was jedoch nicht für jeden Anwendungsfall erstrebenswert erscheint. Ähnliche Tendenzen lassen sich aber auch den Statistiken zur ungleichen Einkommensverteilung entnehmen.[33] Ein auf dieser Grundlage basierendes KI-System könnte daher eine Verfestigung des status quo verursachen.

Representation Bias hingegen entsteht erst bei der Selektierung von Personengruppen für einen Datensatz. Sind Teilmengen dieses Datensatzes, wie z.B. eine bestimmte Bevölkerungsgruppe unterrepräsentiert, so kann dies zu verfälschten Ergebnissen führen. Ein solcher Bias zeigt sich häufig im Bereich der Bilderkennung. In einer Studie, die drei kommerzielle Gesichtserkennungstools evaluierte, wurde teilweise eine Fehlerrate bei dunkelhäutigen Frauen von knapp einem Drittel festgestellt. Im Vergleich dazu lag die Fehlerquote bei weißen Männern unter 1%.[34] Nichtrepräsentatives Datenmaterial kann daher zu Diskriminierungen führen.[35]

Eine Entfernung der expliziten Indikatoren, wie etwa Geschlecht, Name oder Personalpronomen ist zudem nicht zwingend ausreichend, da geschlechtsspezifische Unterschiede dadurch nicht vollständig negiert werden.[36] KI-Systeme sind grundsätzlich dazu fähig in einem solchen Datensatz für Menschen bereits nicht mehr ersichtliche Zusammenhänge dennoch zu erkennen, weshalb ein ausgeglichenes Trainingsdatenmaterial stets zur Minimierung der Gefahr eines Representation Bias genutzt werden kann.

Measurement Bias kann auftreten, wenn zwar eine korrekte Messung gewisser Eigenschaften erfolgt, diese aber nicht repräsentativ sind. Häufig stehen die gemessenen Daten dann nur stellvertretend für die eigentlich gewünschte Eigenschaft.

Die bisher genannten Bias-Arten treten während der Zusammenstellung der Trainingsdaten auf. Nach dieser Phase sind vor allem der Evaluation Bias oder der Aggregation Bias relevant.

Der **Evaluation Bias** tritt während der Iteration und der Evaluierung des KI-Systems auf. Wurden die Testdaten nicht repräsentativ ausgewählt, beeinträchtigt dies die Funktionalität des Modells bei den betroffenen Teilmengen. Dabei ergibt sich häufig ein Zusammenhang mit dem Representation Bias.

Im Rahmen des Beispiels zur Gesichtserkennung waren Bilder von Frauen bereits in den Trainingsdaten unterrepräsentiert (Representation Bias). Werden bei der Evaluation des KI-Systems nun erneut zu wenige Bilder von Frauen verwendet (Evaluation Bias), so wird die Erkennung der Unterrepräsentation erschwert.[37]

Der **Aggregation Bias** dagegen entsteht dadurch, dass Rückschlüsse aus der betrachteten (ggf. zu diversen) Personengruppe auf Einzelne bezogen werden. Diese ggf. fehlerhafte Pauschalisierung beeinflusst dann die Funktionalität des KI-Systems. Ein solcher Bias ist besonders bei medizinischen Anwendungen relevant, da dieselbe Krankheit bei jedem Patienten mit anderen Symptomen auftreten kann.[38]

Die oben dargestellte Einordnung der Bias-Arten lässt detaillierter erkennen, anhand welcher Kriterien die Qualität des Trainings- und Testdatenmaterials beurteilt werden sollten, um insbesondere geschlechterdiskriminierende Folgen zu verhindern.

---

[32] DIW Berlin, Frauenanteil in Vorständen der DAX-Unternehmen (DAX-30) in Deutschland von 2011 bis 2018, Statista, https://de.statista.com/statistik/daten/studie/409010/umfrage/frauenanteil-in-dax-vorstaenden/ (aufgerufen am 26.11.2019), 2019; vgl. aber auch DIW Berlin, Frauenanteil in den Vorständen der 200 größten deutschen Unternehmen bis 2018, https://de.statista.com/statistik/daten/studie/180102/umfrage/frauenanteil-in-den-vorstaenden-der-200-groessten-deutschen-unternehmen/ (aufgerufen am 21.11.2019), 2019.

[33] Statistisches Bundesamt, Gender Pay Gap: Verdienstabstand zwischen Männern und Frauen in Deutschland von 1995 bis 2018, https://de.statista.com/statistik/daten/studie/3261/umfrage/gender-pay-gap-in-deutschland/ (aufgerufen am 21.11.2019), 2019.

[34] Buolamwini/Gebru, Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification, Proceedings of the 1st Conference on Fairness, Accountability and Transparency, in PMLR 81:77-91, 2018, S. 77, (S. 77, 87).

[35] GI-Gutachten «Algorithmische Entscheidungsverfahren», S. 34f.

[36] De-Arteaga et al., Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting, in: Proceedings of the 1st Conference on Fairness, Accountability and Transparency, S. 120–128, ACM, 2019, S. 127.; Vgl. zum Thema Scrubbing auch Kleinberg/Ludwig/Mullainathan/Rambachan, Algorithmic fairness, In AEA Papers and Proceedings, Vol. 108, May 2018, S. 22.

[37] Suresh/Guttag, Framework, S. 5.

[38] Suresh/Guttag, Framework, S. 5.

4. Darstellung existierender Regulierungsvorschläge

Im Folgenden werden ausgewählte Regulierungsvorschläge beschrieben. Entscheidend ist dabei für die Zwecke dieser Darstellung, inwieweit die genannten Maßnahmen die Eintrittswahrscheinlichkeit von geschlechterspezifischen Diskriminierungen minimieren können. Festzuhalten ist dabei zunächst, dass im Grundsatz eine EU-weit einheitliche Regulierung angestrebt wird.[39] Dies ist jedoch nicht nur angesichts der dadurch erhofften Vorteile einer Harmonisierung der Regulierung und entsprechender Effektivität bei der Durchsetzung zu begründen. Vielmehr ist es den einzelnen Mitgliedstaaten aufgrund der bestehenden europarechtlichen Rechtsakte in einigen Rechtsgebieten nicht möglich, im notwendigen Maße nationale Regelungen zu erlassen, was insbesondere auf den Bereich des Datenschutzrechts zutrifft.[40]

4.1. Kennzeichnung

Die Einführung einer Kennzeichnungspflicht kann es potentiell Betroffenen grundsätzlich erst einmal ermöglichen, von einer Ungleichbehandlung bzw. von dem KI-System als denkbare Ursache zu erfahren. Eine Kennzeichnung von automatisierter Entscheidungsfindung oder -vorbereitung kann daher einen relevanten Beitrag zu effektiverem Rechtsschutz leisten.

Abzugrenzen ist eine bloße Kennzeichnung von KI-Systemen allerdings von Informationspflichten, die, wie etwa nach Art. 13 Abs. 2 lit. f, Art. 14 Abs. 2 lit. g DSGVO, den Betroffenen im Rahmen des Anwendungsbereichs des Art. 22 DSGVO zwar prinzipiell bereits eine Kenntnis über das Bestehen einer automatisierten Entscheidungsfindung i. S. d. Art. 22 DSGVO ermöglichen. Unter Kennzeichnung wird in diesem Beitrag allerdings die leicht zugängliche und leicht verständliche Information des Nutzers über den Einsatz eines KI-Systems verstanden, wie es etwa auch über ein vereinheitlichtes Symbol erfolgen kann. Solche simplen Darstellungen entsprechen dagegen nicht den umfangreichen Informationskatalogen aus Art. 13, 14 DSGVO. Zwar ist es den Verantwortlichen möglich im Rahmen des Art. 12 Abs. 7 S. 1 DSGVO in Kombination zu den Informationen nach Art. 13, 14 DSGVO standardisierte Bildsymbole bereitzustellen, durch welche bei entsprechendem Einsatz durchaus eine leicht zugängliche und leicht verständliche Information des Nutzers erfolgen kann. Die Regelung stellt jedoch nur eine zusätzliche Option des Verantwortlichen dar und ist insoweit nicht verpflichtend.[41]

Gleichzeitig ist festzuhalten, dass im Falle fehlender weiterer Informationen eine reine Kennzeichnung den Betroffenen im Wesentlichen nur vor die Wahl stellt, soweit dies praxisgerecht im konkreten Fall überhaupt möglich ist, vollständig auf die Nutzung des Systems zu verzichten. Des Weiteren wird eine effektive Handhabung einer Kennzeichnungspflicht durch die fehlende Einigkeit bei der Definition ihres Anwendungsbereichs erschwert, sodass eine Differenzierung bei der Ermittlung der Notwendigkeit einer Kennzeichnung in der Praxis zu Umsetzungsschwierigkeiten führen dürfte.[42]

4.2. «Algorithmic Responsibility Codex»

Martini schlägt zudem vor, das Konzept des «Corporate Governance Kodex», welches § 161 AktG zugrunde liegt, zu übertragen und eine entsprechende «Regierungskommission Algorithmic Responsibility Codex»

[39] Vgl. etwa Gutachten der Datenethikkommission, 2019, S. 180–182 (im Folgenden zitiert als: DEK-Gutachten); Martini, Kontrollsystem für algorithmenbasierte Entscheidungsprozesse, Speyer, 2019, S. 67f.

[40] Detailliert zu den Gestaltungsoptionen der nationalen Gesetzgeber: Martini, Kontrollsystem für algorithmenbasierte Entscheidungsprozesse, Speyer, 2019, S. 66–75.

[41] BeckOK DatenschutzR, Quaas, 29. Ed., 1.8.2019, DS-GVO Art. 12 Rn. 53; Kühling/Buchner DS-GVO, Bäcker, Art. 12 Rn. 20; Martini, Kontrollsystem für algorithmenbasierte Entscheidungsprozesse, Speyer, 2019, S. 10f.

[42] Eine kritische Beurteilung findet sich etwa im GI-Gutachten «Algorithmische Entscheidungsverfahren», S. 162.

einzurichten.[43] Der Regelungskodex folgt dabei dem comply-or-explain Prinzip. Das bedeutet, dass die betroffenen Unternehmen den Kodex entweder implementieren müssen oder eine Begründung der fehlenden Umsetzung veröffentlichen müssen. Im Falle des Algorithmic Responsibility Codex sollen nach diesem Vorschlag fehlerhafte Erklärungen bußgeldbewehrt sein.[44] Um die Wirkung eines solchen Kodex' noch zu verstärken, wäre zudem denkbar, den betroffenen Unternehmen, die sich ohne Einschränkungen zur Einhaltung dessen verpflichten, ein entsprechendes Siegel bzw. Symbol für Kennzeichnungszwecke zur Verfügung zu stellen. Dies könnte einen zusätzlichen Anreiz für die Erfüllung der Empfehlungen aus dem Kodex bewirken.

4.3. Anpassung der Beweislastregelungen

Als weiterer, für die Themensetzung dieses Beitrags relevanter Regulierungsansatz ist die von Martini vorgeschlagene Anpassung der Beweislastverteilung im Haftungsprozess zu nennen. Demnach soll es in den einschlägigen Fällen ausreichen, dass Betroffene Tatsachen vortragen, die zu einer überwiegenden Wahrscheinlichkeit einen Schluss darauf zulassen, dass die Verarbeitung mittels unzulässiger Parameter (wie beispielsweise dem Geschlecht) erfolgte.[45] Eine solche Anpassung kann eine effektive Maßnahme gegen die oben dargestellte Intransparenz von KI-Systemen sein und wesentlich zu einer vermehrten Wahrnehmung von Rechtsschutzmöglichkeiten führen.

5. Fazit

Die aus der Komplexität des technischen Aufbaus resultierende Intransparenz von KI-Systemen, sowie die Analyse der vielfältigen Ursachen für Bias in den Trainings- und Testdaten, lassen auf die Entstehung von für den Einzelnen nicht mehr erkenn- bzw. überschaubaren Diskriminierungsrisiken schließen.

Ein dahingehender Regulierungsbedarf zum Schutz der Betroffenen beim Einsatz von KI in der Privatwirtschaft ist daher ersichtlich. Die verschiedenen publizierten Ansätze geben einen Ausblick auf die Vielfalt der Möglichkeiten. Zum Einstieg in den umfangreichen Regulierungsprozess erscheinen die oben genannten Vorschläge zum aktuellen Zeitpunkt als sinnvoll.

Bei der Zielsetzung der Regulierungsmaßnahmen ist jedoch ebenfalls zu beachten, dass Digitalisierungsprozesse nicht allein bereits existierende geschlechterdiskriminierende Phänomene auflösen können. So ist Frauenfeindlichkeit anerkanntermaßen ein allgemeines gesellschaftliches Problem. KI-Systeme verdeutlichen oftmals nur diese bereits angelegten Probleme und stellen daher einen Spiegel der Gesellschaft dar. Auf diese Weise können sich auch unbeabsichtigt diskriminierende Tendenzen in KI-Systemen wiederfinden.

Daher sollte ein regulatorischer Fokus darauf liegen, dass die bisherigen diskriminierenden gesellschaftlichen Strukturen möglichst nicht durch die technische Implementierung verfestigt werden.

---

[43] Martini, Kontrollsystem für algorithmenbasierte Entscheidungsprozesse, Speyer, 2019, S. 40f.; Martini, Algorithmen als Herausforderung für die Rechtsordnung, JZ 2017, S. 1017 (S. 1022f.); die Überlegungen fanden unter der Bezeichnung «Algorithmic Accountability Codex» Eingang in das Gutachten der DEK, DEK-Gutachten, S. 202.

[44] Martini, Kontrollsystem für algorithmenbasierte Entscheidungsprozesse, Speyer, 2019, S. 41.

[45] Vgl. Martini, Kontrollsystem für algorithmenbasierte Entscheidungsprozesse, Speyer, 2019, S. 36; Martini, Algorithmen als Herausforderung für die Rechtsordnung, JZ 2017, S. 1017 (S. 1023f.).