Paola Lopez

Artificial Intelligence und die normative Kraft des Faktischen

Den Gesundheitszustand von Patienten zu prognostizieren, um medizinische Präventionsmaßnahmen möglichst sinnvoll zu verteilen, ist schwieriger, als man denkt. Dabei scheint die Problemstellung denkbar simpel: Es sollen diejenigen Patientinnen zusätzliche Präventionsmaßnahmen erhalten, deren Gesundheitszustand sich zu verschlechtern droht. Doch der Gesundheitszustand ist, wie die meisten menschlichen Angelegenheiten, zu komplex, um ihn einheitlich messen und quantifizieren zu können. Das gilt auch dann, wenn alle Patientendaten vollständig zur Verfügung stehen und mit Big-Data-Methoden verarbeitet werden können.

Ein vielfach genutztes algorithmisches *health care management*-System aus den USA reduzierte die Komplexität dieser Aufgabe, indem es anstelle des Gesundheitszustands die aufzuwendenden medizinischen Kosten prognostizierte. In der Tat sollte man meinen, dass diese beiden Größen korrelieren, dass man also aus einer Kostenprognose auf den wahrscheinlichen Gesundheitszustand rückschließen könnte.

Die Sache hat allerdings einen Haken: Gesundheitskosten sind keine neutrale statistische Größe. Der Zugang zu medizinischen Ressourcen und damit die aufgewendeten Kosten hängen stark mit der sozioökonomischen Position von Patienten zusammen. Das ist überall so und erst recht in den USA, die kein flächendeckendes solidarisches Gesundheitssystem besitzen. Arme Menschen können sich weder eine gute Krankenversicherung noch teure Behandlungen leisten.

Schwarze Menschen sind in den USA sozioökonomisch systematisch schlechter gestellt als andere. Es korrelieren also Schwarzsein mit Armut und Armut mit geringen Ausgaben im Gesundheitsbereich. Dieser Effekt wird, wie empirische Studien zeigen, noch dadurch verstärkt, dass auch in der individuellen Beziehung zwischen Ärztinnen und Patienten schwarze Patientinnen weniger präventive und weiterführende Behandlungen erhalten.

All das sind Gründe, warum die statistische Größe der aufgewendeten Gesundheitskosten einem gravierenden *racial bias* unterliegt. Für das *health management tool* bedeutete das im Ergebnis: Schwarzen Patienten wurden systematisch weniger medizinische Präventionsmaßnahmen zugeteilt, da ihr Gesundheitszustand als weniger gravierend eingeschätzt wurde – auf Basis der Kostenprognose. Dieser Fehler wurde von Wissenschaftlerinnen durch

Zufall entdeckt und behoben.[1] Nach der Korrektur dieses Fehlers verdoppelte sich der Anteil der schwarzen Patienten, die nun zusätzliche Präventionsmaßnahmen erhielten.

Die Wissenschaftlerinnen, die den Fehler behoben, präsentierten diesen Fall auf der FAccT-Konferenz der Association for Computing Machinery. FAccT ist ein Akronym für »Fairness, Accountability and Transparency in Machine Learning«, und die jährlich stattfindende FAccT-Konferenz bündelt ein inter- und multidisziplinäres Forschungsfeld im Überschneidungsbereich von Technologie, Informatik und Mathematik, das auch gegenüber rechts-, geistes- und sozialwissenschaftlichen Ansätzen offen ist. Im Kern geht es dabei um die Frage, wie sich unerwünschte Effekte von algorithmischen Systemen erkennen, messen und mit ihrerseits technischen Mitteln einhegen und konterkarieren lassen.[2] Auch die EU-Kommission beschäftigt sich mittlerweile mit diesem Thema. Sie hat »Ethik-Guidelines« für datenbasierte algorithmische Systeme und Künstliche Intelligenz entworfen, genauer gesagt entwerfen lassen, und zwar von einem Gremium namens »High-Level Expert Group on Artificial Intelligence« (AI HLEG).[3]

Die Tatsache, dass falsch konzipierte algorithmische Systeme durch diskriminierenden Bias gravierenden Schaden anrichten können, ist also bekannt und wird diskutiert. Was aber, wenn AI-Systeme genau so gebaut werden, wie sie entlang ihres grundlegenden mathematischen Paradigmas gebaut werden sollen? Trotzdem – und gerade dann – sind Bias und diskriminierende Effekte in AI-Systemen in vielen Anwendungszusammenhängen bereits in der mathematischen Struktur angelegt und damit buchstäblich

1 Ziad Obermeyer / Brian Powers u. a., *Dissecting racial bias in an algorithm used to manage the health of populations*. In: *Science* vom 25. Oktober 2019 (science.sciencemag.org/content/366/6464/447).

2 Schon die Frage nach der Messung von diskriminierenden Effekten durch Daten-Bias ist nicht unkompliziert. Es hängt einiges davon ab, welche der zahlreichen Definitionen von »Fairness« man dabei anwenden möchte. Vgl. das *Translation tutorial: 21 fairness definitions and their politics* von Arvind Narayanan auf der ACM Conference for Fairness, Accountability and Transparency in Machine Learning 2018 (www.youtube.com/watch?v=wqamrPkF5kk).

3 European Commission, *Ethic Guidelines for Trustworthy AI* (ec.europa.eu/futurium/en/ai-alliance-consultation). Derartige Best-Practice-Richtlinien sind gleich doppelter Kritik ausgesetzt: Zum einen, weil sie fast immer auf Freiwilligkeit setzen und nicht juristisch durchsetzbar sind; zum anderen, weil sie in der Regel Individuen adressieren und nicht Organisationen oder Institutionen, als ließen sich gesamtgesellschaftlich begründete Probleme durch das Verhalten von Einzelnen lösen. Gegen den damit verbundenen Vorwurf des »ethics washing« wiederum wird eingewandt, dass mit den Ethik-Guidelines wenigstens ein Anfang gemacht sei.

»vorprogrammiert«. Denn AI-Systeme haben in ihrer mathematischen Architektur angelegte epistemische Grenzen, die in einem fundamentalen Missverhältnis zu ihren gegenwärtigen Anwendungszusammenhängen stehen. Diese Grenzen werden deutlich, wenn man sich eine einfache Frage stellt: Welches Wissen können diese Systeme produzieren – und welches nicht?

Paradigmenwechsel

Die AI-Forschung hat seit ihren Anfängen mindestens einen fundamentalen Paradigmenwechsel erlebt. War es zu Beginn der Entwicklung noch das Ziel, mittels Computerprogrammen einen kognitiven, menschenähnlichen Gedankenprozess zu simulieren (was in der Regel nicht gut funktionierte),[4] so gründet der neueste Stand der AI-Forschung heute auf Methoden des datenbasierten Machine Learning. Die Stärke und der springende Punkt der heutigen AI sind der Umgang mit großen Datenmengen. Durch gesteigerte Datenproduktions-, Datenspeicherungs- und Datenverarbeitungskapazitäten ist es nun möglich, ein Machine-Learning-System dahingehend zu trainieren, dass es das richtige (im Sinne von: erwünschte) Ergebnis für eine bestimmte Fragestellung produziert, indem riesige, aggregierte Datenmengen mit statistischen Methoden verarbeitet werden. Heutige AI-Programme »lernen« letztendlich durch die schiere Menge an Datenbeispielen. Die »Trainingsphase« der Entwicklung eines Machine-Learning-Systems bezeichnet genau das. Nach dem Wechsel von einem regelbasierten zu einem datenbasierten Paradigma wird nun nicht mehr die Frage gestellt, ob der Weg zu einem Ergebnis sinnvoll ist. Vielmehr wird entlang verschiedener mathematischer Gütekriterien beurteilt, ob das Ergebnis hinreichend zufriedenstellend ist. Ist das der Fall, dann wird stillschweigend unterstellt, dass auch der Weg dorthin sinnvoll gewesen sein muss.

Ein illustratives Beispiel ist das automatisierte Übersetzen von Text. Nach dem älteren AI-Paradigma funktionierte das Übersetzen von Text in eine andere Sprache durch das explizite Formalisieren und Einprogrammieren grammatikalischer Regeln. Das Programm sollte also letztendlich, wenn auch in einem sehr reduktiven Sinn des Wortes, »verstehen«, wie ein Satz zu übersetzen ist. Nun gibt es unzählige grammatikalische Regeln, zugleich

4 Vgl. den von Joseph Weizenbaum 1966 erstellten Chatbot »ELIZA«, der ein Gespräch mit einem Therapeuten simulieren und damit laut Weizenbaum die Oberflächlichkeit solcher Systeme aufzeigen sollte. Die Grenzen des Programms werden tatsächlich nach einigen Eingaben schnell deutlich (psych.fullerton.edu/mbirnbaum/psych101/Eliza.htm).

gibt es nur durch Deutung Erkennbares wie Sarkasmus und Ironie, sowie überhaupt kontextuelle Faktoren. Entsprechend ineffizient und fehlerhaft waren die Ergebnisse im Großen und Ganzen. Nach dem aktuellen AI-Paradigma funktionieren automatisierte Übersetzungen dagegen durch »Training«: Sie verarbeiten ein riesiges Textkorpus in den jeweiligen Sprachen, so dass das Programm am Ende zwar nicht »versteht«, warum ein Satz so oder so übersetzt wird. Anhand der riesigen Menge an Trainingsbeispielen, die es statistisch verarbeitet hat, kann es in vielen Fällen jedoch ein halbwegs zufriedenstellendes Ergebnis produzieren.

Auf diese Weise hängen die Begriffe Artificial Intelligence und Big Data zusammen, denn Big Data ist sozusagen das Material einer AI-Struktur. Im Englischen setzte sich der Begriff des *data-driven system* durch, der insofern unglücklich gewählt ist, als er suggeriert, dass Daten selbst Akteure sind und etwas tun. Da von den Daten selbst tatsächlich überhaupt kein »drive« kommt,[5] verwende ich lieber den Begriff des datenbasierten algorithmischen Systems oder der datenbasierten Methode. Dennoch sind es selbstverständlich Menschen, die Daten produzieren und aufbereiten, und Menschen, die entscheiden, dass datenbasierte Methoden zur Wissensproduktion eingesetzt werden, die entsprechende statistische Modelle und mathematische Gütekriterien wählen und – das ist an dieser Stelle der springende Punkt – Menschen, die entscheiden, für welche Fragen datenbasierte Methoden die Antworten liefern sollen. Das sind Dinge, zu denen die Techniksoziologie sowie die Science and Technology Studies seit Jahrzehnten arbeiten.

Muster werden zu Regeln

Machine-Learning-Systeme funktionieren so, dass in den zugrundeliegenden Daten gefundene Muster zu Regeln für die zukünftige Ergebnisproduktion werden. Das bedeutet zweierlei: Erstens können diese Systeme nur Wissen produzieren, das die Vergangenheit betrifft, denn Daten können nur Vergangenes beschreiben. Prognostiziert man etwa Erdbeben, so wird auf die Zukunft geschlossen, indem Daten über vergangene Erdbeben verarbeitet werden. In vielen Naturzusammenhängen ist diese Vorgehensweise unproblematisch. Eine Erdbebenprognose hat keinerlei Auswirkungen auf das

5 Anders sehen das Machine-Learning-Enthusiasten wie etwa Ethem Alpaydin, der in seiner populärwissenschaftlichen Einführung schreibt: »Data starts to drive the operation; it is not the programmers anymore but the data itself that defines what to do next.« Ethem Alpaydin, *Machine Learning. The New AI*. Cambridge/Mass.: MIT Press 2016.

tatsächliche Erdbeben. Werden jedoch seismografische Modelle benutzt, um *hot spots* der Kriminalität vorherzusagen, wie es in den USA unter dem Stichwort des *predictive policing* getan wird, so ist das eine ganz andere Sache. Denn die aufgrund der Prognosen abgeleiteten Maßnahmen, wie beispielsweise verstärkte Polizeipräsenz, beeinflussen sehr wohl das Geschehen, so dass es in diesen Fällen zu selbsterfüllenden Prophezeiungen kommen kann.[6]

Zweitens basieren Machine-Learning-Systeme auf probabilistischen, statistischen Methoden. Betrachtet und analysiert wird dabei eine große Menge an Daten. Datenbasierte algorithmische Systeme »sehen« also in einem gewissen Sinn nur die Masse und nicht das Individuum. Sie sind epistemisch so angelegt, dass nur dann etwas als Muster und damit als Regel erkannt wird, wenn es in großen Mengen statistisch lokalisiert werden kann. Freilich kann sich das Wissen über die große Masse und damit die statistische Norm dann in einem *targeting* des Individuums bündeln. Beispielhaft etwa bei Methoden der *anomaly detection*, die gezielt nach statistischen Ausreißern suchen. Ein Beispiel aus den Niederlanden ist das System »Sy RI« (system risk indication), das beim Aufspüren von Betrug am Sozialstaat helfen sollte und mittlerweile aufgrund der Inkompatibilität mit Artikel 8 der Europäischen Menschenrechtskonvention, der das Recht auf Achtung des Privat- und Familienlebens schützt, gerichtlich abgeschafft wurde. Aufsehenerregend war dabei die dystopisch anmutende Tatsache, dass der Wasserverbrauch eines Haushalts als einer der Indikatoren für einen potentiellen Betrugsfall diente: Verbrauchte ein Haushalt, der Sozialhilfe beantragt hatte, viel mehr oder viel weniger Wasser als die große statistische Norm, wurde auf eine Falschangabe bei der Anzahl der Personen in diesem Haushalt und damit auf potentiellen Betrug beim Bezug der sozialstaatlichen Leistungen geschlossen.

Epistemische Grenzen

Machine-Learning-Systeme produzieren also vergangenheitsbezogene Aussagen über die große Masse. In vielen gegenwärtigen Anwendungszusammenhängen und vor allem in den Fällen, die stark kritisiert werden, werden sie jedoch für zukunftsbezogene Aussagen über Individuen eingesetzt. Das ist eine fundamentale Diskrepanz zwischen dem, was diese Systeme können,

6 Vgl. das Kapitel »This Is a Story About Nerds and Cops: PredPol and Algorithmic Policing« in: Jackie Wang, *Carceral Capitalism* (South Pasadena: Semiotext(e) 2018); Ruha Benjamin, *Race After Technology* (Cambridge: Polity 2019); Danielle Ensign/Sorelle A. Friedler u. a., *Runaway Feedback Loops in Predictive Policing*. In: *PMLR*, Nr. 81, 2018 (proceedings.mlr.press/v81/ensign18a.html).

und dem, was ihnen zugetraut wird. Eines der bekanntesten und am stärksten kritisierten Beispiele ist das COMPAS-System in den USA: Algorithmische Prognosen werden angewendet, um *risk assessments* durchzuführen, die vorhersagen sollen, wie wahrscheinlich es ist, dass ein Straftäter erneut straffällig wird.

Eine vieldiskutierte journalistische Arbeit zeigte durch statistische Analysen auf, dass die Fehler, die dieses System bei der Risikoklassifikation macht, einem *racial bias* unterliegen.[7] Schwarze Menschen werden systematisch häufiger – nämlich doppelt so oft – als »falsch positive«, weiße Menschen systematisch häufiger als »falsch negative« Fälle klassifiziert. Wird auf Grundlage dieser Prognosen über das Strafmaß, Bewährungsauflagen oder über andere Maßnahmen entschieden, fallen sie für schwarze Straftäterinnen systematisch fälschlicherweise drastischer aus. Das ist gerade in einem stark privatisierten Strafvollzugssystem mit seinen oft katastrophalen Lebensbedingungen in den Gefängnissen gravierend. Die Frage nach den genauen systematischen Gründen des Bias und der dadurch verursachten Fehlprognosen kann letztgültig nur durch eine Analyse des Systems selbst beantwortet werden, das aber nicht transparent gemacht wird.

Man kann jedoch Schlüsse ziehen, wenn man die epistemischen Grenzen dieses Systems betrachtet. Genau genommen prognostiziert ein derartiges algorithmisches System nämlich nicht, ob ein Mensch eine Straftat begehen wird. Die zugrundeliegenden Daten sind vorliegende Verhaftungs- und Verurteilungsdaten, also fragt dieses System präzise betrachtet nicht danach, ob eine Straftat begangen werden, sondern ob eine Verhaftung und Verurteilung aufgrund einer vermeintlichen Straftat stattfinden wird. Die Datenbasis enthält damit den aggregierten *racial bias* der Verhaftungskultur in den USA. In den Dateneinträgen, die abgefragt werden, wird zudem sichtbar, wie eng der hier zur Anwendung kommende Kriminalitätsbegriff mit Prekarität und Armut verknüpft ist:[8] In die Prognose der Gefährlichkeit eines Menschen fließt ein, ob er oder Familienmitglieder oder Freunde in der Vergangenheit jemals verhaftet (nicht etwa verurteilt) wurden, ob er einen Schulabschluss hat, welche Noten er in der Schule hatte, ob er arbeitslos ist, ob seine Eltern getrennt sind und wenn ja, wie alt er bei der Trennung der Eltern war, ob er oft umzieht, ob es in seiner Wohngegend Gang-Kriminali-

7 Julia Angwin / Jeff Larson u.a., *Machine Bias*. In: *Pro Publica* vom 23. Mai 2016 (www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing).

8 Im Rahmen einer künstlerischen Intervention erstellten Brian Clifton, Sam Lavigne und Francis Tseng demgegenüber eine *risk map* für *White Collar Crime Risk Zones* (whitecollar.thenewinquiry.com/).

tät gibt, ob er als Kind jemals von der Schule suspendiert wurde, ob jemand aus seiner Familie jemals Opfer einer Straftat war, und weitere Faktoren.

Es ist ein Teufelskreis mit selbstverstärkenden Effekten: Eine schwarze Person wird systematisch fälschlicherweise härter bestraft, weil aufgrund von *racial profiling* und der im System zum Einsatz kommenden Definition von Kriminalität in der Vergangenheit mehr schwarze Menschen verhaftet und verurteilt worden sind. Ein kollektives rassistisches Übel aus der Vergangenheit, das mit der konkreten Person nichts zu tun hat, produziert maßnahmenrelevante Konsequenzen für die nahe Zukunft dieser Person – etwas, das in einem Rechtsstaat nicht passieren darf.

Deutlicher, weil transparenter, ist der Fall eines algorithmischen Prognosesystems aus Österreich, der sich derzeit in einer rechtlichen Aushandlungssituation befindet. Das österreichische Arbeitsmarktservice (AMS), das der deutschen Arbeitsagentur entspricht, verkündete Ende 2018, dass ein algorithmisches Klassifizierungssystem bei der Beratung der Erwerbsarbeitslosen zum Einsatz kommen soll. Es soll deren Arbeitsmarktchancen prognostizieren und entsprechend eine Segmentierung der Menschen in drei Gruppen anleiten: diejenigen mit hohen, diejenigen mit mittleren und diejenigen mit niedrigen Chancen. Die verfügbaren Förderressourcen sind dann abhängig von der Gruppe, in die man einsortiert worden ist. Dabei soll das algorithmische System nicht automatisiert über die Gruppenzugehörigkeit entscheiden, sondern die Sachbearbeiterinnen bei der Entscheidung als *decision support*-Werkzeug unterstützen. Die Sachbearbeiter treffen dann die »letztgültige Entscheidung«.

Erwerbsarbeitslosen mit laut Prognose hohen Chancen werden weniger Ressourcen zugeteilt, da sie diese ohnehin nicht benötigten. Die Gruppe mit vermeintlich niedrigen Chancen soll an Betreuungsagenturen outgesourct werden. Die eingesparten Kapazitäten würden dann effizienter für die Erwerbsarbeitslosen in der mittleren Gruppe aufgewendet. Die Auslagerung stößt dabei selbst bei den dadurch entlasteten Sachbearbeiterinnen auf Kritik, nicht zuletzt weil die Erwerbsarbeitslosen mit niedrigen Chancen damit komplett aus dem Fokus des AMS geraten – in einer Evaluation der externen Betreuungsformate wurde von der Gefahr einer Einbahnstraße gesprochen.

Die Prognose der »Chancen« und damit die Gruppenzugehörigkeit beruht auf einer Schätzung der Wahrscheinlichkeit für die Erwerbsarbeitslosen, innerhalb eines festgelegten Zeitraums für einen wiederum festgelegten anderen Zeitraum in den Arbeitsmarkt integriert zu werden. Es werden dafür Daten, die dem AMS schon immer zur Verfügung standen, statistisch analysiert, und zwar entlang der Frage, welche Menschen mit welchen Dateneinträgen in der Vergangenheit wie schnell eine Arbeit gefunden haben.

Zuvor wurde geprüft, welche verfügbaren Arten von Dateneinträgen überhaupt einen statistisch signifikanten Einfluss auf diese Frage hatten, und nur diese flossen schließlich in das algorithmische Klassifikationssystem ein.

Mit Ergebnissen wie diesen: Ein weiblicher Geschlechtseintrag wirkt sich negativ auf die Chancen aus, ein männlicher demgegenüber positiv. Was das Alter betrifft, werden die Chancen ab dreißig schon schlechter, ab fünfzig sieht die Sache noch problematischer aus. Negativ: eine nichtösterreichische oder gar eine Nicht-EU-Staatsangehörigkeit; ebenso: gesundheitliche Beeinträchtigungen oder Betreuungspflichten – Letztere allerdings nur bei Frauen. Das klingt alles nicht gut, ist aber noch nicht diskriminierend, zumindest wenn man es entlang des epistemischen Fundaments solcher Systeme betrachtet. Es handelt sich schlicht um eine statistische Analyse vergangener, gruppenbezogener Daten. Diese zeigt, dass der Arbeitsmarkt Menschen mit bestimmten Dateneinträgen strukturell besser behandelt als andere.

Diese Vergangenheitsanalyse wird – wie bei allen datenbasierten Methoden – zur Prognose unter der Annahme und Voraussetzung, dass sich in der Zukunft die Dinge wie bisher verhalten. Diese Annahme ist in einer dynamischen Gesellschaft jedoch, vorsichtig gesagt, voraussetzungsreich – von einem pandemiegeprägten Arbeitsmarkt ganz zu schweigen. Die Konstruktion dieses algorithmischen Systems lässt sich überdies auch aus einer technischen Perspektive kritisieren, etwa mit Blick auf die abrupten Altersschwellen oder mögliche Interdependenzen zwischen den Variablen. Im Grunde aber wird jedes algorithmische System, das so konstruiert ist, ähnliche Ergebnisse produzieren. Es wird immer so sein, dass Individuen von Maßnahmen betroffen sind, die sich aus vergangenheitsbezogenen, kollektiven Prognosen ergeben.

Wird eine epistemisch gruppenbezogene Prognose auf ein Individuum angewendet und werden sozialstaatliche Ressourcen entlang dieser Prognosen verteilt – und zwar in einer Weise, die Menschen mit besonders schlechten Chancen aus Kostengründen outsourct –, so können diskriminierende Effekte nicht verhindert werden. Das liegt jedoch nicht an dem algorithmischen System selbst oder an einem »Fehler« in der Entwicklung, sondern an seinem spezifischen Anwendungszusammenhang und an der spezifischen Aufgabe, die dieses System dort zugewiesen bekommt, nämlich über die Ressourcenallokation in Bezug auf Individuen zu entscheiden.[9]

9 Paola Lopez, *Reinforcing Intersectional Inequality via the AMS Algorithm in Austria*. In: *Proceedings of the STS Conference Graz 2019* (https://diglib.tugraz.at/download.php?id=5e29a88e0e34f&location=browse).

»Entscheidungshilfe«

In der rechtlichen Auseinandersetzung um den »AMS-Algorithmus« spielt wie in vielen Kontexten des *decision support* durch algorithmische Systeme die Frage eine große Rolle, ob das algorithmische Ergebnis automatisch und routinemäßig durch die jeweiligen Anwenderinnen und Anwender übernommen wird. Die Datenschutzbehörde in Österreich untersagte im August 2020 die Anwendung des algorithmischen Systems per Bescheid. Argumentiert wurde unter anderem mit entsprechenden Bedenken – nicht zuletzt wegen der teilweise bloß zehnminütigen Beratungszeit pro Klientin. Das jedoch sei datenschutzrechtlich in dieser Form ohne zusätzliche Schutzmaßnahmen unzulässig.

Das österreichische Bundesverwaltungsgericht entschied im Dezember 2020 jedoch gegen dieses Verbot: Die Befürchtung, dass algorithmische Entscheidungen routinemäßig übernommen würden, sei eine bloße Behauptung und müsse erst nachgewiesen werden. Der Bescheid der Datenschutzbehörde wurde aufgehoben und der Einsatz des algorithmischen Systems erlaubt. Im Januar 2021 erhob die Datenschutzbehörde Revision, um diese Entscheidung zu bekämpfen. Es bleibt abzuwarten, wie der Verwaltungsgerichtshof als zuständiges Höchstgericht entscheiden wird.

Die Diskussion um *decision support* durch algorithmische Systeme findet sich an vielen Stellen wieder. Auch die strafrechtlichen *risk assessments* sollen nur eine Entscheidungshilfe darstellen, genauso wie das oben genannte *health care management*-System. Das Argument lautet stets, unter Zuhilfenahme objektiver mathematischer Systeme lasse sich mehr sehen als ohne sie: Sie stützen sich schließlich auf aggregiertes Big-Data-Wissen, können also aus mehr »Erfahrung« schöpfen als Individuen.

An dieser Stelle zeigt sich ein Widerspruch. Denn einerseits wird argumentiert, das System wisse mehr als die Sachbearbeiterinnen, da es mehr Daten zur Verfügung habe. Andererseits wird den Sachbearbeitern aber doch eine Art Meta-Intelligenz unterstellt, die ihnen zu entscheiden hilft, wann das System nichtdiskriminierend verwendet werden kann und wann man von ihm abweichen muss, um Diskriminierung oder schlicht falsche Ergebnisse zu vermeiden. Außerdem ist unklar, bis zu welchem Punkt der Einsatz als Entscheidungshilfe zulässig sein soll und ab wann von einer unzulässigen routinemäßigen Übernahme gesprochen werden kann: Werden die Entscheidungen des Systems zu null Prozent übernommen, dann kann das System auch abgeschafft werden. Da hinter dem System aber immerhin eine kostspielige Entwicklung steht, soll es so umfassend wie möglich zur Anwendung kommen. Würden die Entscheidungen jedoch zu hundert Prozent

übernommen, käme das einem automatisierten, also in dieser Form nicht zulässigen Entscheidungssystem gleich. Wird ein Schwellenprozentwert bestimmt, so wäre das eine systematische Vorgabe, die die stets betonte Entscheidungsfreiheit der Sachbearbeiter erst recht stark einschränken würde. Das sind Dinge, die bei der Entwicklung und spätestens bei der rechtlichen Auseinandersetzung um ein solches *decision support*-System jedenfalls zu bedenken sind.

Normative Kraft vs. emanzipatorisches Potential

Algorithmische Systeme operieren also mit Daten, die bis zu einem gewissen Vereinfachungsgrad die faktische Sachlage der zu untersuchenden Situation in bestimmten Aspekten abbilden sollen. Generieren sie (und sei es nur unterstützend) Entscheidungen, die Auswirkungen auf einzelne Menschen haben, so entsteht damit eine recht wörtliche Version von Georg Jellineks »normativer Kraft des Faktischen«. Weil die Bias- und Diskriminierungseffekte von AI-Systemen nicht »Bugs« sind, sondern gerade das »Feature«, also die logische Konsequenz der Big-Data-Methode, kann die Lösung nicht darin bestehen, die diskriminierungsrechtlich sensiblen Dateneinträge aus dem System zu entfernen.[10] Dadurch würden die Prognosen ihre Aussagekraft und damit ihre Treffsicherheit und Legitimierung verlieren – denn die Kategorie »Geschlecht« hat ja tatsächlich einen signifikanten Einfluss auf die Integration in den Arbeitsmarkt.

Das Problem besteht in der normativen Festschreibung des in statistischer Vereinfachung erfassten Status quo, der schon durchzogen ist von intersektionalen Ungleichheiten. In der Konsequenz werden diese festgeschrieben und verstärkt – mittels vermeintlich objektiver Prognosen. Ein automatisiertes *hiring tool* von Amazon, das mittlerweile abgeschafft wurde, beurteilte Frauen etwa systematisch als weniger geeignet als Männer. Da diesem Tool die Lebenslaufdaten der bisherigen Mitarbeiterinnen und Mitarbeiter zugrunde lagen, kann man den Umkehrschluss ziehen, dass die in den Daten reflektierte Unternehmenskultur systematisch Frauen benachteiligt hat.

10 In vielen algorithmischen Systemen werden diskriminierungsrechtlich geschützte Kategorien gar nicht erst explizit erfasst – durch das sogenannte *proxy*-Phänomen ist es oft dennoch möglich, aus stellvertretenden Merkmalen auf die geschützten Attribute zu schließen. In den USA ist beispielsweise die Postleitzahl ein *proxy* für den Dateneintrag *race*, und in Lebensläufen ist auch bei automatisierter Streichung des Geschlechtseintrags aus Einträgen wie *women's college* oder *women's lacrosse team* eine entsprechende Schlussfolgerung möglich.

Solange solche Systeme auf derartige Weise eingesetzt werden, sorgt das für eine Reproduktion von Ungleichheiten in drei Schritten: Die Realität gesellschaftlicher Benachteiligungen wird in einem ersten Schritt in den Daten erfasst – sei es durch verzerrenden Daten-Bias oder durch eine korrekte Abbildung der Ungleichheiten –, in einem zweiten Schritt als vermeintlich objektive Sachlage normativ verstärkt und schließlich in einem dritten Schritt mittels der resultierenden Maßnahmen – sei es ein ausbleibendes Jobangebot, eine nicht gewährte arbeitsmarktpolitische Ressource, eine längere Gefängnisstrafe, eine nicht geleistete medizinische Präventionsmaßnahme – in die soziale Sphäre zurückgespielt.

Das bedeutet wohlgemerkt nicht, dass algorithmische Systeme nicht auch emanzipatorisches Potential entfalten könnten. Dafür dürfte man sie allerdings nicht mehr, wie bisher, in prognostischer Absicht einsetzen. Man müsste sie vielmehr, ihrer mathematischen Architektur entsprechend, als höchst effiziente Diagnoseinstrumente begreifen. Schließlich kann das Wissen, das sie zur Verfügung stellen, Hinweise auf strukturelle Machtverhältnisse und materielle Ungleichheiten geben, die ohne Einsatz von Big Data unsichtbar blieben. Inwieweit das politisch opportun wäre, ist eine andere Frage.