Algorithmische Differenzierung und Diskriminierung aus Sicht der Menschenwürde

Carsten Orwat

1. Einleitung

Die Effizienz von Entscheidungen zur Differenzierung von Personen soll mit Anwendungen der Künstlichen Intelligenz (KI), algorithmischen Differenzierungen und automatisierten Entscheidungssystemen (AES) verbessert werden. Allerdings sind damit auch neue Risiken für die Grundrechte verbunden, einschließlich des Risikos von Diskriminierungen und Verletzungen der Menschenwürde. Das Antidiskriminierungsrecht soll nicht nur Gerechtigkeit und Gleichbehandlung sicherstellen, sondern auch die Möglichkeiten der freien Entfaltung der Persönlichkeit schaffen und die Menschenwürde schützen. Der Beitrag beleuchtet daher Entwicklungen der KI und von algorithmischen Differenzierungen aus der Perspektive der Menschenwürde.

Algorithmen werden zunehmend zur Unterstützung und automatischen Durchführung von Entscheidungen über die Differenzierung von Menschen eingesetzt, bei so unterschiedlichen Anwendungsfeldern wie z.B. Kreditvergabe, Wohnungssuche, Sozialleistungsbemessung oder Gerichtsentscheidungen. Solche Differenzierungen wirken sich dann auf die Verfügbarkeit und Verteilung von Produkten, Diensten, Positionen, Chancen, Vorteilen oder Belastungen aus, die essentiell für die Persönlichkeitsentfaltung und Realisierungen von Autonomie und Freiheit sind. Hier kommen dann Anwendungen zum Einsatz, bei denen maschinelles Lernen als Analysemethode bei Data Mining, Profiling oder Predictive Analytics angewandt wird und die Ergebnisse der Analysemethoden als Modelle bzw. Algorithmen in Entscheidungsprozessen eingesetzt werden. Dies geschieht entweder als Unterstützung und Beratung menschlicher Entscheidender oder als (voll-)automatisierte Entscheidungssysteme.

2. Algorithmische Differenzierung und Diskriminierung

2.1 Ursachen und Formen von algorithmischen Bias

Die Ursachen von Bias bzw. Verzerrungen bei der Verwendung von Algorithmen, insbesondere des maschinellen Lernens, sind vielfältig. Sie ergeben sich aus menschlichen Entscheidungen über die verwendete Datenbasis, die Entwicklung, den Einsatz oder die Anpassung von Algorithmen. Zu den am häufigsten genannten Ursachen von Bias gehören die mit historischen Ungleichheiten und Ungleichbehandlungen kontaminierte Trainingsdatensätze, die Auswahl ungeeigneter Messgrößen bzw. Label, ungeeignet entschiedene technische Trade-offs oder die Anwendung von Algorithmen in Bereichen, für die sie nicht trainiert bzw. optimiert wurden (z.B. Mehrabi et al. 2021, Pessach und Shmueli 2022).

Bias können dazu führen, dass Produkte und Dienste für verschiedene Bevölkerungsgruppen unterschiedlich gut funktionieren. Oder algorithmische Modelle, die bei Entscheidungen über die Differenzierung von Personen über Zugänge und Verteilung von Gütern, Positionen oder Freiheiten eingesetzt werden, führen zu Ungleichbehandlungen bei Betroffenen. Ein weiteres Problem sind Stereotypisierungen bei Generativer KI (wie z.B. Chat-GPT), vor allem wenn Entscheidungen auf Grundlage ihrer Ergebnisse getroffen werden (z.B. automatisiert erstellte Zusammenfassung von Bewerbungen oder Anträgen). Mittlerweile finden sich zahlreiche Beispiele für Bias und Diskriminierungsrisiken durch Algorithmen, von denen hier nur einige wenige angeführt werden können, etwa bei Gesichtsanalysen (Buolamwini und Gebru 2018), bei der Zuweisung zu bestimmten Therapien in der Medizin (Obermeyer et al. 2019), bei KI-basierten Analysen von Videointerviews im Personalbereich (Köchling et al. 2021) oder für Stereotypisierung bei der Sprachverarbeitung (Bender et al. 2021) (weitere Beispiele in AIAAIC Repository oder Orwat 2019, Kapitel 4).

2.2 Technische Lösungen und Restrisiken

Als Reaktion auf algorithmische Diskriminierungsrisiken werden in Wissenschaft, Forschung und Praxis große Mühen darauf gerichtet, Algorithmen diskriminierungsärmer bzw. »fairer« zu gestalten. Dabei wird vor allem versucht, Bias in Datensätzen zu beseitigen und Modelle zu optimieren. Ferner sind zahlreiche mathematische Fairness-Definitionen bzw. Fairness-

Metriken entwickelt worden, um Ungleichbehandlungen quantitativ auszudrücken und um damit Systeme zu optimieren und zu vergleichen (z.B. Pessach und Shmueli 2022). Allerdings sind diese in ihren Konsequenzen für gesellschaftliche Gerechtigkeit und Anti-Diskriminierung bisher noch wenig in der Öffentlichkeit diskutiert worden.

Des Weiteren sind die technischen Lösungen des »debiasing« in mehrfacher Weise begrenzt. So sind verschiedene Fairness-Metriken nicht gleichzeitig erfüllbar. Entwickelnde und Anbietende müssen verschiedene Trade-offs entscheiden, insbesondere ob und welche Fairness-Definitionen verwendet werden, welche Grenzwerte als Restrisiken für akzeptabel gehalten werden, ebenso sind Abwägungen zwischen einzelnen Zielen und Metriken zu treffen, wie z.B. zwischen Genauigkeit der Ermittlung der Differenzierungsziele und Diskriminierungsrisiken (z.B. Mehrabi et al. 2021, Pessach und Shmueli 2022).

Einige Fairness-Metriken sind Verhältniskennziffern, die aus den Fehlerraten »Falsch-Negative« und »Falsch-Positive« und Raten von richtig erkannten Klassifizierungen mit Bezug zu bestimmten Bevölkerungsmerkmalen gebildet werden. Bisher ist jedoch noch unklar, wie gesellschaftlich mit diesen Fairness-Metriken umgegangen wird, vor allem hinsichtlich der Niveaus an zu akzeptierenden Diskriminierungsrisiken für die Gesellschaft, etwa wer sie festlegt oder in welcher Höhe sie festgelegt werden. Nach Meinung der Europäischen Agentur für Menschenrechte kann, auch beim Vorliegen von Fairness-Metriken, nur von Fall zu Fall entschieden werden, wann eine ausreichend signifikante Diskriminierung vorliegt, nicht aber anhand der Bestimmung eines abstrakten Grenzwertes (FRA 2022, 25).

Im Entwurf der KI-Verordnung der Europäischen Union (2021/0106 (COM); KI-VO-E) wird zwar auf Parameter und Metriken verwiesen, aber nicht geklärt, wer akzeptable Fehlerraten und Grenzwerte festlegt. Zudem ermöglichen Formulierungen wie »Restrisiko« oder »akzeptabel« bzw. »vertretbar« (Artikel 9 (4) KI-VO-E) oder »angemessenes Maß an Genauigkeit« (Artikel 15 (1) KI-VO-E), dass Risikovermeidung mit Wirtschaftlichkeitsüberlegungen abgewogen werden kann. Dies ist vor dem Hintergrund zu sehen, dass für das Testen der KI-Systeme, die Risikovermeidung und andere regulatorische Pflichten Kosten entstehen. Der Entwurf der KI-VO lässt offen, ob die Europäische Kommission, Standardisierungsorganisationen, Entwickelnde, Anbietende, Anwendende oder Einrichtungen, die die »compliance« zertifizieren, die normativen Entscheidungen über Diskriminierungsniveaus und damit über die gesellschaftliche Realisierung von Gerechtigkeit tref-

fen. Dies kann dazu führen, dass Restrisiken von Diskriminierungen auf gesellschaftlich nicht festgelegtem Niveau wahrscheinlich sind.

Im Gegensatz zu menschlichen Entscheidenden, bei denen eher mit punktuellen Diskriminierungen zu rechnen ist, können einzelne AES und KI-Systeme große Reichweiten haben (z.B. durch Marktkonzentration oder bei Verwendung eines Systems in vielen Verwaltungen). Trotz scheinbar geringen Fehlerraten kann dies zu systematischen Diskriminierungen führen. Damit stellt sich die Frage, ob die Verminderung von Diskriminierungen oder Vorurteilen, die gegenüber menschlichen Entscheidenden mit Algorithmen angestrebt wird, nicht wieder durch den Reichweiteneffekt der Restrisiken aufgewogen oder sogar überwogen werden können.

2.3 Restrisiken und Antidiskriminierungsrecht

Die Restrisiken müssen im Lichte der Diskriminierungsverbote des Antidiskriminierungsrechts betrachtet werden. Zu rechtlich definierten Diskriminierungen führen Bias-Probleme, wenn bei Differenzierungen verzerrende Algorithmen verwendet werden und es dadurch zu ungerechtfertigten Ungleichbehandlungen durch die Nutzung von rechtlich geschützten Merkmalen (z.B. Geschlecht, ethnische Herkunft, Religion, Behinderung, Alter oder sexuelle Identität) oder scheinbar neutralen Merkmalen, Verfahren, Vorschriften oder Praktiken, die aber einen Zusammenhang zu den geschützten Merkmalen haben, kommt (Hacker 2018, von Ungern-Sternberg 2022).

Das Antidiskriminierungsrecht zeigt jedoch Schwächen im Umgang mit algorithmischen Diskriminierungen, denn angesichts algorithmischer, oft personalisierter oder individualisierter, Differenzierungen kann es für einzelne Betroffene schwierig sein, eine Ungleichbehandlung im Verhältnis zu anderen Betroffenen als solche wahrzunehmen und die rechtlich notwendigen ersten Nachweise für eine Schlechterstellung gegenüber vergleichbaren anderen Personen zu erbringen. Diese sind jedoch Voraussetzung dafür, dass rechtliche Verfahren eingeleitet werden können, auch wenn dann die einer Diskriminierung beschuldigte Person oder Einrichtung die Beweislast hat, dass sie nicht diskriminiert (Orwat 2019, 107–09 m.w.N., von Ungern-Sternberg 2022, 1141f.). Die ohnehin hohen Hürden für betroffene Individuen, die oft verhindern, dass es zu einem rechtlichen Diskriminierungsfall kommt, werden so noch weiter erhöht.

3. Verständnis der Menschenwürde im Verfassungsrecht

Die Menschenwürde gilt oft als ein abstraktes und unterschiedlich interpretierbares Konzept (z.B. Mahlmann 2008). Dies wird oft kritisch gegen ihre Verwendung eingewandt. Allerdings ist die Menschenwürde in vielen Menschen- und Grundrechtsdokumenten enthalten und durch Umsetzung in Rechtssystemen und der Rechtsprechung konkretisiert worden (z.B. McCrudden 2008, Mahlmann 2012). Im Folgenden wird daher Bezug auf die Entscheidungen des Bundesverfassungsgerichts (BVerfGE) genommen.

Danach ist die Menschenwürde ein »fundamentaler Wert- und Achtungsanspruch […], der jedem Menschen zukommt.« (BVerfGE 87, 209, Rn. 109). Sie umfasst vor allem die Wahrung der personalen Individualität, Identität und Integrität sowie die elementare Rechtsgleichheit (BVerfGE 144, 20, Leitsatz 3 a, Rn. 539). »Dem liegt eine Vorstellung vom Menschen zugrunde, die diesen als Person begreift, die in Freiheit über sich selbst bestimmen und ihr Schicksal eigenverantwortlich gestalten kann« (ebd., Rn. 539). Die Menschenwürde ist dem Menschen inhärent. »Jeder besitzt sie, ohne Rücksicht auf seine Eigenschaften, seine Leistungen und seinen sozialen Status.« (BVerfGE 87, 209, Rn. 107).

Zur weiteren Konkretisierung der Menschenwürde wurde die so genannte »Objektformel« entwickelt (Hong 2019, 672–90). Danach ist es mit der Menschenwürde unvereinbar, den Menschen »zum bloßen Objekt« staatlichen Handelns zu machen (BVerfGE 27, 1, S. 6 und weitere Entscheidungen). Nach der Objektformel darf der Mensch nicht wie eine Sache behandelt, verdinglicht oder zu einem bloßen Gegenstand herabgewürdigt werden (Hong 2019, 418f. m.w.N.). Das Bundesverfassungsgericht hat die Objektformel mit der Subjektformel weiterentwickelt, wonach es untersagt ist, einzelne Menschen einer Behandlung auszusetzen, die ihre Subjektqualität grundsätzlich in Frage stellt, indem sie die Achtung des Wertes vermissen lässt, der ihnen um ihrer selbst willen zukommt (nach Hong 2022, Rn. 26 m.w.N., ausführlich in Hong 2019, 421–28). Nach Höfling (2021, Rn. 16) ist zur Bestimmung einer Verletzung der Menschenwürde in konkreten Entscheidungssituationen zu fragen, ob der Subjektstatus eines Menschen trotz Verobjektivierung in Unterordnungs- und Abhängigkeitsverhältnissen durch Kompensationsmechanismen noch gesichert ist.

Bestimmte Formen von Diskriminierung stellen direkt eine Verletzung der Menschenwürde dar. Ein Verstoß gegen Art. 1 (1) Grundgesetz GG wird unter anderem bei einer unmittelbaren Diskriminierung durch Eingriffe in die je-

dem zustehenden Freiheitsgrundrechte wegen der in Art. 3 (3) GG genannten Kriterien gesehen (Herdegen 2022, Rn. 120). Höfling u.a. sehen eine verbotene Menschenwürdeverletzung vor allem in rassistischer Diskriminierung und ähnlichen demütigen Ungleichbehandlungen (Höfling 2021, Rn. 35) (siehe dazu vor allem auch BVerfGE 144, 20, Rn. 541). Hillgruber betont, dass eine Verletzung der Menschenwürde nicht nur vorliegt, wenn Menschen bestimmter »Rasse«, Hautfarbe, Religion oder Geschlecht als »minderwertig« angesehen werden, sondern auch bei Diskriminierungen von Menschen aufgrund einer körperlichen und geistigen Behinderung, insbesondere wenn eine Ausgrenzung wegen ihrer Behinderung droht (Hillgruber 2023, Rn. 17). Die genannten Merkmale sind besonders relevant, da sie zum einen unveränderliche und persönlichkeitskonstituierende Merkmale sind. Zum anderen sind sie historisch als Abgrenzung von den Gräueltaten des nationalsozialistischen Unrechtsregimes begründet (Lehner 2013, 226–48, Hong 2019, 407).

Eine weitere Konkretisierung der Menschenwürde erfolgt durch ihre Weiterentwicklung zum verfassungsrechtlichen allgemeinen Persönlichkeitsrecht, aus dem vor allem das zur weiteren Realisierung des Menschenwürdeschutzes entwickelte Recht auf informationelle Selbstbestimmung, die Diskriminierungsverbote, aber auch das Recht auf Selbstdarstellung (Britz 2007) für die nachfolgenden Betrachtungen relevant sind. Das Antidiskriminierungsrecht dient so nicht nur der Realisierung des Rechts auf Gleichbehandlung und sozialpolitischer Ziele, sondern auch des Rechts auf freie Entfaltung der Persönlichkeit und dem Schutz der Menschenwürde (ähnlich Baer 2009). Das Recht auf informationelle Selbstbestimmung und die Diskriminierungsverbote dienen u.a. dazu, unangemessene Fremdbilder der Persönlichkeit zu verhindern. Die Rechte sollen dazu befähigen, mitzuentscheiden, was Betroffene als zu ihrer Persönlichkeit gehörig und diese ausmachend ansehen können (Britz 2007). Kerngewährleistung des Persönlichkeitsrechts ist es, »Mechanismen zur Verfügung zu stellen, die den Einzelnen so in die Vorgänge der Konstituierung von Persönlichkeit einbinden, dass er seine Persönlichkeit als frei gewählt begreifen kann […]« (Britz 2008, 191).

In einer Reihe von Entscheidungen hat das Bundesverfassungsgericht das Recht auf Menschenwürde und das Persönlichkeitsrecht konkret auf die Risiken der modernen Informations- und Kommunikationstechnologien bezogen und entwickelt. In der »Mikrozensus«-Entscheidung (BVerfGE 27, 1) hat das Gericht deutlich gemacht, dass es der Menschenwürde widerspricht, den Menschen zu einem bloßen Objekt im Staat zu machen. Es ist mit der Menschenwürde unvereinbar, den Menschen in seiner gesamten Persönlich-

keit zwangsweise zu registrieren und zu katalogisieren und ihn damit wie eine Sache zu behandeln, die einer Bestandsaufnahme in jeder Beziehung zugänglich ist (ebd., S. 6–7).

In der Entscheidung zur Volkszählung (BVerfGE 65, 1) wurde das Grundrecht auf informationelle Selbstbestimmung begründet, das dem Recht auf freie Entfaltung der Persönlichkeit in Verbindung mit dem Recht auf Menschenwürde dient. Es garantiert die Befugnis, grundsätzlich selbst über die Preisgabe und Verwendung seiner persönlichen Daten bestimmen zu können (ebd., Leitsatz 1). Das Recht auf informationelle Selbstbestimmung dient nicht nur der Sicherung der äußeren und inneren Freiheitsdimensionen (Handlungsfreiheit und Identitätsbildung), sondern auch der Vermeidung von Abschreckungseffekten (chilling effects), die durch Unsicherheiten über die Datenverarbeitungen bei den Betroffenen entstehen können (Britz 2010).

In der Entscheidung zum großen Lauschangriff (BVerfGE 109, 279) erkennt das Gericht im Hinblick auf die Unantastbarkeit der Menschenwürde einen Kernbereich privater Lebensgestaltung an, der absoluten Schutz genießt. Was zum Kernbereich privater Lebensgestaltung gehört, hängt davon ab, ob der Sachverhalt einen inhaltlich höchstpersönlichen Charakter hat (ebd., Rn. 123). Ebenso verletzt eine zeitlich und räumliche »Rundumüberwachung« die Menschenwürde, wenn sie über einen längeren Zeitraum erfolgt und alle Bewegungen und Lebensäußerungen des Betroffenen nahezu lückenlos erfasst und zur Grundlage eines Persönlichkeitsprofils werden können (ebd., Rn. 150).

In der Entscheidung zum Recht auf Vergessen I (BVerfGE 152, 152) legt das Bundesverfassungsgericht das Recht auf informationelle Selbstbestimmung im privaten Bereich so aus, dass dem Einzelnen gewährleistet ist, »über der eigenen Person geltende Zuschreibungen selbst substantiell mitzuentscheiden.« (Ebd.). Das Gericht sah, dass in vielen Lebenssituationen private Unternehmen die grundlegenden Dienstleistungen erbringen, die eine entscheidende Rolle bei der öffentlichen Meinungsbildung, der Zuteilung oder Verweigerung von Chancen oder der Ermöglichung der Teilhabe am sozialen bzw. täglichen Leben spielen. In vielen Fällen geschehe dies auf der Grundlage einer umfassenden Datenerhebung und Verarbeitung durch oft marktmächtige Unternehmen, bei der die Offenlegung personenbezogener Daten in großem Umfang kaum zu vermeiden ist, will man nicht von den Produkten und Diensten ausgeschlossen werden (ebd., Rn. 85). In Fällen bei weitreichenden Abhängigkeiten oder Ausgesetztsein von ausweglosen Vertragsbedingungen (ebd., Rn. 85) oder »wenn private Unternehmen in eine staatsähnlich dominante Position rücken […] kann die Grundrechtsbindung

Privater einer Grundrechtsbindung des Staates im Ergebnis vielmehr nahe- oder auch gleichkommen« (ebd., Rn. 88).

Im Urteil zur automatisierten Datenanalyse bei der Polizeiarbeit (BVerfGE 1 BvR 1547/19 und 1 BvR 2634/20 vom 16.2.2023) betont das Gericht u.a., dass aus bereits bestehenden Datensätzen mit automatisierten Datenanalysen neues Wissen, vor allem persönlichkeitsrelevante Informationen, erzeugt werden kann (ebd., Rn. 68). Das Gericht verweist zudem auf die Diskriminierungsrisiken von automatisierten Datenanalysen, die umso weniger hinzunehmen sind, je mehr sich die Wirkungen der Analysen einer unzulässigen Benachteiligung nach Art. 3 (3) GG annähern (ebd., Rn. 76). Darüber hinaus betont es die Nachvollziehbarkeit der Algorithmen für den individuellen Rechtsschutz und die aufsichtliche Kontrolle, um Fehler erkennen und korrigieren zu können. Verlust der staatlichen Kontrolle wird vor allem bei Verwendung von lernfähigen Systemen bzw. KI gesehen, denn diese können im Verlauf des maschinellen Lernprozesses sich von der ursprünglichen menschlichen Programmierung lösen und deren Lernprozesse und Ergebnisse immer schwerer nachzuvollziehen sein (ebd., Rn. 99, mit Verweis auf das EuGH-Urteil Ligue des droits humains, C-817/19).

4. Faktoren der Verletzung der Menschenwürde bei algorithmischen Differenzierungen

4.1 Schwerwiegende und strukturelle Diskriminierung

Algorithmischen Differenzierungen können durch ihre Reichweite, die ganze Populationen umfassen, zu systematischen bzw. strukturellen Diskriminierungen führen. Einige Faktoren können auf eine Verletzung der Menschenwürde durch mögliche systematische algorithmische Diskriminierungen hindeuten, wenn nämlich algorithmische Diskriminierungen nach »Rasse« bzw. Ethnie, Geschlecht, körperlichen und geistigen Behinderungen und den weiteren geschützten Merkmalen des Art. 3 (3) GG erfolgen und wenn algorithmische Differenzierungen in Bereichen eingesetzt werden, bei denen eine starke Abhängigkeit von den Leistungen, Produkten oder Diensten besteht oder besonders Schutzbedürftige betrifft. Dies gilt insbesondere bei Entscheidungen, bei denen es darum geht, ein Leben in Würde zu leben (z.B. bei Empfangenden von Sozialleistungen). Dort kann es bei Diskriminierungen auch zu Formen der Demütigung oder Erniedrigung kommen, dadurch dass eine Behandlung

als Personen mit gleichem moralischem Wert vermissen gelassen wird (siehe z.B. die SyRi und Robo-debt Skandale) (ähnlich Teo 2023, 17).

In dieser Hinsicht sind auch algorithmische Differenzierungen problematisch, die durch negative Rückkopplungsschleifen (feedback loops) strukturelle Ungleichheiten festigen oder ausweiten. Solche negativen Rückkopplungsschleifen können entstehen, wenn Ergebnisse von AES und KI-Systeme, die das Verhalten von Betroffenen vorhersagen, wieder erfasst werden und die Systeme diese Daten unkorrigiert als Datengrundlage für weitere Datenanalysen, Schlussfolgerungen oder Weiterentwicklung bzw. »Lernen« der Algorithmen verwenden. Beispiele finden sich bei Systemen der vorausschauenden Polizeiarbeit (predicitive policing) (Lum und Isaac 2016, FRA 2022). Dies kann ebenso bei generativer KI geschehen, bei der Daten der Kommunikation mit den Nutzenden wieder ausgewertet werden. Darüber hinaus argumentieren Behrendt und Loh (2022), dass negative Rückkopplungsschleifen vor allem bei den ohnehin benachteiligten Bevölkerungsgruppen entstehen, da sie eher dazu tendieren, personenbezogene Daten preiszugeben, was u.a. an der Schwäche des Regulierungsinstruments der informierten Einwilligung (s.u.) und der nur scheinbaren Freiwilligkeit der Preisgabe liegt.

4.2 Generalisierung und fehlende Einzelfallgerechtigkeit

Algorithmische Differenzierung nimmt oft Formen der so genannten statistischen Diskriminierung an und verändert diese (Barocas und Selbst 2016, Binns et al. 2018, Orwat 2019). Statistische Diskriminierung ist eine Art Proxydiskriminierung. Statt mittels einer aufwendigen Einzelfallprüfung das tatsächliche Persönlichkeitsmerkmal bzw. Differenzierungsziel (z.B. das soziale Konstrukt »tatsächliche Fähigkeit, ein Flugzeug zu führen«) zu ermitteln, wird eine vergleichsweise einfach zu erhaltende Proxyinformation (z.B. Alter in Jahren) genutzt. Diese Form der Differenzierung soll ein Informationsdefizit effizient überwinden. Zu einer Diskriminierung kann es kommen, wenn die Proxyinformationen rechtlich geschützte Merkmale sind oder Merkmale enthalten, die eine Korrelation zu geschützten Merkmalen aufweisen (z.B. Hellman 2008, Britz 2008, Schauer 2018). Die Proxyinformationen können aus empirischen Untersuchungen abgeleitet werden oder, im Falle des maschinellen Lernens, auf an Daten trainierten Modellen vorliegen.

Statistische Diskriminierung und Generalisierung (sowohl durch menschliche Entscheidende oder mit Nutzung von Algorithmen) sind jedoch bereits an sich ethisch problematisch, weil Gruppeninformationen auf Individuen über-

tragen werden und so bei Entscheidungen als Quasi-Stereotypen und Vorurteilen bei der Entscheidungsfindung wirken (Gandy Jr. 2010, 34). Prinzipiell ist die Einzelfallgerechtigkeit nicht gewährleistet, da eine Einzelfallprüfung von Personen nicht vorgenommen wird und die individuellen Subjekteigenschaften und individuellen Situationen und Kontexte nicht berücksichtigt werden (Britz 2008). Was in der KI-Forschung und Praxis oft als »Prognose« bezeichnet wird, ist keine Prognose des potenziellen Verhaltens eines Individuums, die aus einer individuellen Prüfung abgeleitet wird. Vielmehr handelt es sich um eine Zuordnung von Personen zu statistisch oder mit maschinellen Lernverfahren gebildeten Kriterien, Kategorien, Scorewerten oder Rangfolgen, von denen bestimmte Ergebnisse in der Zukunft für die zugeordneten Personen erwartet werden.

In vielen Fällen der algorithmischen Differenzierung werden zudem die Kategorien, zu denen Individuen zugeordnet werden, anhand der Daten von Gruppen konstruiert, die nicht die Personen enthalten, über die tatsächlich entschieden wird (Eckhouse et al. 2019, 198f.). Auch sind die konstruierten Kategorien für die betroffenen Personen und Dritte in der Regel nicht nachvollziehbar. Anders als bei der Verwendung von klar kommunizierten Kriterien (z.B. Altersgrenzen in der Verwaltung) entziehen sich solche Entscheidungskriterien und -regeln der Überprüfung und Diskussion durch Wissenschaft und Öffentlichkeit, etwa ob überhaupt ein Kausalzusammenhang zwischen Kriterien und Differenzierungsziel besteht, ob er mit Fakten belegt werden kann oder ob die Verwendung bestimmter Kriterien sozialpolitisch oder moralisch umstritten oder unerwünscht ist.

4.3 Behandlung von Personen als Individuen und mit Achtung

Im Gegensatz zu individualisierten Entscheidungen über Personen werden bei statistischer Diskriminierung und Generalisierung Personen als Informationsobjekte und nicht als Individuen behandelt. Wann es moralisch problematisch ist, Menschen nicht als Individuum zu behandeln, sondern nur als Mitglied einer Gruppe (Stereotypisierung) ist umstritten (z.B. Lippert-Rasmussen 2011, Beeghly 2018).

Häufig wird an den moralischen Überlegungen von Kant angeknüpft, der mit dem Instrumentalisierungsverbot die Achtung des Menschen fordert (nach Hill Jr. 2014, 316f., Dillon 2022, Kapitel 2.2). So ist jeder gehalten, »die Würde der Menschheit an jedem anderen Menschen praktisch anzuerkennen, mithin ruht auf ihm eine Pflicht, die sich auf die jedem anderen Menschen

notwendig zu erzeigende Achtung bezieht.« (Kant 1797/1977, 601). Die Achtung von anderen Personen, die sich Menschen gegenseitig schulden und die Menschen gegenüber anderen gelten machen können, ist die Achtung ihrer Würde (Kant 1797/1977, 600, hier nach Schaber 2016, 256, Ulgen 2017, 2022, 14–15). Eine andere Person (und sich selbst) in seiner Würde achten bedeutet, dass ich andere »jederzeit zugleich als Zweck, niemals bloß als Mittel« (Kant 1786/2021, 429) behandle.

Nach Schaber, der dazu seine Erläuterungen zum falschen Versprechen interpretiert, meint Kant damit auch, dass man eine Person bloß als Mittel behandelt, wenn man sie in einer Weise behandelt, der sie unmöglich zustimmen kann. Das ist der Fall, »wenn sie dazu keinen Grund hat und sie sich nicht rational verhalten würde, wenn sie zustimmen würde.« (Schaber 2016, 256). Eine Person in ihrer Würde achten, bedeutet daher, sie in einer Weise zu behandeln, der sie vernünftigerweise zustimmen kann (ebd.). Bereits Korsgaard (1996), die sich ebenfalls auf Kant bezieht, argumentiert, dass man eine Person dann als bloßes Mittel behandelt, wenn man sie so behandelt, dass die Person der Art der Behandlung nicht zustimmen kann. Dies kann die Person weder bei Zwang noch bei Täuschung, da bei beiden Formen der Behandlung die Person keine Chance bekommen hat, den Zweck auszuwählen. Eine Behandlung ist demnach moralisch schlecht, wenn Personen nicht in der Lage sind zu wählen. Sie schließt daher, dass Zwang und Täuschung, nach der Formel der Menschlichkeit von Kant, die grundlegendsten Formen des Fehlverhaltens gegenüber anderen, die Wurzel allen Übels ist (ebd., S. 137–40). Eine Täuschung ist nach Schaber dann problematisch, wenn sie einen Bereich betrifft, der das Recht der Betroffenen, über wesentliche Teile des eigenen Lebens zu verfügen, beeinträchtigt (Schaber 2013, 134–36).

Ebenso unter Bezugnahme auf Kant entwickelt Ulgen (2022) Anforderungen für die Behandlung von Personen mit Respekt vor der ihnen inhärenten Würde im Hinblick auf KI und AES. Die Würde leitet sich aus ihrer Autonomie und ihren rationalen Fähigkeiten zur Ausübung von Vernunft, Urteilen und Entscheidungen ab. Die menschliche Autonomie ist geschützt, wenn Menschen in der Lage sind, unter dem Einfluss der Vernunft zu handeln, wenn sie die Beweggründe für ihr Handeln erkennen können, oder wenn sie ihre Beweggründe ändern können (ebd., S. 19). Diejenigen KI-Anwendungen und AES, die die Möglichkeiten zur Ausübung von Vernunft, Urteilsvermögen und Wahlmöglichkeiten einschränken, untergraben die Menschenwürde (ebd., S. 27). Neben der Relevanz für technologisch implementierte soziale Regeln

(s.u.) ist dies bereits für die algorithmenbasierte Entscheidungsfindung relevant.

Diese Argumentationen zeigen, wie bedeutend die Vorbedingungen für die Art und Weise der Behandlung sowie funktionsfähige Kompensationsmechanismen sind, um zu verhindern, dass Individuen als bloßes Objekt behandelt werden. Insbesondere gehören dazu die Möglichkeiten der Zustimmung zur Behandlung und der Einwirkung auf die Behandlung. Auch die Sicherung von Wahlmöglichkeiten und die Voraussetzung, über die Wahlmöglichkeiten so informiert zu werden, dass Personen selbstbestimmt handeln können, gehören zum Schutz der Menschenwürde.

Philosophische Erklärungsansätze dafür, wann eine Differenzierung moralisch falsch ist, ziehen vergleichbare Schlüsse. Zu ihnen zählen an Menschenwürde und Missachtung orientierte Ansätze der Diskriminierungstheorie (Khaitan 2015, 6–8), auch wenn nicht immer der Begriff »Würde« verwendet wird. Sie sehen eine Differenzierung als falsch an, wenn die diskriminierende Person den moralischen Wert der diskriminierten Person falsch, vor allem als niedriger, einschätzt oder wenn die diskriminierende Person eine falsche Einschätzung zum Ausdruck bringt, also so handelt, als ob die diskriminierte Person einen geringeren moralischen Wert hat (Thomsen 2017).

Baer (2009) sieht Gleichheitsbedürfnisse nicht allein durch Gleichheitsvorstellungen, sondern besser auch durch Bezugnahme auf Freiheit und Würde befriedigt, da sie als Schutzschild gegen kollektivistische Stereotypisierung dienen. Würde ist das Versprechen der Anerkennung unterschiedlicher Selbstwahrnehmungen, die alle den gleichen Respekt verdienen. Aus dem Zusammenwirken von Gleichheit, Freiheit und Würde erwächst nicht nur das Verständnis, dass Menschenwürde für alle Menschen gleich ist, unabhängig von Status, Klasse oder ähnlichem. Sondern Freiheit stellt auch sicher, dass jeder Einzelne sein eigenes Selbstverständnis definiert, anstatt es von Autoritäten bestimmen zu lassen (ebd., 460). Ebenso kann daraus ein Recht auf Entscheidung unter den Bedingungen der Chancengleichheit frei von Unterdrückung und Unterordnung (Gleichheit) konzipiert werden, unter Achtung und Anerkennung aller Beteiligten (Würde) (ebd., 466).

Hellman (2008) sieht das moralisch Falsche einer Diskriminierung darin, dass sie eine Person erniedrigt. Erniedrigende Regeln oder Praktiken drücken eine Missachtung der moralischen Gleichheit der von Diskriminierung betroffenen Personen aus. Eidelson (2013) fasst das Falsche an Diskriminierung als Fehler, eine Person korrekt als Individuum zu behandeln, auf. Der Fehler liegt darin, die Person nicht als (teilweise) Ergebnis ihrer vergangenen Bemühun-

gen der Selbstentfaltung (self-creation) zu sehen und als autonom Handelnde, deren künftige Entscheidungen sie selbst treffen kann (ebd., 227). So beantwortet er das Problem, dass Personen bei statistischer Diskriminierung und Generalisierung nicht als Individuum behandelt werden, mit der Forderung, dass Personen dann und nur dann als Individuum behandelt werden, wenn (1.) die differenzierende Person X den Nachweisen (evidence), wie die betroffene Person Y ihre Autonomie bei der Gestaltung ihres Lebens ausgeübt hat, ein angemessenes Gewicht beimisst, sofern diese Hinweise in angemessener Weise verfügbar und für die vorliegende Entscheidung relevant sind. (2.) Zusätzlich dürfen die Beurteilungen von X, wenn sie die Auswahlentscheidungen der Person Y betreffen, nicht in einer Art erfolgen, die die Fähigkeit von Y, diese Auswahlentscheidungen als autonom handelnde Person zu treffen, herabsetzen (Eidelson 2015, 144).

Zwar bleiben noch Fragen nach dem Ausmaß, Typen von oder Verpflichtungen zu angemessenen und relevanten Nachweisen offen, doch es kann abgeleitet werden, dass der Gegenstand der Informationen und Entscheidungen die selbstbestimmte Persönlichkeitsentfaltung der Betroffenen, ihre Möglichkeiten der Selbstwahrnehmung, Selbstbestimmung und Selbstdarstellung sein sollte. Allerdings muss ein Dilemma vermieden werden: Sollen möglichst viele personenbezogene Daten über die Selbstbestimmung erfasst werden, um das Problem der Generalisierung zu lösen und Personen als Individuum besser zu achten, kann dies nur mit der Kontrolle der dazu zu erstellenden Daten und Profile durch die Betroffenen selbst geschehen, um eine Verletzung des Rechts auf informationelle Selbstbestimmung zu vermeiden. Statt automatisierte Datenerfassungen und -analysen können auch menschliche Entscheidende erforderlich werden, um kontext-, situations- und personenbezogene Entscheidungen treffen zu können, die einen hohen Grad an situativem Abwägen mit Ermessensspielräumen erfordern.

4.4 Automatisierte Entscheidungen

Automatische Entscheidungen auf Basis von Generalisierungen sind nach Citron und Kaminski moralisch problematisch, da außer den Generalisierungsinformationen keine anderen Informationen über die Betroffenen verarbeitet werden. Werden Individuen bloß algorithmisch gebildeten Kategorien, Scorewerten oder Rangfolgen zugeordnet, werden sie nicht mehr als Individuen behandelt. Wenn es bei algorithmischen Entscheidungen den Individuen nicht mehr möglich ist, ihre Individualität zu verdeutlichen, dann verletzt das ih-

re Würde und verdinglicht sie anhand weniger Merkmale, anstatt sie als ganze Personen zu behandeln. Sowohl die Ausübung menschlichen Ermessens als auch individuelle Verfahrensrechte (des Einspruchs, der Korrektur etc.) sind nicht nur notwendig, um Fehler zu vermeiden, sondern auch, damit die Individualität angemessen anerkannt und respektiert werden kann (Citron 2008, 1304, Kaminski 2019, 1541–45). Des Weiteren sind menschliche Ermessensentscheidungen notwendig, wenn in Entscheidungen auch mildernde Umstände einbezogen werden müssen, die der Algorithmus nicht berücksichtigen kann, ebenso wenn unbestimmte Begriffe in den Entscheidungsregeln bestehen, die vom menschlichen Entscheidenden Abwägungen zwischen gegenläufigen Interessen erfordern (Citron 2008, 1304).

Eine der Beweggründe für die Regulierung der automatisierten Entscheidungsfindung ist der Schutz der Menschenwürde. Dies bezieht sich auf Artikel 22 der Allgemeinen Datenschutzgrundverordnung (Verordnung 2016/679 (DSGVO)) und den Vorgänger, Artikel 15 der Datenschutzrichtlinie (95/46/EG, 1995 (DSRL)). Nach Dammann und Simitis (1997, 218f.) sollte mit dem Verbot automatisierter Entscheidungen (Artikel 15 DSRL) verhindert werden, dass Betroffene bei Persönlichkeitsbeurteilungen nur computergestützt und auf der Grundlage gespeicherter Daten behandelt werden. Dies ignoriere die Individualität der Person und werte die Person zu einem bloßen Objekt von Computeroperationen ab (ähnlich zur DSGVO Martini 2021, Rn. 8, Scholz 2019, Rn. 3, ähnlich Jones 2017, Kaminski 2019).

Nach Martini und Nink wird die Subjektqualität eines Menschen allerdings nicht notwendig dadurch missachtet, dass allein personenbezogene Daten das Objekt einer algorithmischen Analyse sind. Die Subjektqualität wird bei automatisierten (Verwaltungs-)Entscheidungen erst dann tangiert, wenn algorithmische Verfahren den Betroffenen nachteilige Folgen aufbürdet, »ohne ihm die Chance zu eröffnen, sich gegen die Entscheidung in angemessener Weise zur Wehr setzen zu können.« (Martini und Nink 2017, 7). Zur Wahrung der informationellen Selbstbestimmung setzt man in der Rechtspraxis auf das Verfahren, (1) über die automatisierte Entscheidung zu informieren, (2) auf Anfrage die wesentlichen Entscheidungsgründe mitzuteilen und zu erläutern, (3) den eigenen Standpunkt geltend machen zu können, um erforderlichenfalls eine Überprüfung und Neubewertung zu erreichen (ebd., S. 7).

Allerdings lassen einige Defizite der Regulierung von automatisierten Entscheidungen Zweifel aufkommen, ob sie dem Schutz der Menschenwürde noch dienen kann. Das so genannte »Verbot« ist mit umfangreichen Ausnah-

men versehen, insbesondere wenn ein AES zum Abschluss oder der Erfüllung eines Vertrags dient, durch Rechtsvorschriften Zulässigkeit verlangt oder eine ausdrückliche Einwilligung vorliegt. Die Regulierung sieht zwar vor, dass der Betreibenden einer automatisierten Entscheidung über das Bestehen einer automatischen Entscheidung und die so genannte involvierte Logik informieren muss, aber es ist noch unklar, welche Inhalte diese Informationspflicht hat, z.B. ob und wie über Entscheidungskriterien oder mögliche Diskriminierungsrisiken informiert werden muss (Orwat 2019, 114–23 m.w.N.). Zudem wird das »Verbot« oft nur als Eingriffsrecht der Betroffenen in begründeten Einzelfällen interpretiert (Martini und Nink 2017, 4). Dabei müssen die Betroffenen zunächst Kenntnis von dem AES und deren Auswirkungen haben und eine Begründung des Verlangens nach einem Eingreifen eines Menschen und nach Erklärung der involvierten Logik erbringen. Da dies sehr aufwendig sein kann, können Abschreckungseffekte (chilling effects) entstehen, wenn einzelne Personen es als unzumutbare hohe Hürden wahrnehmen, die Regelung in Anspruch zu nehmen.

4.5 Entstehung neuen Wissens sowie umfassender und aussagekräftiger Personen- und Gruppenprofile

Die Möglichkeiten der Datenaggregation, der Wiederverwendung von Daten, der Datenkombination und daraus abgeleitete Schlussfolgerungen, der De-Anonymisierung und der Re-Identifizierung von Personen, der Kategorisierung, Einstufung, Beurteilung und des Individual- oder Gruppen-Profiling von Personen sind mit KI stark angewachsen (z.B. Yeung 2019, FRA 2020, Smuha 2021). Einige KI-Systeme wurden dafür entwickelt, automatisierte Rückschlüsse auf die Identität, persönlichkeitskonstituierenden Merkmale und andere sensible Sachverhalte, wie Emotionen, Charaktereigenschaften, psychische Zustände oder politische Orientierungen zu ziehen (Beispiele in Kosinski 2021, Matz et al. 2023). Die KI-basierten biometrischen und psychometrischen Auswertungen (z.B. emotional AI) können der gezielten Ansprache (z.B. im Marketing), der Risikobeurteilung (z.B. bei der Bewerberselektion, Berechnung der Wahrscheinlichkeit des Studienabbruchs oder Kreditausfalls) und Verhaltenssteuerung dienen (kritisch z.B. Valcke, Clifford, und Dessers 2021). Oft basieren die Systeme auf einer Reduktion der Persönlichkeit auf quantifizierbare Messgrößen und Klassen, die versuchen, die für ein Differenzierungsziel relevanten Persönlichkeitseigenschaften abzubilden. Kritisch wird dazu die Standardisierung von Persönlichkeit (Köchling et al.

2021) oder die pseudowissenschaftliche Herangehensweise (Sloane, Moss, und Chowdhury 2022) angeführt.

Auch wenn Umfang und Arten der Anwendung solche Systeme in der Praxis noch wenig bekannt sind, verdeutlicht dies, dass mit KI aussagekräftige Personenprofile bzw. »Rundum«-Profile erzeugt und verwendet werden können, die geeignet sind, ein (nahezu) vollständiges Fremdbild einer Person überzustülpen, dies mit persönlichkeitskonstituierenden Merkmalen und dies auch ohne dass eine valide Zustimmung durch die Betroffenen vorliegt.

Insgesamt kommt es zu einer weiteren Ablösung der Datenrepräsentation durch die Betreibenden von den Möglichkeiten der Kontrolle der Selbstdarstellung durch die Betroffenen (vgl. auch Teo 2023). Die Identität der Betroffenen ist dann nur noch fremdbestimmt, auch wie sie sich begreifen (als normal, gesund, regelkonform etc.). Die Fähigkeiten, sich selbstbestimmt zu entfalten, sind dann eingeschränkt oder sogar beseitigt. Es liegt nach den oben skizzierten Maßstäben eine Verletzung der informationellen Selbstbestimmung und der Menschenwürde vor. Auch können derartige KI-Anwendungen Abschreckungseffekte und damit Selbsteinschränkungen der freien Persönlichkeitsentfaltung als Form der Würdeverletzung hervorrufen (FRA 2019, 20).

4.6 Strukturelle Überlegenheit

Im Verhältnis Staat gegenüber Betroffenen (z.B. Bürgern oder Migranten) muss man grundsätzlich von der strukturellen Überlegenheit des Staates ausgehen, denn es liegen üblicherweise Situationen mit Gewaltmonopol, fehlende Ausweichmöglichkeiten, Ausgeliefertsein, Nichtverhandelbarkeit und vollständiger, einseitig festgelegter Verbindlichkeit der Regeln vor. Eine Reihe von Faktoren können auch in privaten Verhältnissen die strukturelle Überlegenheit von Anwendenden der KI-Systeme (z.B. Anbietende, Arbeitgebende, Banken) gegenüber den Betroffenen (z.B. Kunden, Bewerbende, Kreditsuchenden) erhöhen.

Erstens werden sowohl im staatlichen als auch im privaten Bereich zunehmend gesellschaftliche Regeln in Software bzw. Algorithmen gefasst. Für die technische Implementierung müssen die Regeln in Programmiersprache umgesetzt werden oder werden durch maschinelles Lernen erzeugt. Bei diesen Formen der Präzisierung gehen aber auch Auslegungs- und Ermessensspielräume verloren, die oft notwendig sind, damit gesellschaftliche Regeln auf viele Situationen, die teils nicht vorhersehbar sind, angewandt werden können. Werden Regeln vollständig automatisiert durchgesetzt, wie z.B. bei

vollautomatisierten Entscheidungen, wird ein Abweichen von den Regeln technisch verhindert. Aber auch bei algorithmischen Auswahlarchitekturen (choice architectures oder nudging) wird der Raum der Auswahlmöglichkeiten technisch vorgegeben und oft verengt. Die Regeln werden dann zumeist einseitig von den Entwickelnden und Anwendenden vorgegeben, Verhandlungs-, Einwirkungs-, Korrekturmöglichkeiten der Betroffenen reduziert oder beseitigt, wodurch sich die strukturelle Überlegenheit erhöhen kann. Dadurch verringern sich auch die Handlungsmöglichkeiten, die Autonomie und die Chancen der Selbstdurchsetzung von Autonomie durch die Betroffenen (Ulgen 2022, Teo 2023, 27–31, Deutscher Ethikrat 2023, 120–37).

Zweitens, wenn die privaten Anwendenden von KI-Systemen gleichzeitig auch diejenigen sind, die Plattformen betreiben (die teilweise auch die Entwickelnden von KI-Systemen sind), können starke Netzwerkeffekte der Plattformen zu verminderten Ausweichmöglichkeiten und stärkeren Abhängigkeiten führen.

Drittens sind, wie eben gezeigt, einige KI-Systeme in der Lage, sensible Persönlichkeitsmerkmale, wie psychische Zustände, Charaktereigenschaften, Emotionen, auch aus scheinbar »belanglosen« Daten, wie der Kommunikation in sozialen Netzen, zu ermitteln. Dadurch kann die Angewiesenheit auf ein Produkt, Dienst oder Position besser ermittelt und ausgenutzt werden (z.B. Härtel 2019), ebenso menschliche Schwächen, insbesondere wenn die Systeme für »Dark Pattern« oder anderen Formen der Manipulation verwendet werden (z.B. Ulgen 2022, 22–24).

Entwicklungen, die die strukturelle Überlegenheit der Anwendenden von KI-Systemen gegenüber den Betroffenen erhöhen, können tendenziell die Menschenwürde und die freie Persönlichkeitsentfaltung gefährden, denn den Betroffenen werden die Möglichkeit der selbstbestimmten Lebensgestaltung eingeschränkt. Dies kann in privaten Verhältnissen auf dem Wege der Störung der Vertragsparität und damit der Verringerung der Möglichkeiten, dass Betroffene ihre Autonomie selbst durchsetzen können, geschehen. Denn auch wenn die Vertragsfreiheit gilt, d.h. dass jeder die Freiheit hat, zu bestimmen, mit wem und unter welchen Bedingungen Verträge abgeschlossen werden, so muss gesichert sein, dass dabei »auch die Bedingungen freier Selbstbestimmung tatsächlich gegeben sind.« (BVerfGE 82, 242, S. 255).

4.7 Fehlende Möglichkeit der validen Zustimmung

Das Instrument der Zustimmung kann moralisch unzulässige Behandlungen in zulässige transformieren. Allerdings kann sie diese moralische Transformationsleistung nur bei Einhaltung bestimmter Voraussetzungen erreichen. Dazu gehört, dass die Betroffenen freiwillig zustimmen und dazu Wahlmöglichkeiten haben, dass sie ausreichend informiert sind, also den Umfang der Datenverarbeitungen und die Konsequenzen daraus verstehen, und dass sie die notwendigen Entscheidungsbefähigungen haben (Bullock 2018). Zum anderen wird aus philosophischer Sicht auch bezweifelt, dass Zustimmung eine Behandlung als bloßes Objekt in eine moralisch zulässige Behandlung transformieren kann, wenn bereits die Behandlung die Pflicht, andere Menschen mit Achtung zu behandeln, verletzt (Fahmy 2023). Das dürfte etwa bei schwerwiegenden algorithmischen Diskriminierungen oder bei Differenzierungen, die auf Profilen mit (nahezu) vollständiger Erfassung und Fremdbestimmung der Persönlichkeit basieren, der Fall sein.

In der Datenschutzpraxis sind die Probleme der informierten Einwilligung seit langem bekannt. Die Wirksamkeit und Aussagekraft wird zunehmend eingeschränkt durch nicht verhandelbare, lange, unverständliche und in juristischer Sprache formulierte Datenschutzerklärungen, die zunehmende Erhebung von Daten auf der Grundlage sogenannter berechtigter Interessen ohne das einer Einwilligung erforderlich ist (Artikel 6 (1) f DSGVO), starke technisch-ökonomische Netzwerkeffekte und dadurch Bindungen von Kunden und Nutzern an Systeme oder Plattformen (abnehmende Freiwilligkeit) sowie durch Schnittstellendesigns, die zu einer Einwilligung in Datenerhebungen verleiten. Den Betroffenen fehlt oft die Kenntnis über die Notwendigkeit und die rechtlichen Möglichkeiten der informierten Einwilligung selbst. Zudem können sie kaum abschätzen, welche tatsächlichen Folgen die Einwilligung im Hinblick auf potenziell nachteilige, teilweise zeitlich weit entfernte Behandlungen, die auch auf Basis schwer nachvollziehbarer Datenakkumulation oder nicht mehr abschätzbarer Weiterverarbeitungen und Datenweitergaben entstehen können, hat. Darüber hinaus können bei komplexen KI-Algorithmen die Entscheidungskriterien unverständlich oder unbekannt sein, insbesondere wenn es sich um selbstlernende bzw. adaptive Systeme handelt. Ebenso kann mit KI-basierten Schlussfolgerungen neues Wissen aus vorhandenen personenbezogenen Daten generiert werden, auch aus anonymisierten Daten und auch für Personen oder Gruppen, die nicht an der ursprünglichen Einwilligung beteiligt waren (s.o.). Es ist dann davon auszugehen, dass die

Betroffenen nicht mehr hinreichend erkennen können, in was sie einwilligen (z.B. Orwat 2019, 106f. m.w.N.). Auf Grund dieser Faktoren wird die informierte Einwilligung als Legitimation der Behandlung von Menschen als bloße Objekte und als Instrument der Selbstbestimmung zunehmend unbrauchbar.

Zarsky konkretisiert noch detaillierter, dass zum Schutz der Menschenwürde ein Verständnis der inneren Abläufe von automatisierten Datenanalysen vorliegen müsse, denn ohne dieses Verständnis können die Ergebnisse immer noch willkürlich und falsch erscheinen. Das Problem ist, dass mit den bestehenden (rechtlichen) Transparenzanforderungen maximal lediglich Informationen über die Korrelationen bzw. Klassifikationen, in denen eine Person eventuell fallen könnte, geliefert werden. Stattdessen muss der automatisierte Prognoseprozess auch interpretierbar sein, d.h. der Auswahlprozess muss erklärbar sein. Daher erfordere der Schutz der Menschenwürde sogar, dass für die Betroffenen kausale Zusammenhänge und nicht bloß Korrelationen feststellbar sein müssten, bevor Schlussfolgerungen und Maßnahmen getroffen werden (Zarsky 2013).

5. Zusammenfassung und Schlussfolgerung

Die Menschenwürdeperspektive ermöglicht es zu bestimmen, wie Menschen oder Maschinen andere Menschen behandeln sollten. Diese Perspektive kann die Arbeiten zur relativen Fairness von Algorithmen und der Verminderung von Diskriminierungsrisiken ergänzen. Die Menschenwürdeperspektive geht über das Bemühen des »debiasing« von Datensätzen und Algorithmen hinaus und fragt, wie algorithmenbasierte Entscheidungsprozesse gestaltet sein sollten und welche Informationsgrundlagen dafür zur Verfügung stehen sollten. Sie weist darauf hin, dass selbst der Idealfall »genauer« Profile als Grundlage algorithmenbasierter Entscheidungen problematisch ist, wenn übermächtige Fremdbilder die informationelle Selbstbestimmung der Betroffenen unterdrücken. Sie liefert auch Begründungen dafür, in welchen Situationen KI und AES nicht eingesetzt werden sollten, weil eine Einschränkung oder Verletzung der Menschenwürde nicht ausgeschlossen werden kann. So sind die Verbote bestimmter KI-Anwendungen im Entwurf der KI-Verordnung auch mit Verweis auf den Schutz der Menschenwürde begründet worden (z.B. in Erwägung 15 und 17). Das Verständnis von Diskriminierung auch als Verletzung der Würde und der moralischen Gleichwertigkeit der Betroffenen ergänzt das Verständnis von Diskriminierung als Schädigungen von Gerechtigkeitsvor-

stellungen oder von sozialpolitischen Zielen. Es gibt Aufschluss darüber, was es bedeutet, Menschen als Individuen und mit Respekt zu behandeln.

Durch den Einsatz von KI und AES kann es zur Verletzung der Menschenwürde kommen. Wie gezeigt, kann dies mit der Überlagerung von problematischen Faktoren bei der Anwendung von KI und AES und die zunehmende Ungeeignetheit von rechtlich etablierten Kompensationsmechanismen zur Abmilderung der Behandlung als bloßem Objekt geschehen. Diese Faktoren umfassen (1.) die Generalisierung und Missachtung der Persönlichkeit in Entscheidungen der Ungleichbehandlung, (2.) die Reichweite von Systemen mit Restrisiken der systematischen und strukturellen Diskriminierung, einschließlich dem Umstand, einige Personen einem höheren Diskriminierungsrisiko auszusetzen und sie wie Personen mit geringerem moralischen Wert zu behandeln, (3.) die immer unzureichender werdende informierte Einwilligung, die sich bei KI-Systemen, deren Entscheidungskriterien und Auswirkungen nicht mehr nachvollziehbar sind, besonders drastisch auswirkt und Betroffene nicht mehr auf die Ergebnisse einwirken können, (4.) die unzureichende Klärung der Regulierung von automatisierten Entscheidungen, der Rolle involvierter menschlicher Entscheidender sowie der informierten Einwilligung dabei, (5.) der Verlust der Kontrolle über die Erzeugung und Verwendung von Persönlichkeitsbildern durch die Betroffenen sowie (6.) die steigende strukturelle Überlegenheit des Staates oder privater Unternehmen durch die zunehmende technische Durchsetzung von gesellschaftlichen Regeln, die Marktkonzentration, die besonderen Fähigkeiten von KI-Systemen Abhängigkeiten und andere menschliche Schwächen zu erkennen und sie auszunutzen und dadurch Situationen mit eingeschränkten Handlungs- und Einwirkungsmöglichkeiten, Unausweichlichkeit und starken Abhängigkeiten. Als Folge kann in den Situationen, in denen Faktoren allein oder zusammenwirken, eine Garantie des Schutzes der Menschenwürde nicht mehr vorliegen. Dies wiegt besonders schwer, wenn es sich um Differenzierungen von Produkten, Diensten, Positionen handelt, die für die selbstbestimmte Lebensgestaltung und Identitätsbildung oder für ein menschenwürdiges Dasein von Menschen mit besonderen Bedürfnissen und Vulnerabilitäten essentiell sind.

Es besteht eine dringende Notwendigkeit, weiter zu klären, wann die Menschenwürde und Persönlichkeitsentfaltung konkret eingeschränkt oder verletzt ist und wie sie zu schützen sind, insbesondere (1.) welche Personen- und Gruppenprofile so umfassend oder so persönlichkeitskonstituierend sind, dass man das Persönlichkeitsbild als fremdbestimmt und den Kernbereich

der privaten Lebensgestaltung als ausgehöhlt bezeichnen muss, (2.) unter welchen Bedingungen schwerwiegende, systematische oder strukturelle Diskriminierungen mit Verwendung von Algorithmen vorliegen, (3.) wie weit und in welcher Form die Persönlichkeit der Betroffenen bei algorithmenbasierten Entscheidungen zu respektieren ist und welche Form von Rechtfertigung für Entscheidungen Betroffene erhalten müssen, (4.) welche Einwirkungsmöglichkeiten auf Entscheidungen und auf ihr Persönlichkeitsbild die Betroffenen haben müssen und wie kommunikative Prozesse dazu aussehen sollten sowie (5.) wie nicht nur die Menschenwürde und Persönlichkeitsentfaltung der direkt Betroffenen (wieder) gestärkt werden kann, sondern auch der Schutz von mitbetroffenen Dritten, die nicht wissen, dass sie betroffen sind.

Anmerkung und Danksagung

Eine ähnliche Version dieses Beitrags soll parallel in englischer Sprache erscheinen. Für wertvolle Hinweise bei der Erarbeitung dieses Beitrags möchte ich meinen Kollegen Reinhard Heil und Philipp Frey danken.

Literatur

Baer, Susanne. 2009. »Dignity, liberty, equality: A fundamental rights triangle of constitutionalism.« *University of Toronto Law Journal* 59 (4): 417–68.

Barocas, Solon and Andrew D. Selbst. 2016. »Big Data's Disparate Impact.« *California Law Review* 104 (3): 671–732.

Beeghly, Erin. 2018. »Failing to treat persons as individuals.« *Ergo: An Open Access Journal of Philosophy* 5 (26): 687–711.

Behrendt, Hauke and Wulf Loh. 2022. »Informed consent and algorithmic discrimination – is giving away your data the new vulnerable?« *Review of Social Economy* 80 (1): 58–84.

Bender, Emily M., Timnit Gebru, Angelica McMillan-Major and Shmargaret Shmitchell. 2021. »On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?« *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.

Binns, Reuben, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao and Nigel Shadbolt. 2018. »›It's Reducing a Human Being to a Percentage‹: Percep-

tions of Justice in Algorithmic Decisions.« *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*.

Britz, Gabriele. 2007. *Freie Entfaltung durch Selbstdarstellung. Eine Rekonstruktion des allgemeinen Persönlichkeitsrechts aus Art. 2 I GG*. Tübingen: Mohr Siebeck.

Britz, Gabriele. 2008. *Einzelfallgerechtigkeit versus Generalisierung. Verfassungsrechtliche Grenzen statistischer Diskriminierung*. Tübingen: Mohr Siebeck.

Britz, Gabriele. 2010. »Informationelle Selbstbestimmung zwischen rechtswissenschaftlicher Grundsatzkritik und Beharren des Bundesverfassungsgerichts.« In *Offene Rechtswissenschaft*, edited by Wolfgang Hoffmann-Riem, 561–96. Tübingen: Mohr Siebeck.

Bullock, Emma C. 2018. »Valid consent.« In *The Routledge Handbook of the Ethics of Consent*, edited by Peter Schaber and Andreas Müller, 85–94. Routledge.

Buolamwini, Joy and Timnit Gebru. 2018. »Gender shades: Intersectional accuracy disparities in commercial gender classification.« *Conference on Fairness, Accountability and Transparency*.

Citron, Danielle K. 2008. »Technological Due Process.« *Washington University Law Review* 85 (6): 1249–313.

Dammann, Ulrich and Spiros Simitis. 1997. *EG-Datenschutzrichtlinie: Kommentar*. Baden-Baden: Nomos.

Deutscher Ethikrat. 2023. *Mensch und Maschine – Herausforderungen durch Künstliche Intelligenz*. Deutscher Ethikrat. Berlin.

Dillon, Robin S. 2022. »Respect.« In *The Stanford Encyclopedia of Philosophy* (Fall 2022 Edition), edited by Edward N. Zalta and Uri Nodelman. Metaphysics Research Lab, Stanford University.

Eckhouse, Laurel, Kristian Lum, Cynthia Conti-Cook and Julie Ciccolini. 2019. »Layers of Bias: A Unified Approach for Understanding Problems With Risk Assessment.« *Criminal Justice and Behavior* 46 (2): 185–209.

Eidelson, Benjamin. 2013. »Treating People as Individuals.« In *Philosophical Foundations of Discrimination Law*, edited by Deborah Hellman and Sophia Moreau, 203–27. Oxford: Oxford University Press.

Eidelson, Benjamin. 2015. *Discrimination and Disrespect*. Oxford: Oxford University Press.

Fahmy, Melissa S. 2023. »Never Merely as a Means: Rethinking the Role and Relevance of Consent.« *Kantian Review* 28 (1): 41–62.

FRA. 2019. *Facial recognition technology: fundamental rights considerations in the context of law enforcement*. European Union Agency for Fundamental Rights (FRA). Luxembourg: Publications Office of the European Union.

FRA. 2020. *Getting the Future Right – Artificial Intelligence and Fundamental Rights*. European Union Agency for Fundamental Rights (FRA). Luxembourg: Publications Office of the European Union.

FRA. 2022. *Bias in Algorithms – Artificial Intelligence and Discrimination*. European Union Agency for Fundamental Rights (FRA). Luxembourg: Publications Office of the European Union.

Gandy Jr., Oscar H. 2010. »Engaging rational discrimination: exploring reasons for placing regulatory constraints on decision support systems.« *Ethics and Information Technology* 12 (1): 1–14.

Hacker, Philipp. 2018. »Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies Against Algorithmic Discrimination Under EU Law.« *Common Market Law Review* 55 (4): 1143–85.

Härtel, Ines. 2019. »Digitalisierung im Lichte des Verfassungsrechts – Algorithmen, Predictive Policing, autonomes Fahren.« *Landes- und Kommunalverwaltung* 29 (2): 49–60.

Hellman, Deborah. 2008. *When is Discrimination Wrong?* Cambridge, London: Harvard University Press.

Herdegen, Matthias. 2022. »Art. 1 Abs. GG.« In *Grundgesetz Kommentar*, edited by Theodor Maunz and Günther Dürig. München: Beck.

Hill Jr., Thomas E. 2014. »In Defence of Human Dignity: Comments on Kant and Rosen.« In *Understanding Human Dignity*, edited by Christopher McCrudden, 313–25. Oxford: Oxford University Press.

Hillgruber, Christian. 2023. »GG Art. 1 Schutz der Menschenwürde.« In *Beck Online-Kommentar Grundgesetz*, edited by Volker Epping and Christian Hillgruber. München.

Höfling, Wolfram. 2021. »Art. 1 GG Schutz der Menschenwürde, Menschenrechte, Grundrechtsbindung.« In *Grundgesetz: Kommentar*, edited by Michael Sachs, 70–102. München: Beck.

Hong, Mathias. 2019. *Der Menschenwürdegehalt der Grundrechte. Grundfragen, Entstehung und Rechtsprechung*. Tübingen: Mohr Siebeck.

Hong, Mathias. 2022. »Grundwerte des Antidiskriminierungsrechts: Würde, Freiheit, Gleichheit und Demokratie.« In *Handbuch Antidiskriminierungsrecht. Strukturen, Rechtsfiguren und Konzepte*, edited by A.K. Mangold and M. Payandeh, 67–123. Tübingen: Mohr Siebeck.

Jones, Meg L. 2017. »The right to a human in the loop: Political constructions of computer automation and personhood.« *Social Studies of Science* 47 (2): 216–39.

Kaminski, Margot E. 2019. »Binary Governance: Lessons from the GDPR's Approach to Algorithmic Accountability.« *Southern California Law Review* 92 (6): 1529–616.

Kant, Immanuel. 1786/2021. *Grundlegung der Metaphysik der Sitten*, edited by Theodor Valentiner, Seitenzahl nach Akademieausgabe Band IV. Stuttgart: Reclam.

Kant, Immanuel. 1797/1977. *Die Metaphysik der Sitten*, edited by Wilhelm Weischedel. Frankfurt a.M.: Suhrkamp.

Khaitan, Tarunabh. 2015. *A Theory of Discrimination Law*. Oxford: Oxford University Press.

Köchling, Alina, Shirin Riazy, Marius C. Wehner and Katharina Simbeck. 2021. »Highly Accurate, But Still Discriminatory.« *Business & Information Systems Engineering* 63 (1): 39–54.

Korsgaard, Christine M. 1996. *Creating the Kingdom of Ends*. Cambridge: Cambridge University Press.

Kosinski, Michal. 2021. »Facial recognition technology can expose political orientation from naturalistic facial images.« *Scientific Reports* 11 (1): Article 100 (7 pages).

Lehner, Roman. 2013. *Zivilrechtlicher Diskriminierungsschutz und Grundrechte. Auch eine grundrechtliche Betrachtung des 3. und 4. Abschnittes des Allgemeinen Gleichbehandlungsgesetzes (§§19-23 AGG)*. Tübingen: Mohr Siebeck.

Lippert-Rasmussen, Kasper. 2011. «We are all Different«: Statistical Discrimination and the Right to be Treated as an Individual.« *The Journal of Ethics* 15 (1): 47–59.

Lum, Kristian and William Isaac. 2016. »To predict and serve?« *Significance* 13 (5): 14–19.

Mahlmann, Matthias. 2008. *Elemente einer ethischen Grundrechtstheorie*. Baden-Baden: Nomos.

Mahlmann, Matthias. 2012. »Human Dignity and Autonomy in Modern Constitutional Orders.« In *The Oxford Handbook of Comparative Constitutional Law*, edited by Michael Rosenfeld and András Sajó, 1–26. Oxford: Oxford University Press.

Martini, Mario. 2021. »DS-GVO Art. 22 Automatisierte Entscheidungen im Einzelfall einschließlich Profiling.« In *Datenschutz-Grundverordnung Bundesdatenschutzgesetz DS-GVO BDSG, Kommentar*, 3. Auflage, edited by Boris P. Paal and Daniel A. Pauly. München: C.H. Beck.

Martini, Mario and David Nink. 2017. »Wenn Maschinen entscheiden… – vollautomatisierte Verwaltungsverfahren und der Persönlichkeitsschutz.« *Neue Zeitschrift für Verwaltungsrecht – Extra* 36 (10): 1–14.

Matz, Sandra C., Christina S. Bukow, Heinrich Peters, Christine Deacons, Alice Dinu and Clemens Stachl. 2023. »sing machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics.« *Scientific Reports* 13 (1): Article 5705 (16 pages).

McCrudden, Christopher. 2008. »Human dignity and judicial interpretation of human rights.« *European Journal of International Law* 19 (4): 655–724.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman and Aram Galstyan. 2021. »A survey on bias and fairness in machine learning.« *ACM Computing Surveys* 54 (6): 1–35.

Obermeyer, Ziad, Brian Powers, Christine Vogeli and Sendhil Mullainathan. 2019. »Dissecting racial bias in an algorithm used to manage the health of populations.« Science 366 (6464): 447–53.

Orwat, Carsten. 2019. *Diskriminierungsrisiken durch Verwendung von Algorithmen*. Studie erstellt mit einer Zuwendung der Antidiskriminierungsstelle des Bundes. Berlin: Nomos.

Pessach, Dana and Erez Shmueli. 2022. »A review on fairness in machine learning.« *ACM Computing Surveys (CSUR)* 55 (3): 1–44.

Schaber, Peter. 2013. *Instrumentalisierung und Menschenwürde*. 2 ed. Münster: Mentis.

Schaber, Peter. 2016. »Menschenwürde.« In *Handbuch Gerechtigkeit*, edited by Anna Goppel, Corinna Mieth and Christian Neuhäuser, 256–62. Stuttgart: J.B. Metzler.

Schauer, Frederick. 2018. »Statistical (and non-statistical) discrimination.« In *The Routledge Handbook of the Ethics of Discrimination*, edited by Kasper Lippert-Rasmussen, 42–53. London: Routlegde.

Scholz, Philip. 2019. »DSGVO Art. 22 Automatisierte Entscheidungen im Einzelfall einschließlich Profiling.« In *Datenschutzrecht. DSGVO und BDSG*, edited by Spiros Simitis, Gerrit Hornung and Indra Spiecker genannt Döhmann. Baden-Baden: Nomos.

Sloane, Mona, Emanuel Moss and Rumman Chowdhury. 2022. »A Silicon Valley love triangle: Hiring algorithms, pseudo-science, and the quest for auditability.« *Patterns* 3 (2): 100425.

Smuha, Nathalie A. 2021. »Beyond the individual: governing AI's societal harm.« *Internet Policy Review* 10 (3).

Teo, Sue A. 2023. »Human dignity and AI: mapping the contours and utility of human dignity in addressing challenges presented by AI.« *Law, Innovation and Technology* 15 (1): 1–39.

Thomsen, Frej K. 2017. »Discrimination.« In *Oxford Research Encyclopedia of Politics* (online), edited by William R. Thompson. New York: Oxford University Press.

Ulgen, Ozlem. 2017. »Kantian Ethics in the Age of Artificial Intelligence and Robotics.« *Questions of International Law (QIL) Zoom-in* 43: 59–83.

Ulgen, Ozlem. 2022. »AI and the Crisis of the Self: Protecting Human Dignity as Status and Respectful Treatment.« In *The Frontlines of Artificial Intelligence Ethics: Human-Centric Perspectives on Technology's Advance*, edited by Andrew J. Hampton and Jeanine A. DeFalco, 9–33. Abingdon, New York: Routledge.

Valcke, Peggy, Damian Clifford and Vilté K. Dessers. 2021. »Constitutional Challenges in the Emotional AI Era.« In *Constitutional Challenges in the Algorithmic Society*, edited by Hans-W. Micklitz, Oreste Pollicino, Amnon Reichman, Andrea Simoncini, Giovanni Sartor and Giovanni De Gregorio, 57–77. Cambridge: Cambridge University Press.

von Ungern-Sternberg, Antje. 2022. »Diskriminierungsschutz bei algorithmenbasierten Entscheidungen.« In *Handbuch Antidiskriminierungsrecht. Strukturen, Rechtsfiguren und Konzepte*, edited by Anna K. Mangold and Mehrdad Payandeh, 1131–80. Tübingen: Mohr Siebeck.

Yeung, Karen. 2019. *Responsibility and AI. A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework*. Council of Europe study DGI(2019)05. Council of Europe, Expert Committee on human rights dimensions of automated data processing and different forms of artificial intelligence (MSI-AUT). Strasbourg.

Zarsky, Tal. 2013. »Transparent predictions.« *University of Illinois Law Review* 2013 (4): 1503–69.