KONRAD ADENAUER STIFTUNG

Vom Prinzip zum Paradigma
KI und Ethik in konkreten Anwendungskontexten

Antonio Bikić

www.kas.de

Auf einen Blick

- Für Technologien, die einen hohen Automatisierungsgrad aufweisen, ist eine erweiterte Anwendung des Vorsorgeprinzips notwendig, um Belastungen oder Schäden im Voraus zu vermeiden oder weitestgehend zuverringern.

- In den hier aufgeführten Beispielen zeigen sich diverse ethische Problemstellungen. Diese können nicht mit lediglich hochrangigen ethischen Prinzipien aufgelöst werden. Die Prinzipien müssen sich in konkreten Anweisungen widerspiegeln.

- Der algorithmischen Implementierung eines Fairness-Prinzips muss eine gerechtigkeitstheoretische Diskussion vorangehen.

- Langfristig bieten sich Kooperationen von Universitäten und Unternehmen an. Auch sind Verbundprojekte mit Forschungsinstituten möglich. Auf diese Weise können Unternehmen die nötige Ethik-Expertise aufbauen, um künstlich intelligente Systeme allgemeinverträglich zu konzipieren.

Inhaltsverzeichnis

Einleitung 3

KI-Technologien erfordern eine erweiterte Anwendung des Vorsorgeprinzips 4

Praxisbeispiel 1: Empfehlungssysteme 4

Praxisbeispiel 2: Automatisiertes Fahren 6

Praxisbeispiel 3: Bewerbungsverfahren 8

Ethische Prinzipien sind für eine Ethik im Kontext Künstlicher Intelligenz nicht ausreichend 11

Einleitung

Der Ruf nach ethischer Bewertung und Regulierung intelligenter Algorithmen wird lauter: vor allem seit dem jüngsten Aufkommen selbstlernender, sogenannter *Machine Learning*-Algorithmen. Algorithmen, die auf Basis von Beispielen lernen. Dafür sind – entgegen weitverbreiteter Ansicht – nicht zwingend große Datenmengen notwendig. Sogenannte *One-Shot-Learning*-Ansätze versuchen mittels weniger Beispiele (kleine Datenmengen), Objekte zu klassifizieren.

Um Systeme, die Künstliche Intelligenz (KI) nutzen, ethisch zu bewerten, muss man zwei Konzepte unterscheiden: Intelligenz und Aktionsfähigkeit.[1] Künstlich intelligente Systeme lösen durch ihre Aktionsfähigkeit Aufgaben. Die Kriterien dafür sind entweder vorgegeben oder werden vom System selbst erarbeitet. Die Lösungen werden im Anschluss daran als intelligent interpretiert, obwohl bei Maschinen nur Aktionsfähigkeit ohne tatsächliche Intelligenz gegeben ist. Eine Maschine kann deshalb nicht als im eigentlichen Sinne intelligent bezeichnet werden. Sonst läge es nahe, auch einen Fluss, der ein Hindernis umfließt, ebenfalls als intelligent zu bezeichnen. Intelligenz basiert auf kognitiven Fähigkeiten, nicht auf bloßer Aktionsfähigkeit.

Es gibt unterschiedliche Methoden, künstlich intelligente Systeme zu realisieren. *Machine Learning* ist einer der dominierenden Ansätze. Mithilfe von *Machine Learning* löst man momentan häufig Klassifikationsprobleme. Dies gelingt mit dem sogenannten *subsymbolischen Machine Learning*-Ansatz, der jedoch die Vorgänge innerhalb der Maschinen nicht mehr nachvollziehbar macht. Die Maschinen werden zu *Black Boxes*. Deutlich länger im Einsatz, weil technisch einfacher zu realisieren, sind sogenannte *symbolische* Ansätze. Man bedient sich ihrer bei sogenannten Expertensystemen (z. B. Programme zur Unterstützung von medizinischen Diagnosen). Symbolische Ansätze eignen sich am besten zur Lösung abstrakter Probleme.

Expertensysteme gibt es bereits seit Jahrzehnten. Ethische, in der breiten Öffentlichkeit wahrgenommene Debatten werden erst mit Aufkommen des subsymbolisch realisierten maschinellen Lernens geführt. Lange erwartete man an neuralgischen Punkten von Entscheidungsketten natürliche Personen. Nun besteht die Möglichkeit, dass Maschinen den Menschen dabei ersetzen, Entscheidungen zu treffen. In Analogie zum menschlichen Handeln „entscheidet“ dann die Maschine.

KI-Technologien erfordern eine erweiterte Anwendung des Vorsorgeprinzips

Die Suche nach einer ethischen Rechtfertigung für den Einsatz von Software entspricht dem aus der Umwelt- und Gesundheitspolitik bekannten Vorsorgeprinzip. Insbesondere gilt dies für Technologien, die als *disruptiv* bezeichnet werden. Wenn Maschinen in Umlauf gebracht werden sollen, in deren Funktionsweisen sich die Werte der europäischen Gesellschaften widerspiegeln sollen, muss hier das Vorsorgeprinzip zum Tragen kommen. Es soll denkbare Belastungen und Schäden, die beim Einsatz z. B. einer Technologie in Betracht kommen könnten, im Voraus möglichst vermeiden oder weitestgehend verringern. In der Praxis ist die Verhältnismäßigkeit von Risikoannahmen für die Nutzenabwägung entscheidend.

Dokumente wie z. B. die *Ethik-Leitlinien für eine vertrauenswürdige Künstliche Intelligenz*[2] der *Hochrangigen Expertengruppe für Künstliche Intelligenz* der Europäischen Kommission haben für eine ethische Reflexion und Bewertung maschineller Funktionsweisen sehr abstrakte Leitlinien formuliert. Sobald jedoch konkrete Situationen betrachtet werden, in denen ethische Expertise notwendig wird, zeigt sich die Unterschiedlichkeit der ethischen Implikationen. Hier muss u. a. entschieden werden, nach welchem ethischen Paradigma gehandelt werden soll. Unter einem ethischen Paradigma versteht man ein Ethik-System (z. B. eine Pflichtenethik). Und es muss klar sein, dass es verschiedene Ethik-Systeme gibt, von denen prinzipienbasierte Ethiken ein System darstellen. Die jeweiligen Unternehmen und Institutionen müssen sich auf ein bestimmtes Paradigma festlegen. Ein ethisches Paradigma kann dabei auch ethische Prinzipien berücksichtigen. Jedoch wird nur durch eine Strukturierung dieser Prinzipien die ethische Reflexion ein integraler Bestandteil der Produktentwicklung und -evaluierung. Einigen Unternehmen ist nicht klar, dass z. B. das Prinzip Fairness im Grunde eine fortgeführte Debatte um Gerechtigkeitstheorien ist. Und oftmals zeigt sich die Verschiedenartigkeit der ethischen Implikationen erst in konkreten Kontexten, in denen künstlich intelligente Systeme eingesetzt werden. Dort erst werden dann beispielsweise Fragen der Gerechtigkeit thematisch. Die Prinzipien selbst sind abstrakte und nicht messbare Größen, weil sie keine Maßeinheit besitzen. Es gibt jedoch Möglichkeiten, die Effekte einer bestimmten Fairness-Konzeption aufzuzeigen. Quotenregelungen können, wenn sie erfolgreich sind, eine Durchlässigkeit bestimmter benachteiligter Gruppen als Folge mit sich führen.

1

Praxisbeispiel: Empfehlungssysteme

Algorithmen der Künstlichen Intelligenz sind oftmals in Systeme integriert, ohne direkt aufzufallen. So sind beispielsweise bei Streaming-Diensten für Videos Empfehlungsdienste bzw. *Recommender Systems* üblich. Diese Programme basieren oft (jedoch nicht ausschließlich) auf verschiedenen Formen des maschinellen Lernens. Wenn Empfehlungsdienste hinreichend viele Informationen zu einer Person gesammelt haben, werden sie in die Lage versetzt, den Nutzerinnen und Nutzern Produkte zu empfehlen. Diese Empfehlung basiert auf dem bisherigen Nutzungsverhalten.

Problematisch werden solche Systeme dann, wenn z. B. auf öffentlich zugänglichen Videoportalen Videos von Minderjährigen hochgeladen werden. Diese Praxis bezieht

sich u. a. auf sogenannte *Family Vlogger*, also Familienangehörige, die Videos hochladen, die das tägliche Leben ihrer Familie dokumentieren. Hierbei handelt es sich oftmals um Bildmaterial, auf dem Kinder z. B. in Strandurlauben nur leicht bekleidet dargestellt werden. Doch diese Videos können auch Personen mit pädophilen Neigungen empfohlen werden, die nach entsprechenden Videos suchen.[3] In der Folge gibt es Videos mit Kindern, die in kurzer Zeit auffallend oft aufgerufen werden. Die Kinder tun kein Unrecht, die Videos werden von ihren Erziehungsberechtigten ins Netz geladen. Hier gerät die Arglosigkeit der Eltern mit der Zielgerichtetheit bestimmter Suchanfragen in Konflikt. Man könnte nun dafür optieren, Videos mit Kindern nicht mehr durch Empfehlungsdienste zu bewerben. Dies hätte auch negative Effekte: Die Gemeinschaft der *Family Vlogger* besteht aus Millionen von Nutzerinnen und Nutzern. Und nicht jede Person, die auf solchen Portalen nach Bildmaterial sucht, hat pädophile Neigungen. Die Resultate der Empfehlungsdienste sind lediglich der Nachweis, dass die Software sehr gut funktioniert. Sie zeigt genau an, was zur Suchanfrage passt. Die Software ist nicht in der Lage, die Intentionen der Suchenden zu dechiffrieren.

Wenn problematische Inhalte von Empfehlungssystemen nicht mehr berücksichtigt werden dürften, beträfe das nicht nur Bildmaterial mit Kindern, sondern auch andere Bereiche: etwa Videos mit verschwörungstheoretischen oder extremistischen Inhalten. Da die Intentionen der Nutzerinnen und Nutzer bei der Suchanfrage jedoch nicht eindeutig sind, erweist sich eine generelle Einschränkung als nicht zielführend.

An diesen Beispielen wird deutlich, wie problematisch die maschinelle Bewertung von Inhalten ist. Auch wenn Maschinen allem Anschein nach für unbedenkliche Aufgaben konzipiert werden, müssen sie in der Lage sein, kontextsensibel Missbrauch vorzubeugen. Da auch diese Software einen Missbrauch nicht ausschließen kann, kann sie nicht vollständig reguliert werden.

Im Fall der Empfehlungsdienste ist den Betreibern von Videoportalen schwer vorzuwerfen, dass ihre Algorithmen gut funktionieren. Auch kann man mit guten Gründen argumentieren, dass mündige Personen die Videos ins Netz stellen. Und doch ist es problematisch, wenn unerwünschte Intentionen – etwa pädophile Neigungen – einen Suchauftrag steuern. Hier wäre eine gesetzlich verpflichtende risikoethische Analyse seitens des Betreibers ein gangbarer Weg. Personen, die Videomaterial auf den Portalen hochladen, müssen von den Betreibern darauf aufmerksam gemacht werden, über welche Suchanfragen ihre Videos gefunden werden können. Der Nutzer muss sich darüber im Klaren sein, mit welcher Wahrscheinlichkeit seine Videos in problematischen Suchanfragen auftauchen. Solche Suchanfragen sollten für die Person, die die Videos hochlädt, deutlich markiert sichtbar sein.

Da vielen Eltern nicht bewusst ist, wofür sie ihre Einwilligung geben, wenn sie auf entsprechenden Portalen Dateien hochladen, muss ihnen vermittelt werden, wen sie – wenn auch unbeabsichtigt – adressieren. Dieses Wissen würde ihr Verhalten signifikant steuern. Wenn sie ihr Handeln ethisch begründen können und über Medienkompetenz verfügen, handeln sie eigenverantwortlich.

Der Bereich, der hier reguliert werden müsste, ist insgesamt nur bedingt regulierbar, da oft nicht klar ist, welche Inhalte problematisch werden können. Hier zeigen sich die Grenzen der Regulierung und der technischen Umsetzung bzw. Automatisierung

bestimmter Vorgänge. Daraus ergibt sich die Pflicht, die Medienkompetenz der Nutzerinnen und Nutzer solcher Portale zu stärken. Als unterstützende Optionen wären Online-Kampagnen oder eine individuelle Beratung der Nutzer und Nutzerinnen mit einem Chat-Assistenten durch den Dienstanbieter denkbar.

2 Praxisbeispiel: Automatisiertes Fahren

Das hoch- und vollautomatisierte Fahren wird unter ethischen Gesichtspunkten oft einseitig betrachtet, geht es doch fast ausschließlich um Dilemma-Situationen, die in bestimmten ethischen Paradigmen auftreten (jedoch nicht in allen). In der Realität gibt es eine Reihe weiterer Probleme, die einer ethischen Analyse bedürfen.

Ein besonderes Thema ist das sogenannte *Nudging:* die bewusste Steuerung des menschlichen Verhaltens hin zu einer Entscheidung, die einer anderen Person oder Personengruppe als wünschenswert erscheint. Ein *Nudge* ist eine zwanglos erzeugte minimale Beeinflussung des Verhaltens einer Person.

Diverse Techniken der Verhaltensbeeinflussung erhöhen die Wahrscheinlichkeit, dass eine gewünschte Entscheidung getroffen wird. So werden teurere Produkte in den Regalen eines Supermarkts auf Sichthöhe platziert, wohingegen die günstigeren unten stehen. Ohne Zwang auf die Kundschaft auszuüben, werden die teureren Produkte häufiger gekauft.

Bei Pilotinnen und Piloten konnte bereits nachgewiesen werden, dass *Nudging* Treibstoff sparen kann.[4] Im Kontext des hoch- und vollautomatisierten Fahrens stellt sich hingegen die Frage, ob man Fahrerinnen und Fahrern anbieten darf, bestimmte Funktionen zu aktivieren. Angenommen, es wäre statistisch nachgewiesen, dass ein von einem Assistenzsystem durchgeführtes Überholmanöver um den Faktor 100 sicherer ist, dann wäre dies ein gutes Argument, ein solches System so oft wie möglich zu nutzen. Und stellen wir uns zudem eine Fahrsituation vor, in der das Fahrzeug erkennt, dass die Konzentration des Fahrers/der Fahrerin abnimmt (z. B. durch Sekundenschlaf). Die Augen des Fahrers/der Fahrerin werden durch Kamerasysteme kontrolliert: eine Technik, die übrigens von deutschen Herstellern stammt.[5] Die signifikant erhöhte Sicherheit durch das Nutzen des Assistenzsystems für Überholmanöver und das Wissen um seine/ihre verringerte Konzentration sind gute Gründe, das Verhalten des Fahrers/der Fahrerin ohne Zwang zu beeinflussen. Wie das Fahrzeug eine solche Verhaltensbeeinflussung veranlasst, ist variabel.

Ein weiterer Punkt ist in diesem Zusammenhang auch der Gedanke der Verkehrserziehung: Dürfen Fahrzeuge negative Anreize setzen, um bestimmte Manöver einer Fahrerin oder eines Fahrers zu erschweren? Wenn Manöver sehr riskant oder nicht konform mit der Straßenverkehrsordnung sind, spricht einiges dafür.

Eine ethische Bewertung ist im Fall des *Nudgings* für hoch- und vollautomatisierte Fahrzeuge nur dann möglich, wenn die genauen Ziele der Feinjustierung des Verhaltens offengelegt sind. So wird *Fremdnudging* vermieden und *Selbstnudging* gefördert. Doch nicht alle Verhaltensweisen der Fahrerinnen und Fahrer sollten durch die Automobilhersteller oder den Staat beeinflusst werden. Sollte es ver-

schiedene Feinjustierungen seitens der Algorithmen des Fahrzeugs geben, muss den Fahrerinnen und Fahrern auch aufgezeigt werden, in welcher Weise das Fahrzeug das Fahrverhalten insgesamt reguliert. Wenn ein Fahrzeug für bestimmte (z. B. riskante) Fahrmanöver negative Anreize produziert, müssen diese so gestaltet werden, dass sie eine unterstützende, niemals aber eine einschränkende Funktion haben. Die Fahrerin oder der Fahrer muss in der Lage sein, bestimmte *Nudges* seitens des Fahrzeugs ohne besonderen Aufwand zu umgehen. Dies ist dann der Fall, wenn nach ihrem situativen Ermessen die Sicherheit des Fahrverlaufs von einer Umgehung der *Nudges* abhängt. So kann durch ein zwar kurzzeitig riskantes, aber notwendiges Fahrmanöver die Sicherheit wiederhergestellt werden. Deswegen dürfen die Anreize zu bestimmtem Verhalten seitens des Fahrzeugs nie zu sehr einschränkend sein.

An dieser Stelle sollten auch wirtschaftliche Interessen thematisiert werden: Versicherungsbeiträge könnten signifikant sinken, würde der Nachweis erbracht, dass z. B. eine Fahrerin oder ein Fahrer kontinuierlich und in den richtigen Momenten die Assistenzsysteme ihres Fahrzeugs nutzt. Die Sicherheit der Insassen und aller Verkehrsbeteiligten würde erhöht. Auch nicht alle Assistenzsysteme permanent zu nutzen, kann sich als sinnvoll erweisen: Fahrerinnen und Fahrer müssen Routinen trainieren, um ihr Fahrzeug zu lenken. In diesem Kontext kann jedoch die Verquickung ökonomischer und privater Interessen problematisch werden. Es muss klar sein, ob das Fahrzeug z. B. hin zu einer Senkung der Versicherungsbeiträge optimiert oder ob es um eine erweiterte Erhöhung der Sicherheit der Insassen und Verkehrsbeteiligten geht. Im Idealfall sind diese Ziele deckungsgleich, aber die Mittel zur Erreichung der Ziele müssen eben nicht notwendig identisch sein. Die Frage wäre hier, ob zwischen diesen beiden Optionen überhaupt gewählt werden dürfte. Auch müssten bei einer solchen Option bestimmte Fahrdaten gespeichert und ausgewertet werden. Insgesamt gilt natürlich auch hier das Prinzip der Verantwortbarkeit: Das Befolgen einer Verhaltensbeeinflussung seitens eines hochautomatisierten Systems muss auch im Nachhinein von den Fahrern und Fahrerinnen verantwortet werden können. Gemeint ist damit, dass es nicht zu einer Verhaltensbeeinflussung kommen darf, die z. B. strafrechtliche Konsequenzen oder ein unzumutbares Risiko mit sich führt.

Die einwandfreie ethische Bewertung hängt einerseits von den impliziten Zielen der *Nudges* seitens des Fahrzeugs ab. Andererseits muss klar sein, ob die Produkthaftung oder ob die Eigenverantwortung der Fahrerinnen und Fahrer maximiert werden soll. Dies hat etwas mit der übergeordneten Frage zu tun, inwiefern die Fahrzeuge Manöver und Verhaltensweisen anbieten, die verantwortbar sind. Hierbei muss die Maxime gelten, dass nur von der Fahrerin oder dem Fahrer verantwortbare Verhaltensweisen angeboten werden dürfen, die angemessen und im Sinne der Fahrerinnen und Fahrer sind. Auf diese Weise kann sukzessiv sichergestellt werden, dass Menschen an eine Fahrsituation herangeführt werden, die sie kognitiv bewältigen können. Nur dann können sie das begründete Gefühl entwickeln, verantwortbare Entscheidungen mittels der Maschine anzustoßen, da sie selbst die Implikationen einsehen können. Die Kontrolle über das Fahrgeschehen ist zentral für die Verantwortungszuschreibung. Auf diese Weise kann auch die Entwicklung des *Sense of Agency*[6], also die Entwicklung des Gefühls, Autor oder Autorin[7] der eigenen Handlungen zu sein, im Kontext hoch- und vollautomatisierter Systeme gefördert werden.

Ein Fahrzeug, das standardmäßig die Steuerung hin zu einer Haftung seitens des Herstellers maximiert, erzeugt eine zutiefst unmoralische Situation: Bei dem Fahrer eines solchen Fahrzeugs stellt sich dann das Gefühl ein, keine echte Kontrolle über die Maschine zu haben. Dies unterminiert die Übernahme von Verantwortung. Gleichzeitig würde vom Fahrer bei hoch- und vollautomatisierten Fahrzeugen allerdings verlangt, dennoch der Fahraufgabe nachzukommen, die Verantwortung impliziert. Es ist daher wichtig, dass nicht nur dokumentiert wird, ob die Assistenzsysteme oder die Person die Steuerung des Fahrzeugs zu einem bestimmten Zeitpunkt maßgeblich kontrolliert haben. Es müssen zudem auch Informationen festgehalten werden, aus denen sich ableiten lässt, ob die Empfehlungen der Assistenzsysteme verantwortbar waren. Die Kriterien müssen hier nicht nur gesetzlichen, sondern auch risikoethischen Kriterien entsprechen. Die risikoethischen Kriterien sollten von den Fahrerinnen und Fahrern ausgewählt werden können.

3 Praxisbeispiel: Bewerbungsverfahren

Es liegt nahe, selbstlernende Algorithmen, die sich insbesondere für Klassifikationen sehr gut eignen, für die Optimierung von Bewerbungsprozessen einzusetzen. Ohne sehr komplexe Algorithmen nutzen zu müssen, können Lebensläufe problemlos „aussortiert“ werden, die bestimmten Anforderungen nicht entsprechen.

In Zusammenhang mit Bewerbungsverfahren wird oft der Begriff der Fairness thematisiert. Dieser Begriff ist seit geraumer Zeit aus der Gerechtigkeitsdebatte übernommen worden und wird mittlerweile fast als Synonym für Gerechtigkeit gebraucht. Allerdings muss man beachten, dass es zahlreiche andere Gerechtigkeitstheorien gibt, die heute Anwendung finden und nicht primär Fairness zum Inhalt haben.[8]

Fairness wird insbesondere bei selbstlernenden Systemen regelmäßig mit dem Begriff *Bias* in Verbindung gebracht. Gemeint sind damit kognitive Verzerrungen bzw. Vorurteile.[9] Wenn behauptet wird, dass bestimmte Datensätze, auf Basis derer maschinelles Lernen möglich wird, einen *Bias* aufweisen, dann bedeutet dies nicht, dass die Daten vorurteilsbelastet sind. Gemeint ist damit, dass sich aus den Daten Vorurteile ableiten lassen: Wenn z. B. eine Merkmalsausprägung in einem Datensatz besonders dominant ist, dann kann sich bei dem selbstlernenden System das Urteil einstellen, dass diese Merkmalsausprägung gewünscht ist. Wenn lediglich festgestellt wird, dass ein Unternehmen Angestellte braucht, stellt diese Norm kein Problem dar. Wenn jedoch zur Norm erhoben wird, dass die Angestellten jedes Unternehmens ein bestimmtes Geschlecht haben sollten, liegt ein Problem vor.

Bewerbungsprozesse können auf verschiedene Variablen hin optimiert werden. Es kann versucht werden, den Bewerbungsprozess insgesamt gerecht zu gestalten oder es kann versucht werden, das Resultat des Prozesses gerecht zu gestalten. Dazu gibt es diverse Fairness-Metriken, durch die gemessen werden kann, inwiefern die durch einen Algorithmus erzeugten Ergebnisse einer bestimmten Gruppe gegenüber fair sind. Diese Metriken messen also, inwiefern alle Benutzerinnen und Benutzer an den Ressourcen eines Systems (z. B. Arbeitsstellen) angemessen bzw. fair teilhaben können. Es gibt ebenfalls Methoden, die den *Bias* senken können – also einer bestimmten Gruppe gegenüber faire Bedingungen der Beurteilung schaffen. Werden diese Metri-

ken jedoch ohne Gesamtkonzept angewandt, kann der Prozess, auch wenn er in sich einzelne Vorurteile eliminiert, im Resultat unfaire bzw. ungerechte Ergebnisse produzieren. Das bloße Vorhandensein dieser Metriken ist daher kein Garant für Fairness. Durch die diversen Ziele, auf die hin optimiert wird, kann insgesamt ein nicht-intendiertes Gesamtergebnis erzielt werden.

Eine Form von positiver Diskriminierung stellen beispielsweise Quotenregelungen dar. Dies ist eine Form der distributiven bzw. der Verteilungsgerechtigkeit. Hier wird darauf hingearbeitet, dass der Nutzen der Gesellschaft allen Menschen in gleichem Maße zukommt. Eines der Probleme bei diesen Maßnahmen ist jedoch beispielsweise im Rahmen der Frauenquote, dass Formulierungen wie „bei gleicher Eignung werden Frauen bevorzugt behandelt“ nicht einheitlich formalisiert sind. Durch eine Formalisierung können Formulierungen durch Algorithmen verarbeitet werden. Wann aber sind zwei Personen gleich geeignet? Es gibt zahlreiche Kriterien, die herangezogen werden können, um die fachliche Gleichheit zweier Personen festzustellen. Diese Kriterien (Abschlussnote, Studienort, Berufserfahrung etc.) sollten jedoch normiert sein und die Situation der jeweiligen Personen berücksichtigen. Wenn die Bedingungen für die Feststellung der Gleichheit zweier Personen jedoch nicht mehr vergleichbar sind, wird eine als ungerecht empfundene Situation erzeugt. Es ist gut möglich, dass einheitliche Standards bei der Feststellung der gleichen Eignung zweier Personen für eine Stelle derzeit nicht einmal in den einzelnen Unternehmen einheitlich umgesetzt werden. Gerade durch den Einsatz von algorithmisch gesteuerten Lösungen könnte dieser Umstand ausgebessert werden.

Wenn Bewerbungsprozesse automatisiert werden, gilt es, eine Gesamtperspektive auf das System zu entwickeln. Dies fängt bereits damit an, dass Ausschreibungstexte, wenn sie automatisch generiert werden, neutral formuliert werden müssen. Anderenfalls werden durch die Formulierung selbst bestimmte Gruppen ausgeschlossen, die eigentlich zum Kreis der Bewerberinnen und Bewerber gehören. Zudem müssen allgemeine Vorgaben entwickelt werden, die sowohl für eine Verteilungsgerechtigkeit (distributiv[10]) sorgen als auch eine Regelgerechtigkeit (prozedural[11]) befördern. Je nachdem, wie groß ein Unternehmen ist, müssen auf diversen Ebenen verschiedene Gerechtigkeitsansätze implementiert werden, um beispielsweise die Aufstiegsmöglichkeiten aller sozialen Gruppen in einem Unternehmen zu garantieren. Fairness als Gerechtigkeitstheorie kann insgesamt unfair werden, wenn die Fairness-Metriken nicht kohärent eingesetzt werden. Wenn sehr viele Prozesse in Bewerbungsverfahren hochautomatisierten Systemen überlassen werden, können sich über Jahre hinweg ungerechte Strukturen etablieren, deren Aufdeckung durch den unpersönlichen Charakter von Algorithmen erschwert wird.

Dem kann man entgegenwirken, wenn man gerechtigkeitstheoretische Standards etabliert. Eine gerechte Behandlung kann durch Maschinen in bestimmten Punkten sogar einheitlicher umgesetzt werden als durch natürliche Personen.

Zudem müssen für eine Automatisierung von Bewerbungsprozessen verbindliche Standards für Formalisierungen wie „bei gleicher Eignung“ ausgearbeitet werden. Es muss klar sein, ob und inwiefern dabei z. B. nur auf den Studienabschluss und -ort geachtet wird oder ob auch zusätzlich z. B. die Abschlussnote, die Berufserfahrung, das Eintrittsalter und Ähnliches eine Rolle spielen dürfen. Die Bewerberinnen und

Bewerber müssen sich objektiv und überall gleichbehandelt fühlen. Nur dann kann auch die Überzeugung entstehen, dass die künstlich intelligenten Systeme bestimmte Gerechtigkeitsstandards wirkungsvoll umsetzen.

Eine gesonderte Debatte stellen Systeme dar, die Video- und Audiodaten verarbeiten. Auf diese Weise kann u. a. die Körpersprache einer Person untersucht werden. Auch ist es möglich, auf Basis der Audiodaten mittels *Machine Learning* abzuschätzen, ob eine Person möglicherweise an einer Depression leidet.[12] Diese Systeme flächendeckend für Bewerbungsgespräche einzusetzen, ist problematisch. Hier deckt die geltende Rechtslage bereits viel ab und regelt, was erlaubt ist. Eine Person müsste sehr wahrscheinlich zustimmen, dass ihre Video- und Audiodaten verarbeitet werden und u. a. auch hin auf eine Depression geprüft werden. Problematisch wäre es dann, wenn die Personen diese Zustimmung ablehnen. Daraus können bestimmte Intentionen der Bewerberinnen und Bewerber abgeleitet werden, die gar nicht zutreffen müssen. Die Nutzung solcher Systeme wäre also flächendeckend bei Bewerbungsverfahren kaum zu verantworten. Bei solchen Systemen sollte daher ein Konsens gefunden werden, der eine bestimmte Anzahl an Analysen zulässt.

Es muss auch überprüft werden können, ob ein System, das zur Automatisierung von Bewerbungsprozessen eingesetzt wird, zwischen einer *erwünschten* und einer *faktischen* Norm unterscheiden kann. So kann es faktisch Norm sein, dass das Geschlecht darüber entscheidet, ob man eine Stelle erhält oder nicht. Aber es ist die erwünschte Norm, dass diese Information bei der Beurteilung keine Rolle spielen darf. Daher sollte es also die Möglichkeit geben, dass die Maschine hier auch autonom zwischen diesen Normen unterscheiden kann. Erwünschte Normen könnte man in diesem Kontext auch Werte nennen. Für die Aufrechterhaltung und Durchsetzung dieser Werte gibt es gute Argumente. Das macht sie zu erwünschten Normen.

Alle Akteurinnen und Akteure, die bei der Entwicklung, Herstellung, dem Betrieb, der Evaluierung und Nutzung der Anwendungen beteiligt sind, haben durch die Nähe zum automatisierten Bewerbungssystem eine Expertise, die externe Prüfungen nur schwerlich erreichen können. Ethik-Expertise wird in den Prozessstrukturen eines automatischen Bewerbungssystems daher nur dann sichtbar, wenn die an dem System beteiligten Personen Ethik-Kenntnisse vorweisen können. Hier wäre es sinnvoll, wenn es zu einer Zusammenarbeit zwischen Hochschulen und Unternehmen kommen könnte. Dabei könnten konkrete Probleme, die sich empirisch erheben lassen, durch die Ethik-Expertise von Universitäten bearbeitet werden.

Es wären aber auch Verbundprojekte mit Forschungsinstituten möglich. Dabei wären zum einen Kooperationen zur Formalisierung und Implementierung bestimmter Quotenregelungen denkbar. Zum anderen wäre es sinnvoll, wenn für den gesamten Bewerbungsprozess, falls er automatisiert ablaufen soll, einheitliche Ansätze für die Implementierung von Fairness eingesetzt würden. Auf diese Weise wäre es möglich, alle Formen künstlich intelligenter Systeme hin auf ihre gesellschaftliche Gesamtverträglichkeit zu prüfen. Dies ist vor allem dann geboten, wenn diese Systeme für sehr viele Bürgerinnen und Bürger an neuralgischen Stellen in Entscheidungsketten zum Einsatz kommen.

Ethische Prinzipien sind für eine Ethik im Kontext Künstlicher Intelligenz nicht ausreichend

Eine zu starke Fokussierung auf ethische Prinzipien ist für die momentane Umgebung, in die künstlich intelligente Algorithmen implementiert werden, nicht optimal.[13] Dies bedeutet nicht, dass ein prinzipienbasierter Ansatz grundsätzlich falsch ist, sondern dass es viel konkreterer Ausformulierungen der Prinzipien bedarf. Die Praxisbeispiele haben u. a. gezeigt, dass im Fall der automatisierten Bewerbungsprozesse auch gerechtigkeitstheoretische Fragestellungen in den Vordergrund rücken. Der bloße Einsatz von Fairness-Metriken liefert auf diese Fragestellungen keine Antwort. Vor allem heißt das auch, dass ethische Paradigmen formuliert werden müssen, mittels derer ethisch richtige von falschen Entscheidungen unterschieden werden können. Die Notwendigkeit dieses Ansatzes zeigt sich vor allem dann, wenn konkrete Anwendungskontexte von künstlich intelligenten Algorithmen untersucht werden.

Letzter Abruf der in den Quellenangaben genannten Links: 06.03.2020

1 Luciano Floridi, J.W. Sanders: „On the Morality of Artificial Agents“, *Minds and Machines* 14, 349–379 (2004). https://doi.org/10.1023/B:MIND.0000035461.63578.9d
2 https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60664
3 https://www.nytimes.com/2019/06/03/world/americas/youtube-pedophiles.html
4 Vgl. dazu http://www.lse.ac.uk/GranthamInstitute/wp-content/uploads/2017/01/Working-paper-262-Gosnell-et-al.pdf
5 https://www.nytimes.com/2017/03/16/automobiles/wheels/drowsy-driving-technology.html
6 Vgl. dazu https://www.cell.com/current-biology/pdf/S0960-9822(12)00191-1.pdf
7 Zu dem Begriff der Autorschaft über die eigenen Handlungen: Julian Nida-Rümelin: *Philosophie einer humanen Bildung*, Hamburg 2013.
8 Diese Theorien gehen im Wesentlichen auf Rawls zurück, vgl. dazu John Rawls: *Eine Theorie der Gerechtigkeit*, Frankfurt am Main 1975 (Original: *A Theory of Justice*, 1971).
9 *Bias* tritt auch unabhängig von selbstlernenden Systemen auf. Populär ist u. a. die Darstellung von Kahneman, vgl. dazu Daniel Kahneman: *Schnelles Denken, langsames Denken*, München 2012.
10 Vgl. dazu https://arxiv.org/pdf/1710.03184.pdf
11 Vgl. dazu http://mlg.eng.cam.ac.uk/adrian/AAAI18-BeyondDistributiveFairness.pdf
12 Vgl.dazu die MIT-Publikation: http://news.mit.edu/2018/neural-network-model-detect-depression-conversations-0830
13 https://www.nature.com/articles/s42256-019-0114-4

Der Autor

Antonio Bikić
geboren 1987, ist Doktorand am Münchner Graduiertenkolleg für Ethik in der Praxis und promoviert an der LMU München und der ETH Zürich zur Implementierbarkeit ethischer Paradigmen. Er hat einen Studienhintergrund in Philosophie und Computerlinguistik/Informatik und arbeitete für das Rechenzentrum der Max-Planck-Gesellschaft, an Lehrstühlen für Praktische Philosophie/Ethik (LMU München) sowie u. a. für den Verband der Automobilindustrie, das Bauhaus Luftfahrt, PwC München und für das Fraunhofer-Institut für Arbeitswirtschaft und Organisation. Er hält regelmäßig an Universitäten in Deutschland, Luxemburg und Österreich Seminare und Vorträge zur Philosophie des Geistes und zur Ethik im Kontext Künstlicher Intelligenz.

Impressum

Ansprechpartner:
Tobias Wangermann
Abteilung Wirtschaft und Innovation
Hauptabteilung Analyse und Beratung
T +49 30 / 26 996-3380
tobias.wangermann@kas.de

Herausgeberin:
Konrad-Adenauer-Stiftung e. V. 2020 Berlin

Umschlagfoto: © adobeStock/sdecoret; graphiCrash
Gestaltung und Satz: yellow too, Pasiek Horntrich GbR

Der Text dieses Werkes ist lizenziert unter den Bedingungen von „Creative Commons Namensnennung-Weitergabe unter gleichen Bedingungen 4.0 international", CC BY-SA 4.0 (abrufbar unter: https://creativecommons.org/licenses/by-sa/4.0/legalcode.de).

ISBN 978-3-95721-780-6

www.kas.de