Über das Verhältnis von Ethik und Algorithmen
Ein Problemaufriss

Michael Reder, Nicholas Müller, Robert Lehmann

Gegenwärtig wird intensiv darüber diskutiert, ob künstliche Intelligenz eine sehr spezifische menschliche Eigenschaft übernehmen kann, und zwar die, moralisch zu handeln (Floridi und Sanders 2004; Misselhorn 2018; Brieger 2018; Weber 2018). Hintergrund dieser Diskussion sind technische Entwicklungen, die der Maschine scheinbar mehr und mehr so etwas wie autonomes Handeln ermöglichen. Maschinen haben im Zuge dessen nicht nur automatisierte Prozesse übernommen, sondern durch die ihnen spezifische Trainings- und Lernfähigkeit auch neue Aktionsmöglichkeiten erworben, die handlungsanalog zu sein scheinen. Mit diesen Möglichkeiten zu eigenständigem Handeln wird nun die Frage relevant, ob die Maschine auch etwas tun kann, was oftmals nur dem Menschen zugeschrieben wird: und zwar moralisch zu handeln. Im Kern der Diskussionen steht die Frage, ob Maschinen moralisch kalkulieren, entscheiden und handeln können und ob ihnen deshalb auch Verantwortung für ihr Handeln zu geschrieben werden kann (Dignum 2018).

Katharina Zweig (2018) hat in diesem Zusammenhang auf einige grundlegende Probleme hingewiesen. Denn die Informatik braucht eine formalisierte Definition von normativen Begriffen. Erst dann können diese in einen digitalen Code übersetzt werden. Die philosophische Frage nach der Normativität ist jedoch selten so binär, wie die Informatik sie gerne hätte. Normative Konflikte sind oft sehr kompliziert, auch weil sie sich zeitlich verändern und deswegen nur selten in einer eindeutigen Heuristik konzeptualisiert werden können. Dies gilt auch für Normen selbst. Viele Normen sind oft deshalb so formal, weil nur so ihre Allgemeingültigkeit begründet werden kann. Wenn diese Normen material gefüllt werden, ist diese Eindeutigkeit allerdings begrenzt. Mit Blick auf verschiedene soziale, kulturelle oder zeitliche Kontexte einerseits und mit Blick auf den Einzelfall andererseits werden Normen oft unterschiedlich ma-

terial gefüllt. Diese Kontextualität und inhärente Dialektik von Normativität widersprechen der binären Logik der digitalen Technologie.

Der vorliegende Band will diese Frage nach dem Verhältnis von KI und ethischer Verantwortung von unterschiedlichen Seiten aus – teils anhand von Fallbeispielen, teils in systematischer Hinsicht – analysieren und kritisch diskutieren. Es geht darum, Chancen und Risiken beim Umgang mit KI-Technologien auszuloten und nach ethisch verantwortlichen Formen des Umgangs, v.a. in Konfliktsituationen, zu suchen.

Dem Band liegen die Arbeiten eines interdisziplinären Forschungsverbundes zugrunde.[1] In diesem geht es um ein Feld institutionellen Handelns angesichts von Konflikten mit massiver Reichweite (Gutwald et al. 2021). Dabei handelt es sich um die Bewertung von Fallakten durch Jugendämter im Hinblick auf Risiken der Kindeswohlgefährdung. Prinzipiell sieht die grundgesetzliche Regelung in Deutschland vor, dass es das Recht und die Pflicht der Eltern ist, ihre Kinder zu erziehen und sich um ihr Wohlergehen zu kümmern. Die Intervention durch staatliche Stellen ist nur dann vorgesehen, wenn eine Gefährdung des Kindeswohls vorliegt, die die Eltern nicht abwenden können oder wollen. Erst dann wird das Wächteramt des Staates relevant und in Verantwortungsgemeinschaft mit dem Familiengericht greift das Jugendamt als öffentlicher Träger der Jugendhilfe in die Familie ein. Die Jugendämter schlagen im begründeten Krisenfall verschiedene Maßnahmen zum Schutz des Kindes vor: von Beratungsdiensten bis zur Inobhutnahme des Kindes. Dazu wurde durch den Bundesgerichtshof konkretisiert, dass eine Kindeswohlgefährdung vorliegt, wenn »eine gegenwärtige, in einem solchen Maße vorhandene Gefahr [festgestellt werden kann], dass sich bei der weiteren Entwicklung eine erhebliche Schädigung mit ziemlicher Sicherheit voraussehen lässt.«(BGH FamRZ 1956: 350).

Der Forschungsverbund untersucht, ob und inwiefern normative Kriterien, die das Handeln von Jugendämtern leiten, in Algorithmen übersetzt werden können und ob digitale Tools das institutionelle Handeln unterstützen können (Gutwald und Reder 2023). Dies ist einerseits angesichts der Komplexität der Situation, des betroffenen Rechtskonflikts und der Reichweite

1 Dabei handelt es sich um den vom bidt finanzierten Forschungsverbund *Kann ein Algorithmus im Konflikt moralisch kalkulieren*, der von 2021–2023 an der Schnittstelle von Philosophie, Informatik und Sozialer Arbeit angesiedelt ist. Das Projekt ist eine Kooperation der Hochschule für Philosophie in München, der Technischen Hochschule Würzburg-Schweinfurt und der Technischen Hochschule Nürnberg Georg Simon Ohm.

der Entscheidung nicht unproblematisch. Es erscheint aber andererseits angesichts eben dieser Tragweite des Konflikts und begrenzter personeller und Zeitressourcen in den Institutionen auch sinnvoll zu fragen, ob digitale Systeme eine Hilfe sein können, um die Entscheidungen mit Bezug auf gesellschaftliche Normen transparent und fundiert zu treffen.

In einer Welt, in der künstliche Intelligenz und Algorithmen immer mehr Entscheidungen übernehmen, stellt sich mit Blick auf solche Beispielfelder und angesichts dieser Eigenart ethischer Urteilsfindung die Frage, inwieweit diese Systeme moralische und ethische Aspekte berücksichtigen können. Eine zentrale Frage ist, ob und wenn ja wie Algorithmen entwickelt werden könnten, die moralisch kalkulieren und in komplexen ethischen Situationen Entscheidungen treffen können. Die folgende Einleitung stellt einen Problemaufriss zu diesem Themenfeld vor dem Hintergrund der Forschungen des genannten Verbundes dar.

Zur Kalkulationsfähigkeit von Algorithmen im Kontext ethischer Entscheidungen

Aus technischer Perspektive gilt es als erstes zu betonen, dass innerhalb des Themenfeldes der KI zwischen den Basistechnologien und Anwendungen zu unterscheiden ist. Christen et al. (2020) postulieren in diesem Zusammenhang, dass vier Basisfunktionen, und zwar Mustererkennung, Klassifikation, Prognose und Synthese aus unterschiedlichen Input-Daten, z.B. Text, Bild oder Ton, einen Anwendungskontext erzeugen. Muster werden in Texten oder Bildern erkannt, Geräusche durch Klassifizierung als Vogelgeräusche präzisiert (Xie et al. 2023), das Wetter durch Datenaggregation vorhergesagt oder neue Musik durch Synthese generiert (Rutherford 2023). Entsprechend weisen Christen et al. (2020) darauf hin, dass durch KI-Technologien Systeme spezifiziert werden können, die mittels Berechnungen Artefakte generieren können. Um diese generierten Daten zu klassifizieren oder solche Systeme für die Gesellschaft nutzbar zu machen, wird jedoch weiterhin menschliche Expertise benötigt. Im Berufsfeld *Clickwork* werden zum Beispiel die Trainingsdaten generiert (Beuth et al. 2023). Im nächsten Schritt wird eine der Basisfunktionen genutzt, um einen neuen Output zu generieren, der wiederum von Menschen genutzt wird, z.B. beim autonomen Fahren. Dieser kann wiederum Ausgangspunkt für weitere menschliche Entscheidungen sein.

Gemeinsam ist diesen Anwendungen, dass sie Daten aggregiert auswerten und eine Schlussfolgerung aus diesen ziehen. Je nach Anwendungsszenario enthalten sowohl die Eingangsdaten als auch die Ergebnisdarstellung Aspekte, die aus ethisch-moralischer Sicht einer besonderen Prüfung bedürfen. Insbesondere dann, wenn diese bereits bei der rechnerbasierten Aggregation der Daten berücksichtigt werden müssen. Denn die Integration von moralischen Aspekten in statistischen Berechnungen ist eine komplexe Herausforderung. Um normative Kriterien in Algorithmen zu übersetzen, müssen sie zunächst quantifiziert werden. Ein Ansatz zur Quantifizierung von moralischen Aspekten besteht darin, ethische Prinzipien in messbare Indikatoren oder Kriterien zu übersetzen und diese in statistische Modelle einzubeziehen. Dabei ist zu berücksichtigen, dass derartige Modelle nur eine Annäherung an die moralische Abwägung darstellen und nicht alle ethischen Nuancen erfassen können. Diese Einschränkungen finden sich auch in Überblicksstudien zur Anwendung von KI-basierten Assistenz- oder Entscheidungssystemen.

Der Bericht von *AlgorithmWatch* (Matzat et al. 2019) leistet an dieser Stelle einen wesentlichen Beitrag, da die dort beschriebene Datenbank eine Übersicht zu vorrangig in Deutschland etablierten Algorithmen, Richtlinien und Akteuren darstellt. Insbesondere bei den Softwarebeschreibungen wird deutlich, dass oftmals Assistenzsysteme zur Anwendung kommen. Darüber hinaus sind Studien von Interesse, welche den konkreten Einsatz von Softwaresystemen zur Entscheidungsfindung beschreiben. Basierend auf den vorliegenden Recherchen wird dies insbesondere im Personal- und Bewerbungsmanagement eingesetzt. Hunkenschroer und Luetge (2022) betrachten beispielsweise 51 Studien mit Bezug zum Personalmanagement. Sie konstatieren, dass eine normative Einschätzung der Algorithmen bislang im wissenschaftlichen Diskurs unterrepräsentiert ist. Darüber hinaus schlussfolgern sie, dass die Diskussion zu ethischen Herausforderungen zu stark an Richtlinien ausgerichtet ist und nicht konkret auf spezifische Anwendungsfelder bezogen werden.

Eine weitere Überblicksstudie in diesem Bereich von Will, Krpan und Lordan (2023) untersucht, inwiefern Menschen oder Algorithmen hinsichtlich der Entscheidungsfindung besser, gleich oder schlechter handeln. Sie analysieren dazu die Effizienz, Performanz, Diversität und wie die Entscheidungsfindung durch einen Algorithmus wahrgenommen wird. In ihrer Zusammenfassung zeigen die Autoren auf, dass insbesondere bezogen auf Effizienz und Performanz die Algorithmen einen Vorteil haben und auch die Diversität besser gesichert wird. Die Wahrnehmung der Algorithmen-Entscheidungen wird

jedoch generell durch Menschen als kritisch eingeschätzt, was gegebenenfalls an fehlenden Erklärungsansätzen zur Funktion der Software oder der Entscheidungsfindung liegt.

Der Einsatz von KI in ethischen Entscheidungssituationen ist also nicht einfach. Bei allen Problemen, die damit verbunden sind, legen die Studien allerdings auch nahe, dass sie in Form von Assistenzsystemen sehr wohl Entscheidungsträger*innen in konfliktiven Situationen unterstützen können. Sie können beispielsweise Muster in großen Datenmengen erkennen, die für Menschen schwer zu erfassen sind, und so wertvolle Informationen für die Entscheidungsfindung liefern. Dabei ist es wichtig, die Rolle von Algorithmen als Unterstützung und Ergänzung menschlicher Expertise zu verstehen, anstatt sie als Ersatz zu betrachten. Denn menschliche Fähigkeiten wie Empathie, Intuition und Urteilsvermögen bleiben bei ethischen Entscheidungen zentral, da diese von Algorithmen nur schwer nachgebildet werden können. Daher sollten Algorithmen darauf abzielen, die menschliche Entscheidungsfindung zu verbessern, anstatt sie zu ersetzen, und die Transparenz und Nachvollziehbarkeit algorithmischer Entscheidungen sicherstellen. Dies beinhaltet auch die Implementierung von Mechanismen zur Überprüfung und Anpassung von Algorithmen, um sicherzustellen, dass sie ethischen Standards entsprechen und kontinuierlich verbessert werden.

Eine solche Strategie ethischer Verantwortung im Kontext von KI-Entwicklungen ist sinnvoll, da das Hauptproblem auf Algorithmen basierten Entscheidungssystemen darin besteht, dass sie einen gültigen Wert benötigen, um ein Ergebnis zu erzeugen. Fehlende Werte müssen interpoliert werden, was die Ergebnisse verfälschen kann. Eine Vielzahl von Studien beschäftigt sich mit den daraus resultierenden Effekten, z.B. Fragen der Gleichbehandlung verschiedener Gruppen. Dabei zeigt sich jedoch, dass es unter Umständen keine wirklich fairen Algorithmen geben kann (Kleinberg et al. 2016) oder nur ein zeitlich nachgelagerter Auswertungsschritt helfen kann, fehlende Daten zu korrigieren (Chouldechova 2016). Zu diesem Zeitpunkt ist die Entscheidung des Algorithmus jedoch längst gefallen, woraus wieder (problematische) gesellschaftlichen Folgen resultieren können.

Um eine Antwort auf dieses Problem zu finden, wurde im Rahmen des KAI-Mo-Projektes ein Drei-Agenten-Lösungsansatz entwickelt, auf den im weiteren Verlauf noch detailliert eingegangen wird. Ziel ist es, dass das *Agentenarray* aus unabhängigen Ansätzen Lösungen entwickelt und der eingeschlagene Weg transparent dokumentiert wird.

Damit erfüllt ein solcher Ansatz wesentliche Forderungen, die in verschiedenen Fachgremien zur Thematik der ethisch-moralischen Herausforderungen von KI und deren Einsatz in einem digitalen gesellschaftlichen Kontext diskutiert werden. So definiert beispielsweise die *High-level expert group on artificial intelligence* der Europäischen Kommission in ihrer *Assessment List for Trustworthy AI* (ALTAI) (AI HLEG 2020) sieben Punkte: (1) menschliches Handeln und Kontrolle, (2) technische Robustheit und Sicherheit, (3) Datenschutz und Datenmanagement, (4) Transparenz, (5) Vielfalt, Nichtdiskriminierung und Fairness, (6) Umwelt- und gesellschaftliches Wohlergehen, (7) Verantwortlichkeit. Der Deutsche Ethikrat wiederum bezeichnet Entscheidungsunterstützungssysteme in der Stellungnahme *Mensch und Maschine – Herausforderungen durch Künstliche Intelligenz* (Deutscher Ethikrat 2023) als begrüßenswert, um die menschliche Entscheidungsfindung durch einen verbesserten Zugang zu Daten zu unterstützen. Eingeschränkt wird die Empfehlung jedoch durch den Hinweis, dass diese Systeme stets das Ziel verfolgen müssen, die menschliche Entscheidungsfindung zu verbessern und nicht die Effizienzsteigerung oder Personaleinsparung durch den Einsatz von Computersystemen voranzutreiben. Darüber hinaus wird in der Stellungnahme auf die potentielle Gefahr hingewiesen, dass Menschen eine Computerentscheidung unreflektiert übernehmen. Folgerichtig wird in diesem Zusammenhang empfohlen, dass »geeignete technische und organisatorische Instrumente zur Vorkehrung gegen die manifeste Gefahr eines Automation Bias bereitgestellt werden, die es den Fachkräften erschweren […] der algorithmischen Entscheidungsempfehlung unbesehen zu folgen.« (Deutscher Ethikrat 2023: 249) Diese Bedingung gilt für alle gesellschaftlichen Felder, in denen KI-Systeme zum Einsatz kommen. Es geht in diesem Zusammenhang beispielsweise um die Notwendigkeit, den Datenschutz und die Privatsphäre der betroffenen Personen zu wahren, sowie die Sicherstellung, dass die Algorithmen (beispielsweise durch die Zusammenstellung der Trainingsdaten) bestehende soziale Ungleichheiten nicht verstärken.

Dementsprechend ist es wichtig, die Ergebnisse und Anforderungen der Informatik in Bezug auf ethische Algorithmen kontinuierlich zu evaluieren und in den Dialog mit den anderen beteiligten Disziplinen einzubringen. Hierzu gehört auch die Bereitschaft, die entwickelten Systeme und Ansätze kritisch zu hinterfragen und gegebenenfalls anzupassen, um die bestmögliche Unterstützung für die Entscheidungsfindung in der Sozialen Arbeit zu gewährleisten.

Digitale Assistenzsysteme in der Sozialen Arbeit – ein Beispielfeld

In der Sozialen Arbeit ist die Entscheidungsunterstützung durch KI grundsätzlich ein sehr kontroverses Thema. Anhand der Debatte um die Entscheidungsfindung zur Kindeswohlgefährdung kann dies besonders deutlich illustriert werden, da hier die Fachkräfte im Jugendamt gewichtige Entscheidungen über die Zukunft junger Menschen und ihrer Familien treffen müssen (Bathke et al. 2019).

Schon lange vor der Entwicklung von Systemen der KI in diesem Bereich war die Soziale Arbeit auf der Suche nach Instrumenten, die die Prognosequalität in Kinderschutzfällen verbessern könnten. Dabei waren zwei große Linien zu erkennen: Einerseits zeigte sich in der Praxis, dass Fachkräfte mit Erfahrung deutlich bessere Prognosen generieren, als Personen, die neu in dem Feld arbeiten. Die große Bedeutung des Erfahrungswissens wird nicht zuletzt im §8a SGB VIII deutlich, der vorschreibt, dass Träger, die Leistungen im Kinder- und Jugendbereich erbringen, eine »insofern erfahrene Fachkraft« (§8a SGB VIII, Abs. 4, Nr. 2) bei Fällen einer vermuteten Kindeswohlgefährdung hinzuziehen müssen.

Andererseits spielen neben dem individuellen Erfahrungswissen in Fällen der Kindeswohlgefährdung wissenschaftlich abgesicherte Erkenntnisse eine besondere Rolle. In umfangreichen Studien konnten empirisch Faktoren herausgearbeitet werden, die das Risiko einer Kindeswohlgefährdung erhöhen oder vermindern. Aufbauend auf diesen Ergebnissen wurden mit statistischen Verfahren Modelle entwickelt, die unter Berücksichtigung der relevanten Faktoren mit transparenten Algorithmen Prognosen zur Risikostruktur liefern. Für die Praxis der Sozialen Arbeit wurden sie in Checklisten übersetzt, die teilweise in Papierform, teilweise digital bearbeitet werden können (Ackermann 2020).

Es ist sehr bemerkenswert, dass einerseits in vielen methodisch hochwertigen Metastudien herausgearbeitet wurde, dass diese Instrumente eine deutlich höhere Prognosequalität aufweisen als Fachkräfte, die intuitiv-diskursiv agieren (Grove et al. 2000; van der Put et al. 2017), andererseits im Fachdiskurs dennoch der Standpunkt vertreten wird, dass Ansätze, die eine individuelle Expertise der Fachkräfte in den Vordergrund stellen, zu favorisieren seien (Bastian 2012; Schroth 2021). Hier wird deutlich, dass in der Disziplin der Sozialen Arbeit auf der inhaltlichen Ebene eine große Sorge vor einer De-Professionalisierung durch die Nutzung standardisierter Verfahren besteht (Schrödter et al. 2020). Insofern überrascht es nicht, dass die Anwendung von KI-Sys-

temen im deutschen Kinderschutz sehr kritisch diskutiert wird und die meisten Autor*innen eher die Risiken der Technologie betonen (Görder 2021).

Im internationalen Kontext liegen dagegen bereits einige Ansätze vor, die größere Datenbestände und komplexere Algorithmen für die Risikoprognostik nutzen (La Valle et al. 2016). In Neuseeland wurde z.B. ein *Predictive Risk Modelling to Prevent Child Maltreatment* (PRM) entwickelt. Hier wurden Daten aus den verschiedenen staatlichen Systemen, wie der Sozial- und Kinderfürsorge und dem Gesundheits- und Erziehungssystem verknüpft und darauf aufbauend ein Risikoscore ermittelt (Gillingham 2021). Einen ähnlichen Ansatz verfolgen die Kinderschutzbehörden in den USA. Dort wird z.B. im *Allegheny Family Screening Tool* (AFST) eine Verknüpfung unterschiedlicher Datenquellen zur Berechnung eines *Familien-Screening-Scores* verwendet. Bei einem hohen Score-Wert wird ein höheres Risiko der Kindeswohlgefährdung angenommen und entsprechende Interventionen eingeleitet (Holstein 2022).

Die praktische Anwendung dieser Verfahren zeigt jedoch einige grundlegende Probleme, die mit den skizzierten ethischen Bedenken in Zusammenhang stehen. So zeigte sich, dass das PRM einen sehr starken Zusammenhang zwischen Armut und dem Merkmal ›Alleinerziehend‹ und dem Risiko einer Kindeswohlgefährdung herstellt. Dies lässt den Rückschluss zu, dass in den verwendeten Daten Fälle mit dieser Kombination von Merkmalen häufig vorkamen. Die Interpretation, dass Armut und Alleinerziehend ursächlich für die Kindeswohlgefährdung sind, ist allerdings eine Verwechslung von Korrelation und Kausalität. Aufgrund dieser Problematik wurden Entwicklung und Implementation dieses Verfahrens eingestellt. Ähnliche Vorwürfe wurden gegen das AFST vorgebracht, das ebenfalls vor allem als Armuts-Profiling funktioniert hat (Eubanks 2018).

Die internationalen Erfahrungen machen deutlich, dass die Einführung KI-basierter Risikoprognoseverfahren mit vielfältigen Problemen einhergehen. Eine besondere Rolle kommt den verwendeten Trainingsdaten zu. In den genannten Anwendungen wurden die Algorithmen mit Datensätzen trainiert, in denen durch die Auswahl der betroffenen Personen oder durch implizite Stereotype bei den bearbeitenden Personen Verzerrungen von den Algorithmen erlernt wurden. Bei zukünftigen Entwicklungen ist es daher entscheidend, schon bei der Wahl des Trainingsmaterials auf die Einhaltung von Gerechtigkeitsprinzipien und eine diskriminierungsfreie Auswahl der Daten zu achten (Görder 2021).

Für die Soziale Arbeit in Deutschland stellt sich die Frage, wie dieses Trainingsmaterial aufgebaut sein müsste, um Algorithmen zu trainieren, die Er-

gebnisse produzieren, die sowohl frei von Diskriminierung als auch mit hoher Vorhersagequalität ausgestattet sind. Ausgehend von dem bestehenden Widerspruch im Fachdiskurs zwischen statistischen Prognosemodellen und erfahrungsbasierten Ansätzen könnte der Einsatz von KI-Systemen hier zu einer Art Synthese führen. Bisher bestand bei der Nutzbarmachung des Erfahrungswissens der Fachkräfte das Problem, dass das zugrunde liegende implizite Wissen nur schwer explizierbar war (Böhle 2020). Da dieses Wissen jedoch die Grundlage von Entscheidungen darstellt, die in Akten dokumentiert sind, wird die Nutzung solcher prozessgenerierten Textdokumente in der Professionsforschung der Sozialen Arbeit schon länger diskutiert.

Mit den klassischen qualitativen Analyseverfahren konnten hier bisher keine umfangreichen Studien durchgeführt werden (Lehmann und Klug 2019). Verfahren aus dem KI-Forschungsbereich können neue Erkenntnisse generieren. So könnten die großen Textmengen, die in deutschen Jugendämtern vorliegen mit den verschiedenen Verfahren untersucht und entsprechende Muster aus den Daten extrahiert werden. Es ist anzunehmen, dass das bisher implizite Erfahrungswissen der Fachkräfte auf diese Weise zumindest ansatzweise verarbeitet und zum Ausdruck gebracht werden könnte (Vladova et al. 2019). Es besteht eine große Chance, mit maschinellen Lernverfahren aus der dokumentierten Fachlichkeit in Akten die Grundlagen der Entscheidungsfindung erfahrener Fachkräfte zu erlernen und darauf aufbauend Assistenzsysteme zu generieren.

Auch wenn dieser Ansatz zunächst sehr naheliegend erscheint, birgt er jedoch ebenfalls große Risiken. So sind in den Akten der deutschen Jugendämter ähnlich wie in USA und Neuseeland bestimmte Bevölkerungsgruppen überrepräsentiert (Jugendinstitut eV et al. 2020). Auch die Arbeit im Jugendamt vor Ort ist nicht frei von Vorurteilen und Ressentiments, die sich in der fachlichen Entscheidungsfindung widerspiegeln (Harrer-Amersdorffer 2022; Herzog 2022). Es kann also nicht empfohlen werden, bestehende Daten unreflektiert zum Training einer KI zur Entscheidungsunterstützung zu nutzen. Als Vorstufe wäre eine Mustererkennung durch maschinelle Lernverfahren in den Dokumenten bestehender Praxis allerdings sehr hilfreich. Die erkannten Muster würden sichtbar und könnten damit Gegenstand einer fachlichen Debatte sein. Aufbauend auf einer in diesem Sinne von sozialarbeiterischer Fachlichkeit geprägten Entwicklungsarbeit ist ein System zur Entscheidungsunterstützung in Kinderschutzfällen sehr wünschenswert und denkbar (Linnemann et al. 2023; Steiner und Tschopp 2022).

Abgesehen von der inhaltlichen Qualität von KI-gestützten Assistenzsystemen stellt sich die Frage ihrer Einbindung in die Entscheidungspraktiken der Jugendämter. Selbst wenn ein solches Verfahren extrem gute Prognoseergebnisse liefern sollte, sehen einerseits die geltende Rechtslage, andererseits auch tiefgreifende ethische Erwägungen (Deutscher Ethikrat 2023) vor, dass so lebensentscheidende Entscheidungen in Letztverantwortung von Menschen getroffen werden müssen. Daher muss auch hier überlegt werden, wie eine sinnvolle Integration in Entscheidungsprozesse aussehen kann. Dabei ist im Kontext der Sozialen Arbeit zu beachten, dass hier seit langem eine hohe Skepsis gegenüber digitaler Technologie vorliegt (Bertsche und Como-Zipfel 2017), sodass ein entsprechendes Unterstützungssystem evtl. weniger stark genutzt werden würde als in anderen Bereichen. Gleichzeitig ist bekannt, dass Menschen dazu neigen, ihrer eigenen Expertise weniger zu vertrauen, als einem maschinellen System und im Zweifel eher der Einschätzung einer Maschine folgen, selbst wenn sie dem Urteil der Maschine nicht völlig zustimmen (Banbury 2021; Gapski 2020). Daher besteht das Risiko, dass einem KI-System zu viel Vertrauen entgegengebracht wird. Die Einbindung eines KI-Systems in einen Entscheidungsfindungsprozess muss also sowohl sicherstellen, dass die Expertise des Systems in ausreichender Intensität berücksichtigt wird, als auch verhindern, dass die Menschen unreflektiert der Einschätzung des technischen Systems folgen.

Aus der derzeitigen Perspektive sind daher einfache Empfehlungssysteme im Kinderschutz, die sich deutlich für bestimmte Maßnahmen aussprechen, sehr kritisch zu sehen. Aktuell existiert im deutschsprachigen Raum kein KI-System, das mit einem Datensatz trainiert wurde, der den Ansprüchen an die inhaltliche Qualität und Diskriminierungsfreiheit genügt. Weiterhin liegt in der Interpretation der Ergebnisse und der Interaktion zwischen Mensch und Maschine aktuell ein großes Fehlerpotenzial. Daher erscheint es sinnvoll, einerseits die Entwicklung fachlich kontrollierter und diskriminierungsfreier Systeme voranzutreiben und andererseits Konzepte zu entwickeln, die eine zielführende Interaktion von Mensch und Maschine ermöglichen. Ein Ansatz könnte eine sehr defensive Integration einer KI sein, die nie eine eigene Empfehlung ausspricht, sondern den Menschen bei seiner Entscheidungsfindung durch gut aufbereitete Informationen und fachliche Hinweise maßgeblich unterstützt.

Ausblick

Vor dem Hintergrund der skizzierten Problemstellung versammelt der Band Beiträge aus dem Feld der Sozialen Arbeit – aber auch anderen Themenfeldern, wie u.a. der Medizin oder der Politik. Die Beiträge rekonstruieren soziale Praktiken und Institutionen, in denen Algorithmen bereits zum Einsatz kommen oder dies für die Zukunft geplant ist. In dem Band bieten Expert*innen aus verschiedenen Fachrichtungen fundierte Einsichten in die KI-gestützte Entscheidungs- und Urteilsfindung. Von der digitalen Operationalisierung über die Rolle des Menschen im Zentrum des technischen Fortschritts bis hin zur Konzeption von vertrauenswürdigen Systemen. Anhand dieser Beispiele werden sowohl Chancen als auch Grenzen des Einsatzes von KI-Systemen diskutiert, v.a. in hoch konfliktiven Situationen. Es geht letztlich um die Frage nach der ethischen Verantwortung im digitalen Zeitalter. Für die Diskussion dieser Frage will der Band einige grundlegende Impulse geben.

Literatur

Ackermann, Timo. 2020. »Digitalisierung in der Kinder- und Jugendhilfe und im Kinderschutz: Von Risikoeinschätzungsbögen über Fallbearbeitungssoftware bis zu Big Data.« *Soziale Passagen*, 12 (1): 171–177.

AI HLEG. 2020. »Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment | Shaping Europe's digital future.« Accessed October 6, 2023. https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

Awad, Edmond, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff et al. 2018. »The Moral Machine experiment.« *Nature* 563 (7729): 59–64. https://doi.org/10.1038/s41586-018-0637-6.

Bastian, Pascal. 2012. »Die Überlegenheit statistischer Urteilsbildung im Kinderschutz–Plädoyer für einen Perspektivwechsel hin zu einer angemessenen Form sozialpädagogischer Diagnosen.« In *Rationalitäten des Kinderschutzes*, edited by Thomas Marthaler, Pascal Bastian, Ingo Bode, Mark Schrödter, 249–267. Wiesbaden: Springer.

Bathke, Sigrid A., Milena Bücken, Dirk Fiegenbaum. 2019. »Die Grundlagen: Kinderschutz, Kindeswohl und Kindeswohlgefährdung aus rechtlicher und fachlicher Perspektive.« In *Praxisbuch Kinderschutz interdisziplinär,*

edited by Sigrid A. Bathke, Milena Bücken, Dirk Fiegenbaum, 5–106. Wiesbaden: Springer VS.

Bertsche, Oliver, Frank Como-Zipfel. 2017. »Sozialpädagogische Perspektiven auf die Digitalisierung.« *Soziale Passagen*, 8(2): 235–254. https://doi.org/10.1007/s12592-016-0244-z.

Beuth, Patrick, Heiner Hoffmann, Max Hoppenstedt. 2023. »Das sind die Menschen hinter der KI-Revolution.« Accessed October 6, 2023. https://www.spiegel.de/netzwelt/web/clickwork-und-content-moderation-die-gesichter-hinter-der-kuenstlichen-intelligenz-a-9629ea15-5bc3-42bd-a236-199d606b1a24.

Böhle, Fritz. 2020. »Implizites Wissen und subjektivierendes Handeln – Konzepte und empirische Befunde aus der Arbeitsforschung.« In *Implizites Wissen: Berufs- und wirtschaftspädagogische Annäherungen*, edited by Rico Hermkes, Georg Hans Neuweg, Tim Bonowski, 36–64. Bielefeld: Wbv.

Brieger, Julchen. 2018. »Über die Unmöglichkeit einer kantisch handelnden Maschine.« In *Maschinenethik. Normative Grenzen autonomer Systeme,* edited by Matthias Rath, Friedrich Krotz und Matthias Karmasin, 107–120. Wiesbaden: Springer Fachmedien.

Chouldechova, Alexandra. 2016. »Fair prediction with disparate impact: A study of bias in recidivism prediction instruments.« https://doi.org/10.48550/ARXIV.1610.07524.

Christen, Markus et al. 2020. »Wenn Algorithmen für uns entscheiden: Chancen und Risiken der künstlichen Intelligenz.« Zürich: Vdf Hochschulverlag AG an der ETH Zürich.

Deutscher Ethikrat. 2023. »Mensch und Maschine – Herausforderungen durch Künstliche Intelligenz.« Accessed October 6, 2023. https://www.ethikrat.org/fileadmin/Publikationen/Stellungnahmen/deutsch/stellungnahme-mensch-und-maschine.pdf.

Dietrich, Eric. 2011. »Homo Sapiens 2.0. Building the Better Robots of Our Nature.« In *Machine ethics,* edited by Michael Anderson, Susan Leigh Anderson, 531–538. Cambridge: Cambridge university press.

Dignum, Virginia. 2018. »Ethics in artificial intelligence: introduction to the special issue.« *Ethics and information technology* 20 (1): 1–3. https://doi.org/10.1007/s10676-018-9450-z.

Eubanks, Virginia. 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor*. New York: St. Martin's Press.

Floridi, Luciano, John W. Sanders. 2004. »On the Morality of Artificial Agents.« *Minds and Machines* 14 (3): 349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d.

Gapski, Harald. 2020. »Digitale Transformation: Datafizierung und Algorithmisierung von Lebens-und Arbeitswelten.« In *Handbuch Soziale Arbeit und Digitalisierung*, edited by Nadia Kutscher, Thomas Ley, Udo Seelmeyer, Friederike Siller, Angela Tillmann, Isabel Zorn, 156–166. Weinheim: Beltz Juventa.

Gillingham, Philip. 2021. »Practitioner perspectives on the implementation of an electronic information system to enforce practice standards in England.« *European Journal of Social Work* 24 (5): 761–771. https://doi.org/10.1080/13691457.2020.1870213.

Görder, Björn. 2021. »Die Macht der Muster. Die Ethik der Sozialen Arbeit vor professionsbezogenen und gesellschaftlichen Herausforderungen durch, künstliche Intelligenz ›.« *Ethik Journal* 7 (2). Accessed October 6, 2023. https://www.ethikjournal.de/fileadmin/user_upload/ethikjournal/Texte_Ausgabe_2021_2/Goerder_Ethikjournal_2.2021.pdf.

Grove, William M., David H. Zald, Boyd S. Lebow, Beth E. Snitz, Chad Nelson. 2000. »Clinical versus mechanical prediction: A meta-analysis.« *Psychological Assessment*, 12 (1): 19–30. https://doi.org/10.1037/1040-3590.12.1.19.

Gutwald, Rebecca, Michael Reder. 2023. »How to Protect Children? A Pragmatic Approach: On State Intervention and Children's Welfare.« *The Journal of Ethics* 27 (1): 77–95. https://doi.org/10.1007/s10892-022-09416-3.

Gutwald, Rebecca, Jennifer Burghardt, Maximilian, Kraus, Michael Reder, Robert Lehmann, Nicholas Müller. 2021. »Soziale Konflikte und Digitalisierung. Chancen und Risiken digitaler Technologien bei der Einschätzung von Kindeswohlgefährdungen.« *Ethik Journal* 7 (2). Accessed October 6, 2023. https://www.ethikjournal.de/fileadmin/user_upload/ethikjournal/Texte_Ausgabe_2021_2/Gutwald_u.a._Ethikjournal_2.2021.pdf.

Harrer-Amersdorffer, Jutta. 2022. *Fachliches Handeln in der Fallarbeit: Eine empirische Studie über den Stand der Sozialpädagogischen Familienhilfe*. Leverkusen: Verlag Barbara Budrich.

Herzog, Lucas-Johannes. 2022. »Rassismus im Jugendamt: Vom Nachdenken über eine nicht geführte Debatte.« *Forum Erziehungshilfen*, 10 (1): 11–13. https://doi.org/10.3262/FOE2201011.

Holstein, Kenneth. 2022. »What happens when human workers oversee algorithmic tools?« Accessed October 6, 2023. https://medium.com/@ken

neth.holstein/what-happens-when-human-workers-oversee-algorithmic-tools-bbfc32e8ce61.

Hunkenschroer, Anna Lena, Christoph Luetge. 2022. »Ethics of AI-Enabled Recruiting and Selection: A Review and Research Agenda.« *Journal of Business Ethics*, 178 (4): 977–1007. https://doi.org/10.1007/s10551-022-05049-6.

Jaume-Palasí, Lorena, Matthias Spielkamp. 2017. »Ethik und algorithmische Prozesse zur Entscheidungsfindung oder -vorbereitung.« Accessed October 6, 2023. https://algorithmwatch.org/de/wp-content/uploads/2017/06/AlgorithmWatch_Arbeitspapier_4_Ethik_und_Algorithmen.pdf.

Jugendinstitut e.V., Deutsche, Susanne Lochner, Alexandra Jähnert. 2020. *DJI-Kinder-und Jugendmigrationsreport 2020: Datenanalyse zur Situation junger Menschen in Deutschland*. Bielefeld: Wbv.

Kleinberg, Jon, Sendi Mullainathan, Manish Raghavan. 2016. »Inherent Trade-Offs in the Fair Determination of Risk Scores.« https://doi.org/10.48550/ARXIV.1609.05807.

Kucklick, Christoph. 2016. »Soziologische Aspekte von Big Data.« Audioprotokoll, Deutscher Ethikrat, March 23, 2016. https://www.ethikrat.org/sitzungen/2016/big-data.

La Valle, Ivana, Berni Graham, Lisa Payne. 2016. »*A consistent identifier in education and children's services*.« Department for Education. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/534744/Consistent_identifier__report_July_2016.pdf.

Lehmann, Robert, Wolfgang Klug. 2019. »Die prozessorientierte Aktenanalyse.« In *Sekundäranalysen in der Kinder- und Jugendhilfe* edited by Maik-Carsten Begemann, Klaus Birkelbach, 301–319. Wiesbaden: Springer.

Linnemann, Gesa Alena, Julian Löhe, Beate Rottkemper. 2023. »Bedeutung von Künstlicher Intelligenz in der Sozialen Arbeit.« *Soziale Passagen* 15: 197–211. https://doi.org/10.1007/s12592-023-00455-7.

Lukesch, Helmut. 2006. FEPAA. Fragebogen zur Erfassung von Empathie, Prosozialität, Aggressionsbereitschaft und aggressivem Verhalten. Göttingen: Hogrefe.

Matzat, Lorenz, Lukas Zielinski, Miriam Cocco, Kristina Penner, Matthias Spielkamp, Sebastian Gießler, Sebastian Lang, Veronika Thiel. 2019. »Atlas der Automatisierung/Automatisierte Entscheidungen und Teilhabe in Deutschland.« https://atlas.algorithmwatch.org/wp-content/uploads/2019/07/Atlas_der_Automatisierung_von_AlgorithmWatch.pdf.

Misselhorn, Catrin. 2018a. Grundfragen der Maschinenethik. Ditzingen: Reclam.

Misselhorn, Catrin. 2018b. »Können und sollen Maschinen moralisch handeln?« *APuZ* 68 (6–8): 29–33.

Rutherford, Nichola. 2023. »Drake and The Weeknd AI song pulled from Spotify and Apple.« BBC News. Accessed November 9, 2023. https://www.bbc.com/news/entertainment-arts-65309313.

Schrödter, Mark, Pascal Bastian, Brian Taylor. 2020. »Risikodiagnostik und Big Data Analytics in der Sozialen Arbeit.« In *Handbuch Soziale Arbeit und Digitalisierung*, edited by Nadia Kutscher, Thomas Ley, Udo Seelmeyer, Friederike Siller, Angela Tillmann, Isabel Zorn, 255–264. Weinheim: Beltz-Juventa.

Schroth, Emma. 2021. »Digitale Falldokumentation im Jugendamt.« *Sozial Extra*, *45*(1): 49–52. https://doi.org/10.1007/s12054-020-00350-y.

Shevat, Amir. 2017. Designing Bots—Creating Conversational Experiences. Sebastopol: O'Reilly.

Steiner, Oliver, Dominik Tschopp. 2022. »Künstliche Intelligenz in der Sozialen Arbeit.« *Sozial Extra*, *46*(6): 466–471. https://doi.org/10.1007/s12054-022-00546-4.

van der Put, Claudia E., Mark Assink, Noëlle F Boekhout van Solinge. 2017. »Predicting child maltreatment: A meta-analysis of the predictive validity of risk assessment instruments.« *Child abuse & neglect*, *73*: 71–88. https://doi.org/10.1016/j.chiabu.2017.09.016.

van Wynsberghe, Aimee, Scott Robbins. 2019. »Critiquing the Reasons for Making Artificial Moral Agents.« *Science and engineering ethics* 25 (3): 719–735. https://doi.org/10.1007/s11948-018-0030-8.

Vladova, Gergana, Norbert Gronau, Leo Sylvio Rüdian. 2019. »Wissenstransfer in Bildung und Weiterbildung: Der Beitrag Künstlicher Intelligenz.« In *Digitale Transformation – Gutes Arbeiten und Qualifizierung Aktiv Gestalten*, edited by Dieter Spath, Birgit Spanner-Ulmer, 89–106. Berlin: Gito-Verlag.

Waldrop, Mitchell. 1987. »A question of responsibility.« *The AI Magazine* 8 (1): 28. https://doi.org/10.1609/aimag.v8i1.572.

Wallach, Wendell, Colin Allen. 2009. Moral machines. Teaching robots right from wrong. Oxford: Oxford University Press.

Weber, Karsten. 2018. »Autonomie und Moralität als Zuschreibung. Über die begriffliche und inhaltliche Sinnlosigkeit einer Maschinenethik.« *Maschinenethik. Normative Grenzen autonomer Systeme*, edited by Matthias Rath, Friedrich Krotz und Matthias Karmasin, 193–210. Wiesbaden: Springer Fachmedien.

Wickens, Christopher D., William S. Helton, Justin G. Hollands, Simon Banbury. 2021. *Engineering Psychology and Human Performance*. New York: Routledge.

Will, Paris, Dario Krpan, Grace Lordan. 2023. »People versus machines: Introducing the HIRE framework.« *Artificial Intelligence Review* 56 (2): 1071–1100. https://doi.org/10.1007/s10462-022-10193-6.

Xie, Jiangjian, Yujie Zhong, Junguo Zhang, Shuo Liu, Changqing Ding, Andreas Triantafyllopoulos. 2023. »A review of automatic recognition technology for bird vocalizations in the deep learning era.« *Ecological Informatics* no. 73(March): 101927. https://doi.org/10.1016/j.ecoinf.2022.101927.

Zweig, Katharina Anna, Georg Wenzelburger, Tobias D. Krafft. 2018. »On Chances and Risks of Security Related Algorithmic Decision Making Systems.« *European Journal for Security Research* 3(1): 181–203. https://doi.org/10.1007/s41125-018-0031-2.