Sakata, Tomoki

Warum sich Ethik und KI gegenseitig ergänzen müssen

In:

Düchs, Martin; Illies, Christian; Sakata, Tomoki (Hrsg.), Smart in the City: eine ethische Handreichung für die Digitalisierung der Stadt, Bamberg : University of Bamberg Press, S. 65-78. 2023. DOI: 10.20378/irb-93383

Beitrag im Sammelwerk - Verlagsversion

DOI des Beitrags: 10.20378/irb-94752

Datum der Veröffentlichung: 18.04.2024

Kapitel 3.

Warum sich Ethik und KI gegenseitig ergänzen müssen

Tomoki Sakata 0000-0002-1850-7809

Es gibt gegenwärtig zahlreiche Versuche, die KI ethisch zu bauen, insbesondere da, wo die Transparenz und die Erklärbarkeit als ethische Werte proklamiert werden (vgl. High-Level Expert Group on Artificial Intelligence, 2019). Wir fragen aber hier im umgekehrten Sinne, ob eine Ethik, die in diesem digitalen Zeitalter relevant sein will, der KI bedarf oder nicht. Gegen die intuitive Annahme, dass dies nicht der Fall ist, ist es das Ziel dieses Abschnitts, zu demonstrieren, dass sich die KI-Bedürftigkeit der Ethik von allein ergibt, wenn unsere ethische Idee ihre optimale, ihrer Komplexität gerechte Realisierung erfahren soll, was wiederrum selbst ein ethisches Ziel darstellt.

1. Theorie-Empirie-Dualismus der Ethik

> Man sieht doch auch nicht, dass man bloß aus Büchern ein Arzt wird. Gleichwohl suchen die medizinischen Schriftsteller nicht bloß die Heilmittel anzugeben, sondern auch das Heilverfahren, das man beobachten und die Behandlung, die man den einzelnen Patienten mit Rücksicht auf ihre besondere Konstitution [Habitus] angedeihen lassen muss. – Aristoteles, Nikomachische Ethik[28]

Eine Theorie ist weitgehend nutzlos, wenn sie die konkreten Fragen übersieht, welche sie zu erklären, bestätigen, falsifizieren, oder korrigieren haben. Die Gesundheit ist das allgemeine Telos bzw. das Grundprinzip der Medizin, obwohl mein gesunder Zustand niemals mit dem einer anderen Person völlig kongruieren muss. Aristoteles gebraucht oft solche medizinischen Gleichnisse in seinen biologischen Betrachtungen,

[28] Aristoteles, 1911, S. 223, 1181b.

um auf die Ambivalenz des Allgemeinen und Besonderen hinzuweisen, welche auch die Ethik wie die Politik gleichermaßen erschwert. Mit dem Blick auf die Naturnotwendigkeit lässt sich, so schreibt Aristoteles an einer anderen Stelle (S. 119), ein sicheres theoretisches Wissen erlangen. Im schroffen Gegensatz dazu muss man sich nach Aristoteles im Bereich der Ethik, wo von dergleichen strikter Gesetzmäßigkeit nicht die Rede ist, mit persönlichen Tugenden (je nach dem Habitus/Charakter) zufriedengeben. Nietzsche (1894, S. 257) äußerte ebenso, dass sowohl hinsichtlich der Freiheit (als Moralkonzept), als auch in Bezug auf die Gesundheit „kein allgemein gültiger Begriff" zu postulieren sei, weil jeder „ein eigenes, nur einmaliges Ding ist, das zu allen anderen Dingen eine neue, nie dagewesene Stellung einnimmt". Was wir bisher in unserer Handreichung als „Grundprinzip" und „Bereichswerte" theoretisch konzipiert und postuliert haben, wird hier gründlich herausgefordert.

Dennoch werden wir oder sollten wir zumindest gewahr sein, dass unser gewöhnliches Leben bereits von verschiedenen „Werten" und somit unterschiedlichen „Auswertungen" umgeben ist: Während der vergangenen Covid-19-Pandemie stand ein Schnelltest auf der Tagesordnung, dessen eindeutiges Ergebnis – positiv oder negativ – uns einen besseren Umgang mit Viren ermöglicht hat. Oder denken wir an die Blutuntersuchung, die wir zur gesundheitlichen Abklärung oder Diagnose veranlassen. Obwohl dieser Prozess deutlich komplexer ist, da hier nicht nur einzelne chemisch messbare Werte wie Eisen, Blutzucker usw. festgestellt werden, sondern auch auf das daraus folgende gesamte Krankheitsbild geschlossen werden kann, bleibt der Zweck der Messung unverändert, nämlich die Optimierung künftiger Handlungen. Während es sich bei den genannten Beispielen um die naturwissenschaftliche Voraussage handelt, können im nicht-physikalischen oder nicht-physiologischen Bereich auch universale Werte als Orientierung angewandt werden, selbst wenn der oben von Aristoteles konstatierte Unterschied immer noch besteht und wir darüber hinaus zugeben müssen, dass die Ethik noch schwieriger zu kodifizieren ist. Oder genau deshalb weil die Ethik schwer fassbar ist, können dergleichen Codes als Handlungsanweisungen von großem Belang sein. Als Beispiel lässt sich die „Allgemeine Erklärung der Menschenrechte" nennen, die als Indizien für die Achtung der Menschwürde solche Prinzipien wie Freiheit, Si-

cherheit, Anerkennung, Gleichheit usw. zur Geltung bringt.[29] Im philosophischen Diskurs arbeitet unter anderem Nussbaum (1993, 263-265; 2011, 33f.) an begrenzten Listen von bestimmten Eigenschaften oder Grundfähigkeiten des Menschen, z. B. Sterblichkeit, Leiblichkeit, Schmerz- und Lustempfindung, Emotion, Moralität usf. Diese Ansätze werden hier derart gedeutet und weiterentwickelt, dass wir die Würde des Menschen, die grundlegend, aber zu umfangreich ist, ebenso mittels vorab festgelegter Variablen (d.i. unsere „Bereichswerte") konkretisieren und messen. Wie beim Beispiel der Krankheitsdiagnose muss der Status Quo zuerst unter bestimmten Gesichtspunkten evaluiert werden, um überhaupt die Möglichkeit einer Besserung / Verbesserung buchstabieren zu können. In diesem Sinne ist unsere Vorgehensweise an die platonische Lehre angelehnt, nach der das richtige und ausgeglichene Maß dem Schönen und der Tugend gleichsteht.[30] Unsere Bereichswerte wie Privatheit und Solidarität sind Maßstäbe, anhand derer Probleme abgebildet und erst erkennbar gemacht werden. Wenn ein Projekt z. B. zu sehr auf den Datenschutz achtet und dadurch die soziale Interaktion unterschätzt, wird diese Einsicht uns zur Abwägung der Werte führen. *Jede Forschung ist das Vergleichen*, so sagt bereits Nikolaus von Kues im 15. Jahrhundert.[31] Und jeder Vergleich bedarf eines verlässlichen Maßstabs, den wir hier entwickeln und in die Tat umsetzen wollen.

2. Universale Werte und partikulare Wertschätzungen

> Sehen wir immerfort nur das Geregelte, so denken wir, es müsse so sein, von jeher sei es also bestimmt und deßwegen stationär. Sehen wir aber die Abweichungen, Mißbildungen, ungeheure Mißgestalten, so erkennen wir: daß die Regel zwar fest und ewig, aber zugleich lebendig sei; daß die Wesen, zwar nicht aus derselben heraus, aber doch innerhalb derselben sich in's Unförmliche

[29] Vgl. UN, Die Allgemeine Erklärung der Menschenrechte. https://www.ohchr.org/en/human-rights/universal-declaration/translations/german-deutsch

[30] Vgl. Platon, Philebos, 64d-64e.

[31] Vgl. Nikolaus Von Kues, De docta ignorantia, Kap. 1: „Comparativa igitur est omnis inquisitio medio proportionis utens."

> umbilden können, jederzeit aber, wie mit Zügeln zurückgehalten, die unausweichliche Herrschaft des Gesetzes anerkennen müssen. – Goethe zur Metamorphose (1892, S. 190)

Diese Beobachtung von Goethe, der tief ins Geheimnis des Lebens eingedrungen ist, gibt uns den nächsten Gedankenanstoß. Oben wurde festgestellt, dass eine theoretisch abgeleitete, in sich geschlossene Liste von ethischen Werten unserem Projekt zugrunde liegt. Und wir wollen Nutzer unserer App darauf hinweisen, dass alle Werte ins Auge gefasst und in der Praxis gewürdigt werden müssen. Hier manifestiert sich der *deontologische* Zug der Ethik, d.h. jeder Bereichswert fungiert als kategorischer Imperativ wie „Schütze persönliche Daten!" für die „Privatheit". Alle Pflichte der Menschheit lassen sich in dieser Weise auf das Konzept der Menschenwürde zurückführen und davon ableiten. Jedoch wird nun eine entscheidende Frage aufgeworfen, nämlich, ob alle Gebote immer gleichwertig angesehen und gleichrangig favorisiert werden können (vgl. oben Kap. 2 über die *prima facie* Werthierarchie). Bei genauerem Hinsehen stellt sich aber heraus, dass z. B. der Zugriff auf persönliche Daten, der wegen des Schutzes der Privatsphäre gering gehalten werden sollte, genau die Bedingung dafür bilden, um einen besseren, persönlich angepassten Service zu erhalten (nicht nur bei Online-Werbungen, sondern auch bei alltäglichen Kommunikationen).[32] Teilte das Volk nur das mit, welches über das Nötige nicht hinausgeht, so gäbe es weder freie Marktwirtschaft noch soziale Netzwerke, welche beide aus der freiwilligen Preisgabe der privaten Informationen resultieren. Hier kollidieren oder konkurrieren zwei Werte – die Privatheit und die Partizipation – miteinander.[33] Oder kann das gesteigerte Bewusstsein über die Ressourcenknappheit unsere Autonomie (z. B. die Gas-Heizung oder Dieselautos zu benutzen) einschränken (Suffizienz vs. Autonomie). Die Bereichswerte sind deshalb eng ineinander verflochten und nur in die-

[32] Vorausgesetzt, dass man hier selbst entscheidet, inwieweit man eigene Daten preisgeben will (opt-in)

[33] Amartya Sen (1987) unterscheidet in seiner Theorie des Lebensstandards zwischen konstitutiven und konkurrierenden (competitive) Werten. Unsere Theorie arbeitet ebenfalls an dieser Elaboration, um die Fragen, bei denen die Wertschätzung ermittelt wird, in wechselseitige Relationen zu setzen.

sem Wechselverhältnis fassbar. Nach Goethe (1893, S. 16 und 18) verrät dieses Phänomen im Bereich des Organismus „das Gesetz [...,] daß keinem Theile etwas zugelegt werden könne, ohne daß einem andern dagegen etwas abgezogen werde, und umgekehrt" oder die „Idee eines haushälterischen Gebens und Nehmens". Wie es im obigen Zitat auch deutlich zur Sprache kommt, ist die Herrschaft des Gesetzes nicht stationär, sondern *dynamisch*, weil sich jeder Organismus mit seiner besonderen Umwelt abfinden und dabei zahlreiche Unformen tolerieren muss, welche dennoch vom Gesetz umfasst werden. In unserem Konzept übernimmt die allgemeine Menschenwürde dieselbe Rolle wie die Urpflanze oder das Urtier, dessen Manifestation immer an Einzelexemplaren als messbare Indikationen erfasst werden muss. In anderen Worten wird die deontologische Ethik, bei der alles strikt durch allgemeine Gesetze reglementiert wird, durch diese morphologische Flexibilität aufgelockert und somit mit dem *Konsequentialismus* in Verbindung gesetzt.

Diese Position, welcher der Utilitarismus auch zugeordnet werden kann (vgl. Parfit 2011, S. 374), besagt hier lediglich, dass unsere Bereichswerte der Theorie nach universal und *a priori* begründbar sind, in der Realität aber *unterschiedlich gewichtet* werden müssen. Obwohl das Wasser für jede Pflanze essenziell ist, kann zu viel Wasser sie töten, denn die angemessene Wassermenge ist sowohl von dem einzelnen Organismus als auch von der Umgebung abhängig und durch das Übermaß werden andere wichtige Funktionen beeinträchtigt. An dieser Stelle wird die bereits angeführte Kritik des Aristoteles an der einseitigen Theoretisierung einschlägig, welche durch die empirische Beobachtung ergänzt werden muss. Die Abbildung konkreter Wertschätzungen folgt aus und erfolgt bei einer Auseinandersetzung mit tatsächlichen Lebensbedingungen. Diese Variante der Ethik wird einerseits von Scheler (1916, S. 109) als „materielle Ethik" der Kantischen, formalen Gesinnungsethik (oben der Deontologie) entgegengesetzt und problematisiert. Andererseits geht Weber (1926, S. 65f) davon aus, dass in der „Politik als Beruf" die beiden Typen der Ethik zum Einsatz kommen, denn die an den Konsequenzen orientierte Ethik steht mit dem Begriff der „Verantwortung" aufs engste in Verbindung, während die „Gesinnungsethik" in der Lage ist, absolut gültige Vorsätze zu ermöglichen. Unsere Smart

City Ethik verkörpert ebenfalls diese Dichotomie, die aber nicht den Schluss, sondern den Ausgangspunkt unserer Diskussion bildet. Was wir hier als Versuch anstellen, ist eine neue Methodik, welche den geschilderten Dualismus durch eine smarte, bzw. digital-unterstützte Praxis überwindet. Wie bei dem lebendigen Organismus soll sich die Ethik nicht nur mit universalen Regeln begnügen, sondern sie muss der komplexen Lebenssituation gerecht werden, in der verschiedene Werte miteinander konkurrieren oder kooperieren können. Um diese Lage genau zu erkunden, spielt die KI eine entscheidende Rolle.

3. Die Methodologie unserer Smart City Ethik

Aus dem Gesagten ergibt sich die Dreiteilung der Aufgabe und des Prozederes unserer ethischen Konzeption: 1. Die *Philosophie* ist dafür zuständig, ethische Reflexionen kraft des angemessenen Wertkatalogs zu begünstigen und zu leiten (vgl. Kapitel 2). 2. In Zusammenarbeit mit der *KI-Forschung* wird ein digitales Werkzeug hergestellt, durch welches die Kompatibilität von Wertschätzungen verschiedener Personen, Projekten und Smart Cities ergründet wird. 3. Da jedoch der Wertkatalog eine Hypothese ist und keineswegs endgültig konzipiert werden kann, muss ein realer, zwischenmenschlicher Diskurs bereitgestellt und unter Einschluss der *Kommunikationswissenschaft* mit allen Beteiligten gemeinsam vorangetrieben werden. Obwohl die beiden letzten Schritte nur als Entwürfe dargelegt werden können, ist das Folgende eine detaillierte Skizze jedes Teilaspekts.

- Der erste Schritt: Aufstellung der universalen Regeln (Bereichswerte) und Generierung von Fragen

Diese Arbeit konstituiert den Hauptteil des SCET-Projekts. Dabei wird die philosophische, holistische Überlegung über die Menschenwürde zum Ausgangspunkt gemacht, während verschiedenen Problemen aus der Praxis in vorhandenen Smart Cities ebenso Rechnung getragen wird. Das Entscheidende ist hier jedoch die transzendentale, d.h. idealverbindliche Begründung der Ethik als methodischer Grundsatz. Wir laufen nicht passiv zufällig eintretenden Schwierigkeiten hinterher, sondern identifizieren aktiv Verbesserungsmöglichkeiten anhand unse-

rer eigenen Maßstäbe. Das Ideal des *Menschen* dient als Zielpunkt der ethischen Reflexion, welche durch unsere App angestellt wird. Die tatsächliche Befragung wird aber in konkreten Situationen umrahmt. Die Befragten werden nicht direkt auf die Bereichswerte aufmerksam gemacht, sondern mittelbar darauf hingewiesen, indem sie über mögliche Wertkonflikte nachzudenken bekommen. Einige (noch nicht endgültig formulierte) Beispiele sind:

- Wollen Sie in Smart Cities mehr persönliche Daten abgeben / sammeln, um bessere Angebote zu erhalten / zu kreieren?
- Denken Sie, dass lokale Unternehmen mittels der neuen digitalen Technologie Ressourcen global erschließen können?
- Haben Sie das Gefühl, dass kraft der Online-Tools die Einwohner einer Smart City von ihr physisch unabhängig werden?

- Der zweite Schritt: Suche nach einem an die Tatsachen angepassten Muster der Wertschätzung

Nachdem die ethisch relevanten Gesichtspunkte theoretisch fixiert werden, wird das dynamische Verfahren in Gang gesetzt. Jede Nutzerin oder jeder Nutzer unserer App durchläuft eine Befragung und erfährt daraus am Ende seine oder ihre Zu- und Abneigung zu bestimmten Werten in Zahlenskala, welche in ein Muster zusammengeführt und im Radar-Diagramm visualisiert werden (Abbildung 5). Hier gibt es aber kein endgültiges oder korrektes Muster, welches jeder Mensch, jedes Projekt oder jede Smart City anzustreben hat. Vielmehr handelt es sich um die Frage, inwiefern das Interesse und die Zielsetzung von den Beteiligten (auch Bürgern) in einer konkreten Situation abgestimmt oder harmonisiert sind. Um diese moralische Diagnose zu gewährleisten, werden Daten von Befragten anonym gesammelt, gespeichert und zur Quelle für den Vergleich umfunktioniert. Die App bietet sich somit als *eine fiktive Dialogpartnerin* an, welche entpersönlicht, aber keinesfalls vom Blackbox-Dilemma bedroht ist, denn die KI-Technik inkorporiert und organisiert authentische Stimmen des Menschen, ohne sie zu manipulieren. Wenn z. B. ein Projektleiter sein Projekt primär für die ältere Bevölkerungsschicht entwirft oder ihr Anliegen als Schwerpunkt der Digitalisierung betrachtet, sollte seine Vision idealerweise mit dem Wunschprofil dieser Zielgruppe übereinstimmen. Im umgekehrten Fall

können Bürger auch mit der App erkunden, wie diejenigen, die für Projekte verantwortlich sind, zu den gleichen Bereichswerten Stellung nehmen, und ob zwei Ansichten konvergieren oder divergieren. Und dieser simulative Interessenvergleich kann mit der KI-Technologie fast unendlich erweitert werden, z. B. dadurch, dass man andere Invariablen wie Geschlecht, Beruf, Herkunft, Bildungsniveau, Technikaffinität usw. hinzufügt. Mehr dazu im nächsten Abschnitt.

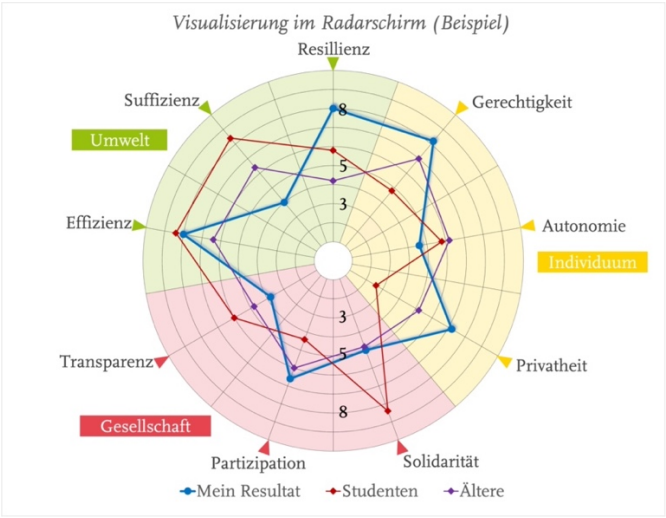

Abbildung 5. Beispiel der Abbildung der Präferenzen unterschiedlicher Gruppen im „Wertradar“

- Der dritte Schritt: Realer Bürgerdialog und Bildung des Vertrauens

Das gelingende soziale Leben besteht, selbst wenn die digitale Technologie unsere alltägliche Kommunikation ständig beschleunigt, erweitert und vervielfacht, letztendlich in einem zwischenmenschlichen Dialog, wo, wie der siebte Brief Platons schön sagt (341c-341d), der Funke der Wahrheit das Feuer unserer Seele anzündet. Während Philosophen über Universalwerte nachdenken und KI-Techniker einen sicheren und effizienten Austausch über diese mittels des Computer-Programms bewerkstelligen, haben alle Menschen *qua* ihres Mensch-Seins Anspruch darauf, eigene Meinungen dazu zu äußern und damit gemein-

schaftlich das Gute in die Wege zu leiten. Unsere App bildet dafür eine Grundlage, auf der sich ein vertrauensvoller Diskurs aufbauen und florieren kann. Workshops und interdisziplinäre oder sektorübergreifende Gespräche bieten uns Gelegenheiten, die Regeln der Ethik je nach Bedarf und Wunsch zu modifizieren, damit sie, um es aus Goethes bereit zitiertem Ausdruck zu entlehnen, „zwar fest und ewig, aber zugleich lebendig" bleiben können.

4. Synergie der Ethik und KI

> A competent judge is required to be a reasonable man [...]: First, a reasonable man shows a willingness, if not a desire, to use the criteria of inductive logic in order to determine what is proper for him to believe. Second, a reasonable man, whenever he is confronted with a moral question, shows a disposition to find reasons for and against the possible lines of conduct which are open to him. Third, a reasonable man exhibits a desire to consider questions with an open mind, and consequently, while he may already have an opinion on some issue, he is always willing to reconsider it in the light of further evidence and reasons which may be presented to him in discussion. Fourth, a reasonable man knows, or tries to know, his own emotional, intellectual, and moral predilections and makes a conscientious effort to take them into account in weighing the merits of any question. – J. Rawls (1951, S. 178f)

Um das Ganze abzuschließen, wird das Verhältnis der Ethik und KI-Technologie zueinander genauer erörtert, die auf den ersten Blick zwei inkompatible Disziplinen zu sein scheinen. Rawls vertritt eine kontraktualistische Theorie der Ethik und überlegt, wie Menschen in einem idealen Gesprächsraum ein faires Miteinander konzipieren würden. Jahrzehnte vor der Erscheinung dieser bekannten Schrift über die Gerechtigkeit als Fairness expliziert er eine induktive Methode der Ethik, welche *in nuce* darin besteht, dass sich alle an Moralentscheidungen Beteiligten im Fall möglicher Interessenkonflikte auf bestimmte Prinzipien einigen, damit ihre Handlungen fairerweise gerechtfertigt werden. Die Bedingung dafür erfüllt ein kompetenter Richter oder ein Vernünftiger, der in der Lage ist, über Meinungsunterschiede derart zu richten,

dass er ein Gesetz, welches virtuell von allen Seinesgleichen anerkannt werden soll, ausfindig macht und proklamiert, obwohl Rawls zugibt, dass keine genau beschreibbaren Methoden der Entdeckung („no precisely describable methods of discovery") solcher Gesetze vorhanden sind (ebd., S. 196). Dieser Gedanke ist augenscheinlich an die Methodik der Naturwissenschaft angelehnt, in der auch objektiv gültige Naturgesetze eruiert und implementiert werden, während sowohl die Ausfindung als auch Applikation bestimmter Gesetze von Fall zu Fall variieren.

In dieser Überlegung liegt bereits die Schnittstelle zwischen der Ethik und dem logisch-induktiven Denken nahe. Um diese Annäherung noch weiter voranzutreiben, wird ein Blick auf interessante Debatten innerhalb der KI-Forschung geworfen über die symbolische und subsymbolische KI. Herkömmliche Computerprogramme werden hier als symbolisch bezeichnet, weil sie mit Zeichen operieren und nach vorgegebenen Regeln einen bestimmten logischen Schluss ziehen, so wie wir normalerweise mit unserer Sprache vorgehen. So wissen wir, wenn jemand uns bittet, den Apfel zu holen, welche Merkmale (wie Form, Farbe, Geruch usw.) ein Apfel besitzt und welchen Akt das Holen bedeutet (nämlich etwas in die Hand nehmen und zur anderen Person überbringen). Allerdings sind wir oft nicht imstande, einen Input eins zu eins mit einem Output zu verknüpfen, und leiten deshalb die Bedeutung eines Phänomens aus verschiedenen Assoziationsbildern ab. Ein Beispiel dafür wäre das einfache Hand-Winken, welches ein breites Interpretationsspektrum eröffnet: Winkt die Person mir oder einer anderen hinter mir? Will sie ein Taxi anhalten? Will sie mich bloß begrüßen oder herbeirufen? Will sie sich lediglich von mir verabschieden? usw. Hier sind keine allgemeingültigen Gesetze vorhanden, um festzustellen, welche Art des Winkens die Person beabsichtigt, während ein angemessenes Verhalten dennoch durch ständiges situatives Lernen antrainiert werden kann. Dieses Maschinenlernen wird im Bereich der Informatik oder Kognitivwissenschaft im Gegensatz zur symbolischen Logik „sub-symbolic" bezeichnet (vgl. z. B. Kelley 2003), da sich in diesem Bereich noch keine feste Sprache bzw. Symbolik definieren lässt. Die Relation beider Verfahrensweisen ist allerdings, wie es aus unseren alltäglichen Erfahrungen ersichtlich wird, keineswegs disjunktiv (entweder-oder), sondern konjunktiv (und). In der Arbeit von Strannegård &

Nizamani (2016) wird gezeigt, inwiefern eine KI, die in einer virtuellen Welt verschiedene Nahrungsmittel für ihre größte Überlebenschance zu gebrauchen hat, beide Denkarten komplementär einsetzen kann: Einerseits (beim regelorientierten Denken) strebt das KI-System nach der Befriedigung einer fixierten Anzahl von Bedürfnissen (wie Trinken und Essen), ohne welche das Leben nicht aufrechterhalten werden kann. Andererseits (beim assoziativen Denken) werden kontingente Faktoren berücksichtigt und hinsichtlich der Überlebenswahrscheinlichkeit kalkuliert, welches zur Entscheidung führt, welches Bedürfnis unter der gegebenen Bedingung zuerst befriedigt werden soll.

Mit dem Blick auf das Gesagte lässt sich, ohne weiter auf technische Details einzugehen, die Anwendungsmöglichkeit der KI in der Ethik folgendermaßen zusammenfassen: Menschliche Entscheidungen erfolgen, wie Rawls in seiner Ethik konstatiert und auch das obige KI-Programm hinsichtlich der Überlebensstrategien zeigt, weder komplett gesetzmäßig, noch völlig willkürlich. In unserem Konzept haben wir auf der einen Seite die universalen Bereichswerte aufgelistet, welche theoretisch im Konzept der Menschenwürde fest verankert sind. Auf der anderen Seite aber muss je nach der Situation eine Präferenz oder Hierarchie zwischen diesen statuiert werden, denn es ist utopisch oder sogar unerwünscht, alle Gebote simultan und äquivalent zu betrachten, weil sie unter variierenden empirischen Bedingungen (wie Altersgruppe, Geschlecht, Beruf usw.) anders gewichtet werden müssen. Dieser Aspekt kann durch die KI-Technologie umfangreicher und ökonomischer beleuchtet werden, da sie schlicht und einfach mehr Probedaten länger speichern und diese Datenbank blitzschnell für einen präzisen Vergleich mobilisieren kann als menschliche Intelligenz. Man muss sich jedoch im Voraus darüber im Klaren sein, dass die KI hier nichts von dem Inhalt, d.h. von der Ethik weiß, sondern lediglich für menschliche Entscheidungen Muster bereitstellt.[34] Die KI kann beispielsweise zeigen,

[34] Bei der sub-symbolischen Logik kann die KI z. B. das Gesicht eines Hundes von dem der Katze *nicht eindeutig* unterscheiden, da diese klare Abgrenzung (welche bei der symbolischen Logik nach einer eindeutigen Regel erfolgt) nicht anzuwenden ist. In anderen Worten wird diese Fragestellung an sich mehr oder weniger aufgehoben, denn die KI versucht lediglich, beliebig gegebene Gesichter nach bestimmten Kriterien zu differenzie-

wie die Gruppe von allein-erziehenden Müttern mit einem Immigrationshintergrund in mittelgroßen (Smart) Cities oder die Minderheit von älteren, in ihrer Mobilität eingeschränkten und technisch-handicapten Menschen in Metropolen die neun Bereichswerte auslotet und dazu kommentiert. Derartige solide Fallanalyse versetzt uns in die Lage, Rawls' vier Kriterien eines guten Richters (d.i., das induktive Denken, die Pro-und-Kontra-Abwägung, die Offenheit und die Achtsamkeit auf eigene Vorurteile) zu genügen, um eine Smart City auf faire Weise zu gestalten.

5. Fazit

In dem Kapitel und auch in unserem Projekt, anstatt ein ethisches Verhalten der KI zu begründen, wird zuerst die Ethik als solche rein aus der Perspektive der Würde des Menschen abgeleitet und dann bei ihrer Umsetzung durch die KI-Technologie auf eine neue Bahn gesetzt. Unsere Bereichswerte sind einerseits unhintergehbar und unabhängig von Raum, Zeit und Kultur gebieterisch. Alle Menschen in Smart Cities für diese klar ausformulierten Regeln zu sensibilisieren, macht die erste Aufgabe unserer App aus (die symbolische KI). Anderseits muss jedoch situativ ausgearbeitet werden, auf welche dieser Regeln die Beteiligten ihre gemeinsamen Handlungen zurückführen wollen. Hierzu gibt die App ein praktisches Werkzeug in unsere Hand. Wie im Gerichtsverfahren, kann eine faire Entscheidung erst dann gemeinschaftlich von Jurys gefällt werden, wenn teils konkurrierenden Aussagen von allen Augenzeugen einer sorgfältigen Betrachtung unterzogen worden sind. Die menschliche Intelligenz bedarf insbesondere in dieser Hinsicht der künstlichen Intelligenz, die unmittelbar den ganzen Korpus von Berichten nach bestimmten Kriterien sortieren und mögliche Hinweise auf relevante Fälle geben kann (die sub-symbolische KI). Zwei Teile der Ethik – Deontologie und Konsequentialismus – entsprechen somit zwei Typen der Computerlogik und die gegenseitige Ergänzung von Ethik und KI liegt uns gar nicht fern, sondern in der Natur der Sache.

ren und ggf. assoziativ in verschiedene Cluster einzuteilen. Es ist dann dem Menschen überlassen darüber eine Aussage zu machen, welches Cluster was darstellt.

Literaturverzeichnis

Aristoteles (1911) *Nikomachische Ethik* (2. Aufl.), übersetzt von E. Rolfes. Leipzig: Felix Meiner.

Goethe, J. W. v. (1892) *Zur Morphologie, II. Teil*. Weimar: Hermann Böhlau (Weimarer Ausgabe, II. Abteilung, 7. Band)

Goethe, J. W. v. (1893) *Zur Morphologie, III. Teil*. Weimar: Hermann Böhlau (Weimarer Ausgabe, II. Abteilung, 8. Band)

Kelley, T. D. (2003) „Symbolic and Sub-Symbolic Representations in Computational Models of Human Cognition: What Can be Learned from Biology?" *Theory & Psychology* 13 (6), 847–860.

High-Level Expert Group on Artificial Intelligence (2019) *Ethics guidelines for trustworthy AI*. European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Nietzsche, F. (1894) *Menschliches, Allzumenschliches. Ein Buch für Freie Geister. Erster Band* (2. Aufl.). Leipzig: C. G. Naumann.

Nikolaus von Kues (1994) *De docta ignorantia. Die belehrte Unwissenheit Buch 1* (4. Auflage), übersetzt und herausgegeben von E. Hoffmann, P. Wilpert und K. Bormann. Hamburg: Felix Meiner.

Nussbaum, M. C. (1993) „Non-Relative Virtues: An Aristotelian Approach". In: Nussbaum, M. C. & Sen, A. (Hg.) *The Quality of Life*. Oxford: Oxford University Press, 242-269. https://doi.org/10.1093/0198287976.003.0019

Nussbaum, M. C. (2011) *Creating Capabilities: The Human Development Approach*, Cambridge, MA [u.a]: Belknap Press of Harvard University Press.

Parfit, D. (2011) *On What Matters. Volume One*. Oxford, Oxford University Press.

Rawls, J. (1951) „Outline of a Decision Procedure for Ethics". *The Philosophical Review* 60(2), 177-197.

Scheler, M. (1916) *Der Formalismus in der Ethik und die materiale Wertethik: Neuer Versuch der Grundlegung eines ethischen Personalismus*. Halle a. d. S.: Max Niemeyer.

Sen, A. (1987). „The Standard of Living: Lecture I, Concepts and Critiques". In: A. Sen, J. Muellbauer, R. Kanbur, K. Hart und B. Williams. *The Standard of Living*. G. Hawthorn (Hg.). Cambridge: Cambridge University Press.

Strannegård, C. und Nizamani, A. R. (2016) „Integrating Symbolic and Sub-symbolic Reasoning". In: Steunebrink, B., Wang, P. und Goertzel, B. (Hg.) *Artificial General Intelligence. AGI 2016. Lecture Notes in Computer Science vol. 9782.* Cham: Springer Nature, 171-180. https://doi.org/10.1007/978-3-319-41649-6_17

UN. „Die Allgemeine Erklärung der Menschenrechte". https://www.ohchr.org/en/human-rights/universal-declaration/translations/german-deutsch

Weber, M. (1926) *Politik als Beruf.* München [u.a.]: Duncker & Humblot.