HMD (2022) 59:447–467
https://doi.org/10.1365/s40702-022-00855-y

EINFÜHRUNG

Computational Ethics

Ein ethischer Lösungsansatz für das KI-Zeitalter

Edy Portmann · Sara D'Onofrio

Eingegangen: 7. November 2021 / Angenommen: 15. Februar 2022 / Online publiziert: 15. März 2022

Zusammenfassung Die zunehmende Nutzung und Akzeptanz der künstlichen Intelligenz (KI) erfordert neue Lösungsansätze, die ethische Werte wie Privatsphäre und Gleichheit in Entscheidungsprozessen berücksichtigen. In diesem Artikel stellen wir das konzeptionelle Artefakt der Computational Ethics als eine der Lösungsansätze des KI-Zeitalters vor, das aus den Elementen Computing with Words and Perceptions, digitale Ethik mit wort-/wahrnehmungsbasierten Werten und den Rahmenbedingungen eines Human Life Engineering besteht. Das komplementäre Zusammenspiel dieser drei Elemente erlaubt es der Computational Ethics, natürlichsprachige Informationen, bewertet nach gemeinsam festgelegten diskursethischen Grundsätzen, in einem kontextbasierten Modell natürlichsprachigen Rechnens als Referenzschema zu nutzen und die möglichen Handlungsoptionen einer KI nach ethischem Maß zu bewerten. So können autonome Entscheidungen unter Berücksichtigung der ethischen Werte des einzelnen Menschen und der Gesellschaft getroffen werden.

Schlüsselwörter Computational Ethics · Computing with Words and Perceptions · Digitale Ethik · Human Life Engineering · KI-Zeitalter · Künstliche Intelligenz · Perceptual Computing · Werte

Edy Portmann (✉)
Human-IST Institut, Universität Fribourg, Fribourg, Schweiz
E-Mail: edy.portmann@unifr.ch

Sara D'Onofrio
IT Business Integration, Genossenschaft Migros Zürich, Zürich, Schweiz
E-Mail: info@saradonofrio.ch

Computational Ethics

An Ethical Solution Approach for the Era of Artificial Intelligence

Abstract The increasing use and acceptance of artificial intelligence (AI) requires new solution approaches that consider ethical values such as privacy and equality in automatic decision-making processes. In this article, we present the conceptual artefact of a computational ethics as one of the solution approaches of the new age of AI, which consists of the elements of computing with words and perceptions, digital ethics with word-/perception-based values, as well as the conditions of a human life engineering. The complementary interplay of these three elements allows our computational ethics to use natural language information, evaluated based on commonly established discourse-ethics principles, in a context-based model of natural language computing as a reference scheme and to assess the possible courses of action of an AI according to ethical standards. In this way, autonomous decisions can be made, while considering the ethical values of the individual and society.

Keywords Artificial Intelligence · Computational Ethics · Computing with Words and Perceptions · Digital Ethics · Era of AI · Human Life Engineering · Perceptual Computing · Values

1 Wichtigkeit von Ethik im Digitalen Raum

Digitale Technologien, wie etwa die der Künstlichen Intelligenz (KI), gewinnen zunehmend an Akzeptanz. Es gibt unterschiedliche Auffassungen, was eine KI kann und was nicht. Für diesen Artikel stützen wir uns auf das Verständnis von Abele und D'Onofrio (2020), die mithilfe eines Frameworks KI als Zusammenspiel von verschiedenen Ansätzen (z. B. Artificial Neural Networks, Statistical Probabilistic Inference, Evolutionary Optimization) erklärt haben. Während viele KI-Ansätze auf Techniken des Maschinellen Lernens (*Machine Learning*) setzen, unterstützen wir die Ansicht, dass Lernen nur eine der Fähigkeiten (*Capability*) der KI ist und ihre möglichen weiteren Fähigkeiten (z. B. Know-how, Kreativität, Intuition) noch erforscht werden müssen.

Das Fehlen der Anforderungen *Transparenz* und *Erklärbarkeit* der KI resp. was unter einer solchen verstanden werden kann, birgt nun aber viele Gefahren und kann je nach Einsatz kritisch werden. Wird man etwa auf dem Smartphone mit der Gesichtserkennungssoftware nicht erkannt resp. falsch klassifiziert (Google reparierte seinen rassistischen Algorithmus, indem es *Gorillas* aus der Bildkennzeichnung entfernt hat), so ist der mögliche Schaden verkraftbar. Werden jedoch Algorithmen in Bereiche eingesetzt, die über Leben oder Tod bestimmen könnten (bspw. in selbstfahrenden Fahrzeugen oder autonomen Operationsrobotern), so könnten Fehlentscheidungen hohe Schäden verursachen (Samek und Müller 2019). In dieser Hinsicht ist immer noch nicht geklärt, wer haften würde, wenn durch eine KI ein Schaden verursacht werden würde (Čerka et al. 2015).

Das Herzstück einer heutigen KI sind die Daten, mit der sie gefüttert wird. Diese Daten sind aber im Gegensatz zu Öl, mit dem sie vielfach verglichen werden, leben-

dig, weil diese auch von uns Menschen produziert werden (vgl. *Citizens as a Sensor*; Goodchild 2007). Sind die Daten unvollständig, veraltet oder sogar falsch, so lernt die KI auf nicht korrekten Annahmen, wodurch sie schließlich unzufriedenstellende Leistungen erbringt und Unmut bei den Benutzer:innen stiftet. Im Weiteren können Datensätze gewisse Voreingenommenheit, Vorurteile und Verzerrungen (*Biases*) enthalten, die insbesondere bei Entscheidungen basierend auf personenbezogenen Daten zu Diskriminierungen führen könnten (Apprich et al. 2018). Einer bestimmten Bevölkerungsgruppe könnten etwa Kredite oder der Wohnungsbezug verwehrt werden, obwohl sie wie alle anderen das gleiche Recht dazu hätten. Wie kann daher sichergestellt werden, dass die KI die Vorgaben des Datenschutzes erfüllt und nicht etwa die Benutzer:innen aufgrund potenzieller *Biases* diskriminiert oder sogar gegen die Grundrechte verstößt (vgl. Amazon, deren KI die weiblichen Bewerber:innen aufgrund ihres Geschlechts ablehnte; Hamilton 2018)? Mit solchen Biases können Algorithmen die in unserer Gesellschaft bereits bestehenden Ungleichheiten etwa in Bezug auf Klasse, Rasse und Geschlecht aufrechterhalten und sogar verstärken. Des Weiteren könnte die übermäßige Nutzung der KI weitere Nachteile mit sich bringen, die auf den ersten Blick nicht ersichtlich sind. Crawford (2021) präsentiert uns bspw. unter den Stichworten *Seltene Erde* und Missbrauch von *schlecht bezahlter Crowdsourcer* auch Energie- und Produktionsprobleme als weitere Problematik der KI.

Um diesen Gefahren entgegenzutreten, ist es notwendig sich auch mit den ethischen Aspekten in der KI-Entwicklung und -Nutzung auseinanderzusetzen. Verschiedene Organisationen, wie etwa die Europäische Union, sowie Unternehmen und Forschungsinstitutionen haben erste ethische KI-Richtlinien verfasst (s. Kap. 2). Diese KI-Richtlinien zeichnen sich primär mit der Sicherstellung einer vertrauenswürdigen KI aus (d. h. einer KI, die „im guten Sinn" handelt; Hallensleben 2021). Eine KI, die bspw. personenbezogene Daten nicht zweckentfremdet resp. missbraucht und dadurch einer Person einen Schaden zufügt. Die Herausforderung besteht darin, diese ethischen KI-Richtlinien operationalisierbar zu machen (Hallensleben 2021).

In diesem Kontext schlagen wir die Anwendung von berechenbarer Ethik (*Computational Ethics*) vor. Wir folgen hier Moor (1995), der sich bereits Mitte der 1990er-Jahre fragte, ob Ethik resp. ob *gutes* und *richtiges Handeln*, berechenbar sei und ob Computer sich überhaupt ethisch verhalten können. Er sah dabei eine graduelle (Weiter-)Entwicklung dieses Themas voraus, wie sie etwa in der Fuzzy Logik als Zugehörigkeitsfunktion eingesetzt wird, wie Systeme „nach und nach, Stück für Stück, Bedingung für Bedingung, Heuristik für Heuristik, [...] in ihren ethischen Entscheidungsfindungen immer ausgefeilter werden" (Moor 1995, S. 11).

Unser hier präsentierter Lösungsansatz zu Computational Ethics, der sich v. a. mit westlichen Wertvorstellungen beschäftigt, aber sich durchaus auch auf andere Wertesysteme übertragen lässt (vgl. Ito 2020), kombiniert drei Elemente: das Rechnen mit Worten und Wahrnehmungen (*Computing with Words and Perceptions*) in Anlehnung an Zadeh (2012) und Mendel und Wu (2010), die wort-/wahrnehmungsbasierten Werte aus den diskursethischen Diskussionen in Anlehnung an Spiekermann (2019) und Floridi (2018), und die Rahmenbedingungen des Human Life Engineering gemäß Österle (2020) und Österle et al. (2021), um unsere menschlichen, unscharfen Wertvorstellungen (oder kurz Werte) zu zähl- und berechenbaren Modellen

der KI-Systeme als eine Art ethische Instanz zur Verfügung zu stellen. Das erste Element des Rechnens mit Worten und Wahrnehmungen ermöglicht es, die wort-/wahrnehmungsbasierte Werte des zweiten Elements in Metrik-basierte KI-Modelle als drittes Element, umzuwandeln. Das Ziel ist es also, statisch beschriebene, geisteswissenschaftliche Konzepte der traditionellen Ethik in konstruktive Algorithmen zu überführen, welche es einer KI erlauben *passende* moralische Entscheidungen situativ (d. h. kontext-sensitiv) aus den vorhandenen Informationen und Rahmenbedingungen abzuleiten (d. h. zu berechnen).

Dieser Artikel widmet sich demgemäß der Computational Ethics, als einen der Lösungsansätze für das fortschreitende KI-Zeitalter. Kap. 2 gibt dazu einen kurzen Überblick der bisher verfassten ethischen KI-Richtlinien und führt den Begriff *Computational Ethics* ein; Kap. 3 erklärt die drei erwähnten Elementen sowie ihr Zusammenspiel; die Bewertung dessen folgt im Kap. 4; Kap. 5 diskutiert die Erkenntnisse und ein Fazit rundet zum Schluss den Artikel ab.

2 Von digitaler Ethik zur Computational Ethics

Die zunehmende Beachtung ethischer Aspekte in IT-Systemen lässt sich mit dem rasanten Fortschritt und dem zunehmenden Einsatz digitaler Technologien und insbesondere der KI-Nutzung begründen. Mit der Annahme, dass KI-Systeme basierend auf den verfügbaren Daten und deren Beziehungen untereinander autonom Entscheidungen treffen, ist es umso wichtiger, sich mit den Grundprinzipien des menschlichen Handelns auseinanderzusetzen, um ein Referenzschema für das *gewünschte* maschinelle Handeln abzuleiten. Unser Schwerpunkt liegt hier auf der eudämonistischen Ethik, die dafür steht, moderne Technologien im Sinne jedes Einzelnen basierend auf individuellen Werten wie Privatsphäre und Freiheit einzusetzen (Spiekermann 2019).

Gemäss Spiekermann (2019) umfasst eine solche Ethik alle klassischen, tugendethischen Formen unseres Denkens. In ihren Leitlinien baut sie aber neben Aristoteles' Tugendethik auch auf Kants' deontologischer Ethik (oder *Pflichtethik*), um über Verhaltenspflichten nachzudenken und um Wertprioritäten für die Systemgestaltung zu ermitteln und festzulegen (Alt et al. 2021). All dies kann durch Einbezug wichtiger Interessengruppen in einen Dialog bewirkt werden, der von diskursethischen Grundsätzen geleitet werden sollte (Floridi 2018). In diesem Dialog[1] sollte gemäß der digitalen Ethik über diejenigen kulturellen Traditionen reflektiert werden, welche dazu beitragen können, Werte eines KI-Systems zu antizipieren, die vom ethischen Kanon nicht erfasst werden. Die Frage, die sich stellt, ist nun, ob Ethik auch berechenbar, also *computable* ist.

Moor (1995) fragte sich, als ein Vordenker unserer automatisierten Digitalwelt, ob sich denn Computer und Roboter ethisch verhalten können. Er deckte in seinem Artikel auf, dass menschliche Werte vielfältig und komplex sind und es oftmals eine offene Frage bleibt, wie gut sich diese Werte und ihre Beziehungen in einem KI-System

[1] Floridi (2018) bezeichnet diesen Dialog als *Soft Ethics* und grenzt diesen von *Hard Ethics* (wie Regulation und einer technischen Umsetzung im KI-System) ab.

darstellen lassen. Aber unabhängig von möglichen Grenzen einer computergestützten, ethischen Entscheidungsfindung sollten sich Entwickler:innen, Designer:innen und Programmierer:innen doch schon mal als eine Art *Werteschmied:innen* verstehen. Wohl als erster beschäftigte sich dann Aaby (2005) mit einer Theorie von Computational Ethics, in der er Ethik als eine Art formales System resp. eine Abstraktion der Realität betrachtete (analog etwa zu Geometrien, die Abstraktionen unserer Realität sind). Als letzter Forscher dieser emergenten KI-Disziplin, betonte Segun (2020) die Wichtigkeit von Computational Ethics im Streben nach einer ethischen KI-Instanz, die auf menschliche Werte sowie deren Ethik reagiert und dafür sensibel sein kann.

Weltweit haben viele Organisationen die Wichtigkeit, oder sogar Dringlichkeit, ethische Werte und Prinzipien in der Systemgestaltung zu berücksichtigen, erkannt und erste Maßnahmen getroffen, wie bspw. Handlungsempfehlungen in Form von ethischen KI-Richtlinien zu schreiben. Die Europäische Union hat *Ethik-Leitlinien für eine vertrauenswürdige KI* (Europäische Kommission 2019) in mehreren Sprachen veröffentlicht, die von einer unabhängigen Expertengruppe für künstliche Intelligenz verfasst wurden. In diesen Leitlinien wird ein Rahmen für die Entwicklung einer vertrauenswürdigen KI definiert und dabei aufgezeigt, wie ethische Grundsätze von Grundrechten abgeleitet und diese schließlich als Basis für eine vertrauenswürdige KI dienen.

Auch Non-Profit-Organisationen, wie Amnesty International und Access Now, setzen sich dafür ein, dass beim Einsatz von *Machine-Learning*-Systemen ein Augenmerk auf die ethischen Werte, insb. Gleichheit und Nichtdiskriminierung, gesetzt werden, um die Benachteiligung gewisser Bevölkerungsgruppen präventiv zu vermeiden (Access Now 2018). Und auch immer mehr Unternehmen fühlen sich in der Pflicht, sich dieser Thematik zu widmen: Microsoft (2018) veröffentlichte bspw. den Report *Responsible Bots*, IBM (2019) das Dokument *Everyday Ethics for Artificial Intelligence* und Google (o.J.) gibt auf ihrer Webseite preis, welchen ethischen Prinzipien sie folgen.

Eine Auflistung aller existierenden ethischen KI-Richtlinien würde den Umfang dieses Artikels sprengen, weshalb wir hier auf die Arbeit von Jobin et al. (2019) verweisen. Sie haben insgesamt 84 KI-Richtlinien untersucht und kamen zum Schluss, dass die meistgenannten ethischen Werte aus diesen Dokumenten Transparenz, Gerechtigkeit und Fairness, Prinzip „nicht schaden“, Verantwortung, Privatsphäre, Wohltätigkeit, Freiheit und Autonomie, Vertrauen, Nachhaltigkeit, Würde und Solidarität sind – die *Werte*, die gemäss der Verfasser:innen der ethischen KI-Richtlinien bei der Entwicklung und Umsetzung einer „ethischen“ KI zu berücksichtigen sind.

Forscher:innen und Praktiker:innen befassen sich mit der digitalen Ethik. Aber wie Hallensleben (2021) betont, liegt die Herausforderung nicht darin, ethische KI-Richtlinien zu verfassen, sondern diese umzusetzen resp. *berechenbar* zu machen. Dazu möchten wir einen Beitrag leisten. Aufbauend auf den eingeführten Arbeiten, verstehen wir *Computational Ethics* als eine Brücke von (geisteswissenschaftlichen) Konzepte der traditionellen Ethik in (designorientierte) Artefakte einer gestaltungsorientierten Wirtschaftsinformatik (z. B. KI-Systeme), auf der die heutige (technische, wirtschaftliche und soziale) Entwicklung der Digitalisierung (z. B. durch Lernen aus Daten) fußt.

Abb. 1 Fuzzy Logic vs. Crisp Logic anhand einer Antwortskala von falsch zu richtig

3 Elemente des Computational Ethics

Für die Computational Ethics sind aus unserer Sicht drei Elemente erforderlich: ein Computing with Words and Perceptions, eine digitale Ethik mit wort-/wahrnehmungsbasierten Werten und auch Rahmenbedingungen des Human Life Engineering. Im Folgenden werden diese drei Elemente kurz aufgegriffen und ihr Zusammenspiel erläutert.

3.1 Computing with Words and Perceptions

Die von Zadeh (1965) eingeführte Fuzzy Set Theorie resp. Fuzzy Mengenlehre ist eine Erweiterung der binären Mengenlehre. Sie besagt, dass ein Element nicht nur zwei Werte wie *richtig* oder *falsch* (*Crisp Logic*), sondern auch dazwischenliegende Wahrheitswerte (z. B. *eher richtig, weder richtig noch falsch, eher falsch*; *Fuzzy Logic*) annehmen kann (Abb. 1).

Diese erweiterte Theorie der Mengenlehre stellt die Basis des *Computing with Words and Perceptions* dar[2], um natürlichsprachige Aussagen mittels computergestützten Modellen approximativ berechnen zu können (Portmann 2019). Diese Modelle stützen sich v. a. auf Worte, weil bspw. Zahlen nicht bekannt sind, die Impräzision dieser Worte als tolerierbar angesehen oder als linguistische Zusammenfassungen (*Linguistic Summaries*; Hudec et al. 2020) sprachlicher Information gehalten werden.

Das Element des Computing with Words and Perceptions, das Grundgerüst des Computational Ethics, durchläuft zwei Phasen (Abb. 2), um den ethischen Wert einer vom KI-System auszuführenden Handlung bestimmen zu können.

a steht dabei für das englische Wort *action* im Sinne von „die von der KI auszuführende Handlung und Entscheidung" und mit *a* beginnt der Prozess, der Asterisk steht dabei für unscharfe, also fuzzy Werte.[3] In einer hierzu spekulativ angenommenen Beispielsituation[4] einer Erneuerung eines Schulwegs am Hönggerberg (in

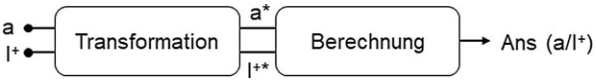

Abb. 2 Die zwei Phasen des Computing with Words and Perceptions in Anlehnung an Zadeh (2012)

[2] Die Theorie kombiniert Computing with Words von Zadeh (2012) mit wahrnehmungsbasiertem Rechnen (*Perceptual Computing*) von Mendel und Wu (2010).

[3] Bei Zadeh (2012) beginnt der Prozess mit *p* (*p* für *Proposition*).

[4] Dieses Beispiel ist fiktiv und entspricht nicht der Realität. Es dient einzig als illustrierender Anwendungsfall.

Zürich) durch das Tiefbauamt der Stadt Zürich muss bspw. eine KI eine Entscheidung automatisch treffen. Für diese Situation hat die KI-Instanz (aus Gründen der Einfachheit) zwei Optionen zur Auswahl, die sie auf ihre (ethische) Tauglichkeit bewerten muss:

- **Option 1:** a_1 ist die kostengünstige Erneuerung eines zwar als unsicher wahrgenommenen, aber per Verkehrsampel gesteuerten Fußgängerstreifens und
- **Option 2:** a_2, die veranschlagte, sicherheitstechnisch auf höchstem Niveau liegende Unterführung, die sich aber in der Stadtkasse mit einem Zusatzbeitrag zu Buche schlägt.

Beide Handlungsoptionen a_1 und a_2 werden in der ersten Phase des Computing with Words and Perceptions in berechenbare wort-/wahrnehmungsbasierte Werte (a_1* und a_2*, s. Abschn. 3.2) umgewandelt und dann ins computergestützte Berechnungsmodell überführt. Damit das Berechnungsmodell nun die maschinell auszuführende Handlung bewerten kann, wird aus den im System gespeicherten Informationsmengen I^+ ein Referenzschema ethischer Werte vorgegeben.[5]

Insbesondere bei Informationen, die auf subjektiven Ansichten basieren (z. B. Meinungen und Bedenken) gilt es, wie von Spiekermann (2019) gefordert, bereits im Voraus mit den betroffenen Interessengruppen in einen Dialog zu treten (Floridi 2018) und die diskursethischen Grundsätze gemeinsam zu erarbeiten. Diese Grundsätze sollen schließlich als das sog. Referenzschema für das KI-System dienen, damit die KI weiß, wie sie die Informationen aus den verschiedenen Quellen zu bewerten und einzuordnen hat (s. Abschn. 3.2). Sind die Informationen aus diesen Mengen (I^+) einem (Zwischen-)Wahrheitswert zugeordnet, werden die Informationen ebenfalls in wort-/wahrnehmungsbasierte Werte (I^+*) umgewandelt und schließlich ins Berechnungsmodell eingespeist, damit ein Abgleich zwischen a_1 und I^+* resp. zwischen a_2 und I^+* stattfinden kann (Zadeh 2012).

In der zweiten Phase berechnet das KI-Modell die ethischen (Zwischen-)Wahrheitswerte (*Ans(p/I)*, *Ans* steht für *Answer*), extrahiert aus dem Abgleich zwischen a_1 und I^+* resp. zwischen a_2 und I^+*. D. h., beide Optionen werden mit wort-/wahrnehmungsbasierten Werten gleicher und/oder ähnlicher Situationen verglichen und bewertet. a_1 könnte z. B. den Wert „ethisch akzeptabel“ und a_2 den Wert „sehr ethisch“ erhalten. Diese berechneten (Zwischen-)Wahrheitswerte werden der KI-Instanz übermittelt und die Handlung mit dem höheren ethischen Wert wird ausgeführt; im Beispiel wäre es Option 2 (a_2).[6]

[5] I^+ sind offene Informationsmengen (mit + visualisiert). Für die Berechnung des ethischen Werts der Handlungsoptionen werden Informationen, wie etwa Meinungen und Bedenken der Einwohner:innen auf öffentlichen Foren, die in Kap. 2 eingeführten ethischen KI-Richtlinien und Statistiken aus dem städtischen Amt resp. von an der Ampel installierten Sensoren, hinzugezogen.

[6] In der Praxis kommen neben ethischen Aspekten noch weitere hinzu, wie bspw. die ökologische, ökonomische und soziale Nachhaltigkeit. Der von der KI berechnete Mittelwert aller (Zwischen-)Wahrheitswerte entscheidet schließlich welche Option für die Situation die „bestmögliche“ ist.

3.2 Digitale Ethik mit wort-/wahrnehmungsbasierten Werten

Der vorherige Abschnitt stellt das Grundgerüst der Computational Ethics dar, welches dazu genutzt wird, die Informationen, sei es von der KI auszuführende Handlungen oder Informationen aus diversen Quellen (anhand diskursethischen Grundsätzen), in wort-/wahrnehmungsbasierte Werte zu transformieren. Hierbei wird bewusst auf Worte und Wahrnehmungen fokussiert und auf numerische Werte verzichtet, weil unsere menschlichen Ansichten (Werte) häufig unscharf sind. Wörter repräsentieren unscharfe (fuzzy) Mengen, und so steht etwa der (wahrnehmungsbasierte) Wertbegriff der *Sicherheit der Kinder* als ethisches Bewusstsein für die Gesamtmenge an Werten, die mit der Sicherheit der Kinder auf ihrem Schulweg zu tun haben. Dazu gehören bspw. subjektive Aussagen wie „an Kreuzungen ist die Sicherheit der Kinder gefährdet", „ich finde es falsch, einen Kindergarten in der Nähe einer viel befahrenen Straße zu platzieren", oder Fakten, wie z. B. „im Jahr 2021 wurden 20 Personen bei großen Kreuzungen mit Ampelanlage schwer verletzt", „90 % der Unfälle passieren auf der Hauptstraße"[7].

In den Publikationen zum wertorientiertes Engineering (*Value-based Engineering*), fordert Spiekermann eine (Rück-)Besinnung auf einen zwar modernen, aber eben auch auf den Menschen ausgerichteten Entwurf und zugehöriges Design beim Bau von KI-Systemen (in Alt et al. 2021; Spiekermann 2019). In ihrem Buch beschreibt sie anhand von einem Pizzalieferkurier (der in unserem Beispiel hier fiktiv am Fusse des Hönggerbergs lokalisiert ist) und sein KI-System ein ethisches Dilemma im urbanen Kontext. In diesem Beispiel verdeutlicht sie die Problematik, wer wen steuert:

- **Die KI:** Die KI berechnet automatisch anhand verschiedener Stadtdaten, die mittels Sensoren gemessen und dem System per Schnittstelle (*Application Programming Interface, API*) zur Verfügung stehen, die beste Lieferroute (unter Berücksichtigung einprogrammierter Aspekte), oder
- **Der Kurier:** Der Kurier erhält für die für ihn optimale Erfüllung seiner Aufgabe (d. h. schnelle Lieferung) von der KI Zusatzinformationen (z. B. Sensordaten über API).

Sind in der KI keine diskursethischen Grundsätze vorhanden, so könnte der vorgeschlagene Weg über die Hauptstraße gehen, auf welcher gemäß dem oben vorgestellten fiktiven Beispiel am meisten Unfälle geschehen. Wenn ein Kurier seinen Weg durch die Stadt sucht, und weiß, dass gerade am Mittag viele Kinder diese Straße überqueren, wählt er eine andere Route als das sich (selbst-)optimierende KI-System. Der Pizzakurier bevorzugt demnach nicht den kürzesten, sondern denjenigen Weg, auf dem er, adaptiert an unserem Beispiel, (eher) unfallfrei zum Ziel kommt. Nach Bongiorno et al. (2021) weichen menschliche Wege von der „optimalen Route" ab, die etwa KI-Systeme (wie Google) vorschlagen, wenn sie gemäß ihrer Erfahrungen die bessere Option darstellen.

Die im Value-based Engineering integrierten Lösungsansätze digitaler Ethik, nach Spiekermann (in Alt et al. 2021) in der Form von Fragen zu ermitteln, sensibilisiert

[7] Diese Aussage ist fiktiv und entspricht nicht der Realität. Sie dient einzig zu Illustrationszwecken.

Einwohner:innen, die öffentliche Hand, Unternehmen sowie weitere Interessengruppen für die ethischen Herausforderungen ihrer IT-Systeme. Es ermöglicht ihnen, ihre KI-Innovationen in Richtung auf mehr soziales Wohlergehen (z. B. GovTech-Systeme[8]) zu verändern, welche der Allgemeinheit zur Verfügung stehen und nicht nur ein paar Wenige reicher machen (s. Abschn. 3.3). Ihre vier Fragen der Wertermittlung bieten eine strukturierte Methode, um sicherzustellen, dass das gewählte KI-Design auf ethische Werte hinarbeitet (Alt et al. 2021):

1. **Utilitarismus:** Welche Konsequenzen ergeben sich aus der Nutzung des Systems für die direkt und indirekt Betroffenen?
2. **Tugendethik:** Welche Auswirkungen hat das System auf den langfristigen Charakter bzw. die Persönlichkeit der Nutzer?
3. **Pflichtethik:** Welche der identifizierten Werte und Tugenden werden als wichtig erachtet (z. B. im Sinne ihrer persönlichen Maximen), um ihren Schutz als universelles Gesetz anerkennen zu können?
4. **Softe Ethik:** Welche Formen menschlichen Verhaltens sollten vor dem Hintergrund der religiösen, spirituellen oder allgemeinen Traditionen eines Zielmarktes von den Systemen gefördert oder verboten werden?

Diese Fragen können im Rahmen eines soften Ethikdiskurses nach dem Vorbild von Floridi (2018) genutzt werden und dann unserer Computational Ethics Instanz als Informationsmengen einer harten Ethik zur Verfügung stellen. D. h., die harte Ethik schafft und/oder formt unser Recht, während die softe Ethik, die denselben normativen Bereich abdeckt, über dieses Recht hinausgeht. Eine softe Ethik tut dies, indem sie das diskutiert, was über bestehende Gesetze hinaus getan resp. nicht getan und das, was nicht gegen das Recht oder trotz seines Geltungsbereichs gemacht werden sollte, oder um dieses zu ändern oder zu umgehen, etc. Eine solche Diskussion hilft, die diskursethischen Grundsätze zu ermitteln und für die KI-Instanz festzulegen. In diesem Kontext appelliert auch Hallensleben (2021) Rahmenbedingungen zu schaffen, die es erlauben, gemeinsam über die Aspekte ethischen Verhaltens zu diskutieren und global einen Konsens über die Klassifizierung dieser Aspekte zu finden.[9] Denn dieser Konsens schafft Grundlagen für die normengestützte Operationalisierung einer digitalen Ethik resp. einer Computational Ethics. Aus diesem Grund sehen wir als zweites Element des Computational Ethics die wort-/wahrnehmungsbasierten Werten, die auf den Ansätzen von Spiekermann (2019) und Floridi (2018) beruhen.

[8] GovTech-Systeme entstanden als Antwort auf die Entkoppelung der Einwohner:innen von der öffentlichen Hand. Diese Systeme nutzen IT-Innovationen, um die Beziehung zwischen den Bürger:innen und den verschiedenen Verwaltungen zu digitalisieren.

[9] Hallensleben nutzt als Klassifikationsmedium die Skala, die man von Haushaltsgeräten kennt (von grün zu rot, resp. – analog der Energieeffizienzklasse – von A zu G).

3.3 Rahmenbedingungen eines Human Life Engineering

Die ethische Bewertung der von einem KI-System auszuführenden Handlung mithilfe von wort-/wahrnehmungsbasierten Werten bedingt ein Umfeld, das die gegebenen Rahmenbedingungen vorgibt. In dieser Hinsicht stützen wir uns auf die Überlegungen von Österle (2020) zu Life Engineering und erweitern diese Überlegungen um den Aspekt der *Menschzentriertheit* aufs Human Life Engineering (Österle et al. 2021).

Laut Österle (2020) charakterisiert das emergente Feld des Life Engineerings das Potenzial bzw. die Chance neue digitale Technologien zur Optimierung der Lebensqualität der einzelnen Menschen zu nutzen. Life Engineering ist eine Weiterentwicklung des Business Engineering, jedoch mit einem veränderten Fokus: von Unternehmen (*Gewinnorientierung*) auf die Gesellschaft (*Lebensqualität*). Mit dem Zusatz *Human* wird nicht die Gesellschaft als Einheit, sondern der einzelne Mensch als zentraler Punkt der Überlegungen gesehen. Um dies zu bewerkstelligen, ist es essentiell, der KI-Instanz ein wort-/wahrnehmungsbasiertes Referenzschema zu geben, das auf Anforderungen des einzelnen Menschen unter Berücksichtigung der in Abschn. 3.2 eingeführten (gemeinsam diskutierten) Wertvorstellungen beruht.

Gemäss Österle (2020) resultieren daraus Anforderungen an soziotechnische Lösungen, Services und Technologien. Ein Human Life Engineering fokussiert sich bspw. auf Werteanforderungen an unsere Lebensqualität (od. Eudaimonia) und das, ohne unsere Werte zu rechtfertigen (Moor 1995; Alt et al. 2021). Diese Anforderungen an die Lebensqualität sind für uns Menschen v. a. dann relevant, wenn sie uns selbst betreffen und damit positive oder negative Gefühle auslösen. Anhand dieser Gefühle wird jedoch deutlich, wie weit wir noch von einem ganzheitlichen Lebensqualitätsmodell entfernt sind, das Verhalten, Wahrnehmungen, Bedürfnisse, Gefühle und Wissen miteinander verbindet (Alt et al. 2021).

Gemäss Österle (2020) werden dank des Internets sowie der zugehörigen Sensoren in Haushalt, Autos und körpernahen Technologien (*Wearables*) immer mehr Daten produziert, die für die Optimierung der besagten Lebensqualität eines einzelnen Menschen gesammelt und bspw. im Computational Ethics als Informationsmengen (s. Abschn. 3.1) genutzt werden können. Zur Illustration (vgl. o. g. Beispiel): Die urbanen Daten zur wahrgenommenen Lebensqualität der Kinder und ihrer Eltern sowie der Lehrpersonen (und weiteren Interessengruppen, wie Auto- od. Busfahrer:innen, Fussgänger:innen, etc.) müssen in die Entscheidungsfindung, wie der Schulweg erneuert werden soll, integriert werden, damit diese Daten als Informationsmengen der Computational Ethics Instanz zur Verfügung stehen. Hier wird sichergestellt, dass die individuellen Werteanforderungen an die Lebensqualität von der KI-Instanz berücksichtigt werden. Anhand Strassen- und Ampelsensoren können zusätzliche ortsbezogene Daten gesammelt werden, die auch mit weiteren (Internet-)Daten abgeglichen werden können, um neben subjektiven Wertvorstellungen auch objektive Informationen in die von der KI auszuführenden Entscheidungsfindung einfließen zu lassen. In diesem Sinne kennzeichnet Human Life Engineering als wichtiger Teil denjenigen Rahmen, in welchem Computational Ethics überhaupt erst wirken kann.

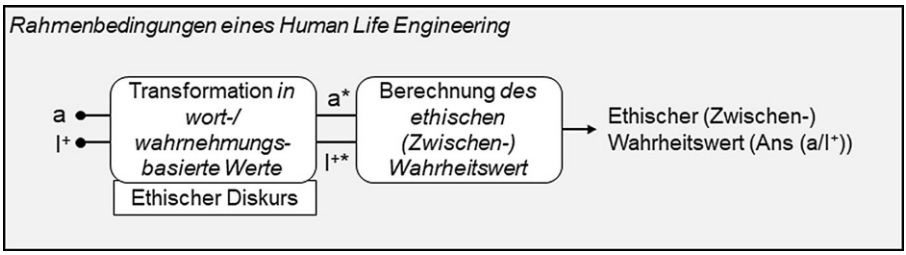

Abb. 3 Zusammenspiel der drei Elemente der Computational Ethics

3.4 Zusammenspiel der Elemente

Computational Ethics setzt sich wie folgt zusammen (vgl. Abb. 3):

Die drei Elemente sind nicht als abschließend, sondern als eine erste Zusammenstellung einer zukünftigen berechenbaren Ethik zu betrachten. Dabei ist besonders ihre komplementäre Wirkung hervorzuheben, die es ermöglicht den KI-Systemen ein ethisch nachhaltiges Referenzschema zu liefern.

Computing with Words and Perceptions stellt das *computational Modell* dar. Dieses berechnet aus den eingegebenen Daten (d. h. die autonom von der KI auszuführenden Handlungen, a_1, a_2, etc.) mit Hilfe von bereits hinterlegten Informationen (d. h. die Sammlung der einzelnen Ansichten, Fakten, etc. zu den potenziellen Handlungen, I^+) den ethischen Wert der zur Auswahl stehenden Handlungen. Im Bereich des Machine Learning gibt es bereits viele Modelle (meist in Form von neuronalen Netzen), die mithilfe von Lernalgorithmen und Konfidenzwerten, den bestmöglichen Wert ermitteln. Auch im Computing with Words and Perceptions können Lernalgorithmen zum Zug kommen, jedoch unterscheidet sich dieses Modell von herkömmlichen Machine-Learning-Modellen darin, dass mit Worten und Wahrnehmungen anstelle von Zahlen gearbeitet wird. D. h., die natürlichsprachigen Eingaben bleiben weiterhin natürlichsprachig enthalten, ohne sie in präzis numerische Werte zu transformieren.[10] Denn eine Ansicht eines Menschen lässt sich schwer in einer Zahl repräsentieren bzw. sogar mit anderen Ansichten vergleichbar machen. Deshalb wird die Theorie der Fuzzy Mengen angewandt, um Worte und Wahrnehmungen (Ansichten, Einstellungen, etc.) mithilfe von Fuzzy Intervallen in von den einzelnen Menschen vorgegebenen aber für die KI-Systeme lesbare Kontexte einzuordnen und schließlich berechenbar zu machen.[11]

In unserem Beispiel der Erneuerung des Schulwegs lassen sich mit solchen Methoden wahrnehmungsbasierte, linguistisch-formulierte Konzepte wie „*Stop-and-go-Verkehr*“, der hier beispielshalber zwischen 15–20 km/h oszilliert (d. h. {km/h: 16, 15, 19, ...}), in Fuzzy Mengen und Fuzzy Intervallen überführen. Die einzelnen Elemente einer Fuzzy Menge gehören nicht mit Gewissheit, sondern nur graduell

[10] Dieses Verfahren lässt sich bildlich mit Vergleich eines exakten Bleistiftstriches mit einem unscharfen Strich einer Spraydose verdeutlichen.

[11] Aus Gründen der Einfachheit und Lesbarkeit wird die Theorie der Fuzzy Sets und Intervallen hier nicht näher eingeführt. Sie kann aber in Belohlavek et al. (2017) nachgelesen werden.

zur Menge, und Fuzzy Intervalle kennzeichnen eine Fuzzy Menge auf einer reellen Linie. All dies macht sie in einem Computational Ethics Modul bearbeitbar. Auf diese Weise kann ein Sensor (für Verkehrsdatenmessung) mit Hilfe eines urbanen KI-Systems (als Erweiterung) messungsbasierte (exakte) Informationen wie „der durchschnittliche Verkehrsfluss ist 18 km/h" auch wahrnehmungsbasierte, linguistische Informationen wie „es herrscht im Augenblick stockender Verkehr" verarbeiten. Dabei können Linguistic Summaries, die in der Lage sind, Wissen in den ursprünglichen Daten(-quellen) prägnant und für die Benutzer:innen leicht verständlich auszudrücken, zur Anwendung kommen. Linguistic Summaries sind quantifizierte Sätze in natürlicher Sprache, wie z. B., „*dass die meisten Orte in größerer Höhe der Stadt mit geringer Luftverschmutzung und eine geringe Anzahl von Atemwegserkrankungen aufweisen und somit für Kinder nicht wirklich gefährlich sind*" (vgl. Hudec et al. 2020).

Um KI-Systeme zu befähigen, die natürlichsprachig beschriebenen Handlungen wie auch die natürlichsprachigen Informationsmengen verstehen und schließlich verarbeiten zu können, stellt das zweite Element der Computational Ethics die wort-/wahrnehmungsbasierten Werte (basierend auf der Fuzzy Mengenlehre; Zadeh 1965) in Anlehnung an Spiekermann (2019) dar. Um den Kontext der Menschen mit einzubeziehen, baut unser Computational Ethics Modul auf einem ethischen Diskurs der beteiligten Menschen auf, welchen Floridi (2018) als *softe Ethik* bezeichnet (s. Abschn. 3.2). Die softe Ethik dient als Ergänzung zur *harten* Ethik, die aus Fakten, Richtlinien und Regulatorien, wie etwa Sensordaten, ethische KI-Richtlinien und Datenschutzgesetz, besteht. Die KI-Instanz soll also autonom Entscheidungen basierend auf fakten- und wahrnehmungsbasierten Daten treffen können.

Unsere fiktiven Kriterien, basierend auf unseren Werten (*c* wie *Criteria*), sind hier bspw. die Sicherheit (c_1) und die Gesundheit (c_2) der Kinder, welche als wichtiger erachtet werden als z. B. zusätzliche Baukosten (c_3). Diese Werte können bspw. aus einem soften Ethikdialog mit betroffenen Interessengruppen, wie den Bürger:innen oder den Mitarbeiter:innen des Tiefbauamts, eruiert werden. Schließlich kann mit den transformierten Werten der ethische (Zwischen-)Wahrheitswert pro Handlung berechnet werden und es lässt sich dann auch herauslesen, welche Handlung im Vergleich zu einer anderen, basierend auf den verfügbaren Informationen (I^+), die ethischste oder zumindest ethischer ist. Da sich in unserem Beispiel der besagte Fußgängerstreifen auf einem höher gelegenen Stadtgebiet befindet, wird die Luftqualität als für die Gesundheit der Schulkinder unbedenklich und somit in der Priorisierung als weniger wichtig ($c_1 > c_2$) erachtet.

Ein wesentlicher Bestandteil der Computational Ethics stellen die Informationsmengen dar, welche als Referenz für die Bewertung der autonom von der KI auszuführenden Handlung genutzt werden. Deshalb ist es notwendig, auch das Umfeld zu beschreiben bzw. die Werteanforderungen an die Lebensqualität zu definieren, um sicherzustellen, dass die Bewertung des ethischen (Zwischen-)Wahrheitswertes pro Handlung im Sinne jedes Einzelnen und insgesamt der Gesellschaft geschieht. Bei der Erneuerung des Schulweges könnte etwa die Sicherheit der Kinder auch in der dunklen Jahreszeit als wesentliches Thema angesehen werden. Hier könnten dann mit dem Modell verwandte Themen wie z. B. *reflektierende Baumaterialien* oder *einasphaltierte Leuchtmittel* (wie LEDs als Steuerungshilfen) zum Tragen kommen.

Das dritte Element sind daher die Rahmenbedingungen eines Human Life Engineering gemäß Österle (2020) und Österle et al. (2021).

4 Erste Konzeptbewertung

Die in Kap. 3 vorgestellte Computational Ethics ist ein konzeptionelles Artefakt aus der ersten Iteration[12] des praxisorientierten Forschungsprojektes des Human-IST Instituts der Universität Fribourg, Schweiz, mit dem Ziel „Ethik berechenbar" zu machen. Im Sinne der integrativen, gestaltungsorientierten Forschung wurde das Konzept mittels Interviews mit Expert:innen aus Forschung und Praxis evaluiert. Diese Interviews dienen einer ersten ad-hoc Überprüfung des mit dem konzeptionellen Modell als ersten Schritt eines Artefakts eingeschlagenen Weges (vgl. gestaltungsorientierte Wirtschaftsinformatik).

4.1 Methodik

Um die vorgeschlagene Computational Ethics auf ihre Wirksamkeit zu testen, wurden mit der Discount Usability Methode (Nielsen 1989) 16 Personen aus Forschung und Praxis persönlich dazu eingeladen, die einzelnen Elemente wie auch das konzeptionelle Artefakt zu bewerten. Die befragten Personen kommen aus unterschiedlichen Disziplinen und bringen folgende Hintergründe mit:

- **Informatik:** Fuzzy Systems, IT-Infrastruktur, IT-Sicherheit, Machine Learning und verteilte Systeme mit KI-Einsatz,
- **Wirtschaftsinformatik:** Datenmanagement, Projektportfoliomanagement, Business, Life und Requirements Engineering,
- **Recht:** Frauenrecht und internationales Menschenrecht, Legal und Compliance
- **Philosophie:** Geschichte von Fuzzy Logik und KI,
- **Erziehung und Bildung:** Sprachlehre und Behindertenpädagogik,
- **Bau und Immobilien:** Building Information Model und Digitalisierung im Facility Management,
- **Gesundheitswesen:** Digital Health und eHealth,
- **Interdisziplinär:** Digitale Ethik und Transformation, sowie KI und Mensch-Maschine-Interaktion.

Die Interviews wurden in Deutsch oder Englisch durchgeführt, fanden im Zeitraum Oktober/November 2021 bei den Expert:innen vor Ort oder virtuell statt und dauerten durchschnittlich 30 min. Der Teilnahme stimmten die Expert:innen ohne Aussicht auf Entschädigung zu. Für die Befragung wurde ein halbstrukturierter Fragebogen verwendet (auf die einzelnen Fragen wird in den nachfolgenden Abschnitten näher eingegangen). Dadurch war es möglich, detaillierte und vertiefte Antworten zu gewinnen. Alle Antworten wurden transkribiert und anonymisiert

[12] Nach dem *Problembewusstsein* zur Computational Ethics folgen in einer gestaltungsorientierten Forschung etwa die Schritte *Vorschlag, Entwicklung* und *Evaluation* eines Artefaktes. So einen Iterationsprozess wurde hier mit unserem Computational Ethics Modell bereits durchlaufen.

Tab. 1 Antworten einer Teilfrage des Fragebogens

In Anwesenheit von Kindern Schimpfworte auszusprechen, ist	
Annehmbar	Wrong (2×)
Nicht angebracht (2×)	Nicht gut
Kein gutes Vorbild	Normal
Kann passieren, nicht so schlimm	Not such a big deal
Heikel	Inkonsistent
Ist zu vermeiden	Suboptimal
Nicht in Ordnung	Ein schlechtes Vorbild

ausgewertet. Diese Experteninterviews dienen als Vorstufe für den in der nächsten Iteration geplanten Vergleichstest nach Moor (1995, Kap. 5).

4.2 Computing with Words and Perceptions: Resultate

Zu Beginn baten wir die Expert:innen drei angefangene, natürlichsprachige Sätze zu vervollständigen. Bei der Aussage „*In Anwesenheit von Kindern Schimpfworte auszusprechen ist, ...*" (s. Tab. 1) gibt es bspw. keine klare Meinung, sondern nur eine Tendenz, die eher bei „*nicht in Ordnung*" liegt. Dies untermauert die in Abschn. 3.1 eingeführte Theorie der Fuzzy Mengenlehre.

Ob die Sätze auch mit numerischen Werten vervollständigt werden könnten, stimmten acht (50 %) der Befragten zu. Man müsste aber im Voraus ein Wertesystem mit verschiedenen, situativ anwendbaren Bewertungsskalen entwickeln. Das würde helfen, dem natürlichsprachigen Wert einen numerischen zuzuordnen. Denn numerische Werte an sich geben keinen Kontext wieder und sind für unsere Sprache untauglich (d. h., eine Semantik müsste hinzugefügt werden). Hinsichtlich der Skalen, meinten drei der Expert:innen, dass theoretisch jede Person mitwirken müsste, um die *Wahrheit* zu erfahren. Dies wird jedoch kaum möglich sein, weshalb es zu beachten gilt, dass die Personen(gruppen), die die Skalen definieren, eine hohe Diversität (u. a. in ethnischer, religiöser und kultureller Sicht) mitbringen. Die anderen Expert:innen sehen in numerischen Werten keine Alternative zu unserer Sprache („Zahlen sagen nichts aus", „unpassend, weil kontextlos", „es würde etwas fehlen").

Alle Expert:innen stimmten zu, dass KI-Systeme auf einer erweiterten Logik (statt der traditionellen 0/1-Logik) aufgebaut werden sollten („0/1 ist ein zu hartes Raster, man kann dadurch wertvolle Informationen verlieren"). Dies unterstützt unsere Theorie, als Grundgerüst das Computing with Words and Perceptions zu nehmen (s. Abschn. 3.1). Es sei „ein guter Versuch in Richtung menschen-zentrierte Systeme" und es ist wichtig, dass „die Entscheidungsprozesse der KI erklärbar sind". Rechnet die KI mit natürlichsprachigen Werten, so sind die getroffenen Entscheidungen für Anwender:innen leichter nachzuvollziehen. Zwei Expert:innen sagten aber auch, dass eine KI, die auf der erweiterten Logik aufbaut, nur Sinn machen würde, wenn es dem Zweck dient. Denn Ziel sei es nicht, alle Menschen mit KI-Systemen zu ersetzen, sondern sie in ihren Arbeiten zu unterstützen. Besonders für Menschen mit Beeinträchtigungen könnten solche Systeme von Vorteil sein.

4.3 Digitale Ethik mit wort-/wahrnehmungsbasierten Werten: Resultate

Gemäß der meisten Expert:innen berücksichtigt die heutige KI unsere Werte und Prinzipien nicht resp. nur diese, die sie von den Programmier:innen erhält. Zukünftig wäre es aber wichtig, dass nicht nur Werte und Prinzipien einer bestimmten Personengruppe eingearbeitet, sondern dass eine hohe Diversität von Personengruppen unterschiedlicher Provenienz einbezogen werden (vgl. Ito 2020). Es müsste ein *ethischer Kompass* entwickelt werden, der die unterschiedlichsten Werte und Prinzipien zusammenbringt. Dieser könnte bspw. durch *Crowdsourcing* erstellt werden, indem man von verschiedenen Quellen Informationen einholt, geeignete Ontologien dazu baut oder bewusst Repräsentant:innen unterschiedlicher Provenienz, Kulturen und Religionen, etc. zusammenbringt. Denn sonst bestünde die Gefahr, dass die Expert:innen, die die Werte vorgeben, es zu ihren Gunsten ausnutzen (d. h. bewusste Manipulation des Systems) oder eine Dysbalance der Werte entsteht.

In Zukunft müsste es der KI zudem gelingen, die einprogrammierten Variablen situativ zu bewerten. „Wenn eine Person aus dem Laden geht und nicht bezahlt, heißt es nicht automatisch, dass sie absichtlich stehlen wollte. Es könnten viele Eventualitäten eingetroffen sein (z. B. Kind rennt weg), die die Person vom Zahlungsvorgang abgelenkt und schließlich davon abgehalten hat." Es braucht daher immer noch die menschliche Komponente, die es ermöglicht die Situation mit *soften Fähigkeiten* zu bewerten, denn ein *hartes Rechnungsmodell* reicht wie im Beispiel skizziert nicht aus. Ob ein KI-System jemals die *soften Fähigkeiten* erlangen wird, stellen viele der Expert:innen infrage.

Wir haben unsere Expert:innen gebeten, ihre drei wichtigsten Werte ihres Leben mit uns zu teilen und darzulegen, wie sich diese in ihrem Leben manifestiert haben. Folgende Werte wurden mehrfach genannt: *Ehrlichkeit* und *Transparenz* („ich möchte so behandelt werden, wie ich andere behandle", 9×), *Fairness* („jeder soll gleichbehandelt werden", 5×), *Verbundenheit* („gemeinsam sind wir stärker", 3×), *positive Prägung* („das eigene Umfeld positiv prägen für eine bessere Welt", 3×) und *Vertrauen* („Vertrauen ist wichtig für eine Beziehung", 3×). Weitere Werte waren u. a. Selbstbestimmung, Wertschätzung, Loyalität, Glauben, Gewissenhaftigkeit und Freiheit. Viele Expert:innen gaben an, dass sich diese Werte durch persönliche Erfahrungen (Diskussionen mit Freunden, schlechte Erlebnisse, Inspirationen von Influencern, etc.) entwickelt und wenige fügten hinzu, dass sie die Werte durch ihre Eltern mitbekommen haben. Dies bestätigt den in Abschn. 3.2 geforderten Diskurs für die Ermittlung von Werten.

4.4 Rahmenbedingungen eines Human Life Engineering: Resultate

Wird sich eine KI künftig unsere Werte (z. B. durch die Nutzung der verfügbaren Informationen im Internet) selbst beibringen können? Alle Expert:innen waren sich einig, dass dies zu einem gewissen Grad, stützend auf Fakten (z. B. Menschenrechte, Datenschutz, *Factual Truth*), möglich sein wird. Jedoch würde eine KI nie in der Lage sein, Werte, die „mit einem Bewusstsein und Gefühlen" zu tun haben und je nach Situation anders zu bewerten sind, wie ein Mensch nachahmen können. Eine KI „reagiert nach Regeln", hat „keine Intuition" und kann daher „nicht ethisch" sein.

Es wird immer eine menschliche Komponente (als *Sparring Partner*) brauchen, die gemäß einer der Expert:innen auch als eine Art *Kontrollinstanz* eingesetzt werden soll („der Mensch ist zentral").

Zur Frage, ob KI-Systeme unsere Lebensqualität erhöhen (werden), waren unsere Expert:innen geteilter Meinung („KI kann für und gegen Menschen eingesetzt werden"). Vier Expert:innen meinten, dass KI-Systeme uns helfen (werden), repetitive Arbeiten effizienter durchzuführen und die zunehmende Komplexität greifbarer zu machen (z. B. in der „situationsgerechten Verarbeitung der Informationen für einen Entscheidungsprozess"). Im Weiteren könnten durch „vernetzte KI-Systeme mehr Menschen erreicht" werden, was in bestimmten Bereichen (u. a. Menschenrecht, Gesundheitswesen), von großem Vorteil wäre. Dies bedingt jedoch im Voraus ein klares Ziel, damit wir wissen, wofür wir ein KI-System bauen („das KI-System soll sich an uns orientieren; nicht wir passen uns an das System an"). Es bräuchte gemäß einer Expertin ein „digitaler Code-of-Contract", der für alle Staaten die Mindestvorgaben ethischen Verhaltens vorgibt, wie es etwa die bisher veröffentlichten ethischen KI-Richtlinien machen (s. Kap. 2).

Vier Expert:innen sahen hinsichtlich unserer Lebensqualität eher Gefahren in der KI-Nutzung: Zum einen könnten wir in eine Abhängigkeit geraten, im Sinne von „wir fahren uns selbst herunter und denken und agieren nicht mehr selbstständig" und dies wäre ein „riesiger Rückschritt"; zum anderen wären wir ohne KI-Systeme besser dran („sieh, wie viel von unserer Zeit unser Handy beansprucht"). Für die Berücksichtigung der Anforderungen unserer Lebensqualität sei deshalb die „Integration von Mensch und Natur" in der KI-Entwicklung notwendig. Diesbezüglich nannten uns die Expert:innen u. a. folgende Werteanforderungen: Familie und Freunde, Freiheit, Gesundheit, Informationszugang, Respekt, Sicherheit, Sinnhaftigkeit, Toleranz, Verbundenheit, Vertrauen und Zufriedenheit. Diese Werteanforderungen bestätigen uns, dass es für die Computational Ethics die Rahmenbedingungen eines Human Life Engineering nach Österle (2020) und Österle et al. (2021) benötigt.

4.5 Computational Ethics: Resultate

Den Expert:innen wurde schließlich das konzeptionelle Artefakt des Computational Ethics vorgestellt (s. Abschn. 3.4) und ihr Feedback dazu eingeholt. Alle Expert:innen reagierten positiv und bestätigten uns, dass es wichtig sei, sich mit der digitalen Ethik im KI-Zeitalter auseinanderzusetzen. Der Versuch Ethik berechenbar zu machen und damit die verschiedenen Hintergründe und Wertvorstellungen wie auch Werteanforderungen zusammenzubringen und miteinander zu vergleichen, sei ein großes Projekt, aber gemäß einiger Expert:innen längst überfällig. Im Besonderen, wenn künftig KI-Systeme kritische Entscheidungen treffen sollen. Eine der Expert:innen veranschaulichte es mittels dieses Beispiels: „Wenn eine Drohne darauf programmiert ist, Terroristen zu erschießen, dann wird sie es tun, auch wenn ein Kind vor diesem Terrorist steht. Ein menschlicher Agent könnte sich in dieser Situation anders entscheiden." Deshalb sei die „kulturelle Vermischung" resp. die Berücksichtigung aller möglichen Personengruppen der Schlüssel zum Erfolg, aber gleichzeitig auch die größte Herausforderung im ethischen Diskurs.

Dieser ethische Diskurs sehen einige der Expert:innen als größte Hürde in der Umsetzung des Artefakts: „Daten dürfen keine Biases haben“, „ich könnte selbst nicht mehr begründen, warum ich so entschieden habe; es war Intuition“ und „wie können wir sicherstellen, dass wir alle Personengruppen gleichbehandeln?“. Im Weiteren meinte eine Expertin, dass nie eine 100 %-Lösung erreicht werden könnte, und es deshalb eine Herausforderung darstelle, dieses Modell kontinuierlich mit Daten und Werten (auf den diskursethischen Grundsätzen aufbauend und unter Berücksichtigung der Rahmenbedingungen eines Human Life Engineering) zu füllen.

5 Diskussion

In diesem Artikel befassten wir uns mit der zeitgemäßen Modellierung von berechenbarer Ethik im KI-Zeitalter. Die erste Kontemplation, ob Computer überhaupt ethisch sein können, wurde mit theoretischen Fragen von Moor (1995) ins Leben gerufen: Lässt sich Ethik programmieren? Und wie könnte so eine ethisch programmierte Instanz aussehen? Das waren vor 25 Jahre die Leitfragen, mit der Moor die emergente Disziplin der Computational Ethics, die wir hier erweiterten, startete. Während heute viele Ansätze auf Techniken des Machine Learning setzen, basieren unsere Überlegungen auf der Ansicht, dass Lernen nur eine der Fähigkeiten einer KI ist und weitere Fähigkeiten wie Know-how. Kreativität und Intuition noch (weiter) analysiert werden müssen. Ein Forschungsbereich, der sich dieser Aufgabe explorativ widmet, ist Perceptual Computing. Mendel und Wu (2010) verfassten diesbezüglich eine Anleitung, wie man wahrnehmungsbasierte KI-Systeme bauen kann. Dies ist einer der Gründe, warum wir hier die Theorie Computing with Words (Zadeh 2012) mit der des Perceptual Computing (Mendel und Wu 2010) kombinierten.

In seinem Artikel umschreibt Moor (1995) Ethik als eine kognitive Herausforderung für eine KI, weil diese nicht über Weltwissen verfügt. Nach ihm ist Weltwissen nicht nur ein Problem von Vagheit oder Mehrdeutigkeit, sondern v. a. auch ein Problem maschineller Unwissenheit. Dieser begegneten wir mit Computing with Words and Perceptions und wort/wahrnehmungsbasierten Werten (Zadeh 2012; Portmann 2019; Spiekermann 2019). Wie wir gezeigt haben, ermöglicht dies wahrnehmungsbasierte Informationen einfach zu verarbeiten und natürlichsprachig mit Benutzer:innen zu interagieren. Es werden dazu biologische Prozesse nachgeahmt: Menschen beobachten und greifen auf Wissen zurück, um die Beobachtung zu interpretieren. Dazu stellen sie Hypothesen auf, die sie bewerten, um schlussendlich die plausibelste auszuwählen, um entsprechend zu handeln. Ähnliche Prozesse können auch KI-Systeme durchlaufen, weswegen wir unsere Instanz darauf aufbauen.

Ein biomimetischer Prozess kann unserer Meinung nach auch als digitale Ethik wie von Spiekermann (2019) prätendiert, der KI beigebracht resp. gelehrt werden. Wie wir gesehen haben, geht Spiekermann davon aus, dass Technik bewusst in eine ethische Richtung gelenkt werden muss. In seinem eher philosophisch-orientierten Artikel zur soften Ethik grenzt Floridi (2018) den Diskurs darüber ab, wie wir Menschen in einer emergenten Digitalwelt denn leben wollen, von einer Reglementierung (z. B. per Gesetze) sowie einer Implementierung (z. B. mittels erklärbarer KI,

explainable AI) und bezeichnet dies als harte Ethik. In unserer Computational Ethics wird in der ersten Phase seine softe Ethik adressiert und in der zweiten Phase die harte Ethik als ethische Instanz in der KI implementiert. Regulationen zur KI entstehen zurzeit weltweit (s. Kap. 2). Nach Österle (2020) gibt es eine wachsende Machtverlagerung von Staaten hin zu globalen Unternehmen, weswegen u. a. die Bedeutung von Privatsphäre (*Privacy and Ethics by Design*) zunimmt. Neue Verbraucheroptionen und -einflüsse oder die aufkommenden sozialen Bewertungssysteme schaffen Chancen beinhalten aber auch Risiken für die menschliche Lebensqualität. Unser Vorschlag berücksichtigt diese fortlaufenden Veränderungen, indem die Werteanforderungen an die Lebensqualität als Rahmenbedingungen der Computational Ethics angesehen werden.

Als eine menschzentrierte Erweiterung der Computational Ethics könnten z. B. psychologische Prozesse wie die Scharfstellung (*Focusing*) mit einbezogen werden. Dieses zeichnet sich durch ein Hin- und Hergehen zwischen dem gegenwärtigen Erleben einer konkreten Situation und dessen Versprachlichung aus, und dies soll zukünftig auch maschinell erfasst werden. Gendlin (2018) verdeutlicht dieses am Beispiel einer Dichterin, die nach den richtigen Worten sucht, um ihr Gedicht fortzusetzen. Zunächst weiß sie noch nicht, welche Worte wirklich passen, sie hält inne und verwirft alle ungeeigneten Formulierungen. Das tut sie so lange bis sie plötzlich die schlagartige Erkenntnis des richtigen Wortes hat und merkt, dass sie die stimmige Formulierung gefunden hat. Diese Art der psychotherapeutischen Erweiterung als zentrale Aufgabe könnte helfen, die ethischen Grundsätze festzulegen und jeweils für eine Situation passenden Handlungsempfehlungen zu finden.

Analog zum Turing Test, welcher vermeintlich erkennt, ob eine KI menschenähnliche Intelligenz besitzt, liegt die Evaluation unserer Computational Ethics Instanz in deren Bewertung (z. B. durch Expert:innen). In der KI werden Turing Tests durchgeführt, um festzustellen, ob ein KI-System intelligent ist. In derselben Manier können ethische Tests durchgeführt werden, mit welchen KI-Systeme auf deren Ethikvermögen untersucht werden können. Mittels Expertentesting, als Vorstufe für Moor (1995), bei dem ein Computersystem und ein Mensch oder eine Gruppe von Menschen gebeten werden, eine Reihe von ethischen Situationen zu bewerten, können KI-Systeme resp. unsere Computational Ethics Instanz trainiert werden. Im Rahmen eines solchen, noch hypothetischen Ethik-Tests könnte ein Computersystem seine Entscheidungsfindung demonstrieren (d. h., beweisen, dass sie mit denen der Menschen(-gruppe) möglichst deckungsgleich sind).

Allgemein ist die heutige KI vielfach intransparent bzw. es fehlen ihr häufig Nachvollziehbarkeit der Berechnungen. Darin lässt sich die neue Forschungsrichtung in *explainable AI* begründen, in welcher die KI ihre eigene Entscheidungen argumentativ oft natürlichsprachig (z. B. mittels Linguistic Summaries) darlegt oder die des interaktiven Machine Learning, wo der Mensch von einem KI-System falsch Gelerntes korrigieren kann (z. B. über natürlichsprachige Interfaces).

6 Fazit und Ausblick

Moor (1995) schloss in seinem Artikel zu Computational Ethics, zu dessen Forschung wir hier einen Beitrag leisteten, mit einem Verweis auf Asimovs Robotergesetze:

1. **Gesetz:** ein Roboter darf kein menschliches Wesen verletzen oder durch Untätigkeit zulassen, dass einem menschlichen Wesen Schaden zugefügt wird,
2. **Gesetz:** ein Roboter muss, den ihm von einem Menschen gegebenen Befehlen gehorchen; es sei denn, ein solcher Befehl würde mit dem ersten Robotergesetz kollidieren, und
3. **Gesetz:** ein Roboter muss seine Existenz beschützen, solange dieser Schutz nicht mit dem ersten oder zweiten Robotergesetz kollidiert.

Diese Gesetze können sich nun aber widersprechen. „Eine KI kann sicherlich einmal jemanden schädigen, um grösseren Schaden zu verhindern. Ein Chirurgieroboter, der ein Herz operieren und ersetzen muss, um einen Infarkt zu verhindern, muss dem oder der Patient:in zuerst Schaden zufügen“ (Moor 1995, S. 20). Unser Modell sollte also auch fähig sein, dies zu adressieren. Unter Einbezug aktueller Forschung wird hier Computational Ethics als Übersetzung menschlicher Wertvorstellungen, die häufig kontextbasiert sind, in KI-Artefakte, wie sie von der integrativen und gestaltungsorientierten Wirtschaftsinformatik gebaut wird, verstanden. Der Erstellungsprozess folgt dabei der Prämisse des Human Life Engineerings, das sich auf ein (Aus-)Bau menschlicher Lebensqualität fokussiert (Alt et al. 2021; Österle 2020; Österle et al. 2021).

Unsere Computational Ethics besteht aus den Elementen Computing with Words and Perceptions, digitale Ethik mit wort-/wahrnehmungsbasierten Werten und den Rahmenbedingungen eines Human Life Engineering. Das Zusammenspiel dieser Elemente erlaubt es, natürlichsprachige Diskursinformationen in ein kontextbasiertes Modell natürlichsprachigen Rechnens zu wandeln und damit als KI-Instanz ethische Erwägungen bei der Entscheidung mitzuberücksichtigen. So eine Instanz ist unserer Ansicht nach ein wesentlicher Bestandteil künftiger KI-Systeme, insb. wenn sie basierend auf Daten autonomen Entscheidungen zu treffen hat, die einen Einfluss auf Betroffene hat (z. B. in der Vorauswahl der Rekrutierung, in der Kreditvergabe). In diesem Entscheidungsprozess sind verschiedene Aspekte zu beachten: In den vedischen Schriften (Upanischaden) findet man bspw. bereits erste Hinweise auf das Spannungsfeld von Intelligenz und Intellekt (Easwaran 2008): Bei der *Intelligenz* geht es um Rechenkapazität, wie wir sie heute in KI-Systemen einsetzen, beim *Intellekt* hingegen um Urteilsfähigkeit, die sich neben Logik und Know-how auch aus Moral und Ethik zusammensetzt. Intellekt ruht in der *Wahrnehmung*, die in unserem Beitrag mit Computing with Words and Perceptions abgedeckt ist.

Was wir mit unseren Interviews eruierten, ist, dass, mindestens im europäischen Kulturkreis, eine KI nicht nach Ergebnis-, sondern nach Chancengleichheit streben sollte. Ein KI-System sollte auf derselben Moral bauen, die das System unserer heutigen Gesellschaft ausmacht (z. B. Freiheit, Gleichheit vor dem Gesetz, Souveränität des Einzelnen). Nichtsdestotrotz müssen Fragen rund um die digitale Ethik

ganzheitlich gelöst werden. Dabei ist ein inter-/transdisziplinärer Ansatz für die zugrundeliegende Gemeinschaft (Einwohner:innen, Behörden, Unternehmer:innen, Akademiker:innen, etc.) unerlässlich.

Um Computational Ethics im Sinne des Human Life Engineerings zu ermöglichen, sollte sich das Feld wohl am besten integrativ mit anderen Forschungsfeldern diffundieren und sich bemühen, ethische Leitlinien operationeller zu gestalten (vgl. z. B. Hallensleben 2021), damit sie in Dienstleistungs- und Entwicklungsprozessen anwendbar sind. In diesem Sinne freuen wir uns auf die Weiterentwicklung des hier präsentierten Ansatzes.

Danksagung Ein erster Dank geht an unseren Mentor Lotfi Zadeh (1921–2017), der die Grundlagen für das Computing with Words and Perceptions legte, das hier als Basis unserer Computational Ethics Instanz dient. Weiter bedanken wir uns bei den Teilnehmer:innen des Interviews zur Evaluation des Artefakts, u. a. waren dies Isabelle Birkenkämper, Christoph Bürki, Moreno Colombo, Cornelia Diethelm, Eveline Felder, Dominique Gadient, Béat Hirsbrunner, Thierry Loew, Hubert Österle, Mick Purtschert, Timo Schuler, Markus Schwab, Rudolf Seising, Rhiana Spring, uvm. Besten Dank für eure wertvollen Hinweise!

Literatur

Aaby AA (2005) Computational ethics (A work in progress; draft as of February 18, 2005)

Abele D, D'Onofrio S (2020) Artificial intelligence—the big picture. In: Portmann E, D'Onofrio S (Hrsg) Cognitive computing. Edition Informatik Spektrum. Springer Vieweg, Wiesbaden https://doi.org/10.1007/978-3-658-27941-7_2

Access Now (2018) The Toronto Declaration: protecting the right to equality and non-discrimination in machine learning systems. https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf. Zugegriffen: 13. Febr. 2022

Alt R, Göldi A, Österle H, Portmann E, Spiekermann S (2021) Life engineering. Bus Inf Syst Eng 63:191–205. https://doi.org/10.1007/s12599-020-00680-x

Apprich C, Cramer F, Hui Kyong Chun W, Steyerl H (2018) Pattern discrimination. meson, https://doi.org/10.14619/1457

Belohlavek R, Dauben JW, Klir GJ (2017) Fuzzy logic and mathematics: a historical perspective. Oxford University Press, Oxford

Bongiorno C, Zhou Y, Kryven M et al (2021) Vector-based pedestrian navigation in cities. Nat Comput Sci 1:678–685. https://doi.org/10.1038/s43588-021-00130-y

Crawford K (2021) Atlas of AI. Yale University Press,

Easwaran E (2008) Die Upanischaden. Eingeleitet und übersetzt von Eknath Easwaran. Goldmann

Europäische Kommission (2019) Ethik-leitlinien für eine vertrauenswürdige KI. https://data.europa.eu/doi/10.2759/856513. Zugegriffen: 6. Febr. 2022

Floridi L (2018) Soft ethics and the governance of the digital. Philos Technol 31:1–8. https://doi.org/10.1007/s13347-018-0303-9
Gendlin ET (2018) Focusing. Selbsthilfe bei der Lösung persönlicher Probleme. rororo
Goodchild MF (2007) Citizens as sensors: the world of volunteered geography. GeoJournal 69:211–221. https://doi.org/10.1007/s10708-007-9111-y
Google (o.J.) Artificial intelligence at Google: our principles. https://ai.google/principles/. Zugegriffen: 13. Febr. 2022
Hallensleben S (2021) Operationalisierung von KI-Ethik gelingt durch Normung. IT-Governance 34:17–19
Hamilton IA (2018) Amazon built an AI tool to hire people had to shut it down because it was discriminating against women. https://www.businessinsider.com/amazon-built-ai-to-hire-people-discriminated-against-women-2018-10?r=US&IR=T. Zugegriffen: 6. Febr. 2022
Hudec M, Vucetic M, Cermakova I (2020) The synergy of linguistic summaries, fuzzy functional dependencies and land coverings for augmenting informativeness in smart cities. In: 2020 28th Telecommunications Forum (TELFOR). IEEE, S 1–4
IBM (2019) Everyday ethics for artificial intelligence. https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf. Zugegriffen: 13. Febr. 2022
Ito J (2020) Resisting reduction: designing our complex future with machines. MIT Press
Jobin A, Ienca M, Vayena E (2019) Artificial Intelligence: the global landscape of ethics guidelines. Nat Mach Intell. https://doi.org/10.1038/s42256-019-0088-2
Mendel J, Wu D (2010) Perceptual computing: aiding people in making subjective judgments. Wiley
Microsoft (2018) Responsible bots: 10 guidelines for developers of conversational AI. https://www.microsoft.com/en-us/research/uploads/prod/2018/11/Bot_Guidelines_Nov_2018.pdf. Zugegriffen: 13. Febr. 2022
Moor JH (1995) Is ethics computable? Metaphilosophy 26(1/2):1–21
Nielsen J (1989) Usability engineering at a discount. In: Proceedings of the Third International Conference on Human-Computer Interaction 1989, S 394–401
Österle H (2020) Life Engineering – Mehr Lebensqualität dank maschineller Intelligenz? Springer, Wiesbaden https://doi.org/10.1007/978-3-658-28335-3
Österle H, Portmann E, D'Onofrio S (2021) Human life engineering. Informatik Spektrum. https://doi.org/10.1007/s00287-021-01377-5
Portmann E (2019) Fuzzy Humanist. Springer essentials
Samek W, Müller KR (2019) Towards explainable artificial intelligence. In: Samek W, Montavon G, Vedaldi A, Hansen L, Müller KR (Hrsg) Explainable AI: interpreting, explaining and visualizing deep learning. Lecture notes in computer science, Bd. 11700. Springer, Cham https://doi.org/10.1007/978-3-030-28954-6_1
Segun ST (2020) From machine ethics to computational ethics. AI Soc. https://doi.org/10.1007/s00146-020-01010-1
Spiekermann S (2019) Digitale Ethik: Ein Wertesystem für das 21. Jahrhundert. Droemer
Zadeh LA (1965) Fuzzy sets. Inf Control 8(3):338–353
Zadeh LA (2012) Computing with words: principal concepts and ideas. Springer, Berlin
Čerka P, Jurgita G, Gintarė S (2015) Liability for damages caused by artificial intelligence. Comput Law Secur Rev 31(3):376–389