ETHIK DER ROBOTIK

Thomas Rusche

Geschäftsführender Gesellschafter SØR Rusche GmbH, Oelde

Privatdozent Universität Siegen, WHU Vallendar, Philosophische Hochschule München

Email: t.rusche@soer.de

INHALT

1. Hinführung zur Roboterethik

Die Robotik gehört mit der Nuklear- und Biotechnologie sowie dem katastrophalen Klimawandel zu den großen Herausforderungen der Menschheit. Während die planetarische Dimension von Atomenergie, Biotechnologie und Klimawandel in Gesellschaft und Wissenschaft breit und intensiv diskutiert wird, steckt die die ethische Reflexion der Robotik noch in den Kinderschuhen. Oftmals werden technologische Fortschrittsperspektiven aus ökonomischer und politischer Sicht unkritisch begrüßt und eingefordert, ohne jedoch die ethische Dimension der Robotik in den Blick zu nehmen.

1.1 Historische Hinführung: Der Roboter als effizientes Rationalisierungsinstrument

Im Todesjahr Max Webers (1864-1920) prägt der tschechische Dramatiker Karel Čapek (1890-1938) den Begriff ‚Roboter' (robota, tschechisch: Plagerei; slawisch: Arbeit). In seinem Drama R. U. R. (Rossum's Universal Robots) erschaffen Vater und Sohn Rossum (slawisch: Vernunft) einen Roboter, der in Serie geht.

Während Vater Rossum erfolglos versucht, einen Menschen in seiner gesamten Komplexität technisch zu reproduzieren, gelingt es dem Sohn, eine Maschine zu erbauen, die auf alles verzichtet, was nicht der Produktivität dient. Menschen sehnen sich nach Glück, spielen Piano und gehen spazieren – „but a working machine must not play the piano, must not feel happy (...). Young Rossum invented a worker with the minimum amount of requirements. He had to simplify him. He rejected everything that did not contribute directly to the progress of work - - everything that makes man more expensive. In fact he rejected man and made the Robot (...). Mechanically they are more perfect than we are, they have an enormously developed intelligence, but they have no soul".[1] Rossum jun. konzipiert seinen Roboter vor der rationalisierungstheoretischen Hintergrundfolie Max Webers. Effizienz, Kalkulierbarkeit, Vorhersehbarkeit und Naturbeherrschung treiben die westliche Zivilisation zu immer größerer Zweckrationalität. In einer erfolgsrationalen Gesellschaft – so führt Maria Bakardjieva (*1959) aus – werden „human faculties that are not essential for survival and domination of nature (playing the piano, going for walk ...) devalued and marginalized in order to bring into prominence qualities that directly serve the rational goals of efficiency and mastery".[2] Indem Rossum jun. von der Fülle des Menschseins abstrahiert und den Menschen auf sein

[1] Čapek, Robots, S. 9.
[2] Bakardjieva, Rationalizing, S. 246.

technisches Machen-Können reduziert, vermag er es, einen Roboter als höchst effizientes technisches Hilfsmittel zur rationalen Zielerreichung zu fertigen.

1.2 Begriffliche Eingrenzung: Roboter und Robotik

Weder ist der Begriff Roboter bzgl. seiner Merkmale und Eigenschaften intensional eindeutig definiert, noch besteht eine transdisziplinäre Übereinkunft bzgl. des extensionalen Begriffsumfangs. Diese definitorische Unschärfe ist für eine junge Wissenschaft typisch. Die wachsende Entwicklungsdynamik der Robotertechnologie sowie die Vielzahl involvierter Disziplinen wie Mechanik, Elektronik, Informatik bzw. Mechatronik, Intelligenzforschung, Psychologie, Soziologie und zunehmende Anwendungsbereiche, wie z. B. Industrie, Medizin, Landwirtschaft, Haushalt, Logistik und Militärwirtschaft, erschweren die Ausbildung einer einheitlichen Nomenklatur. [3]

Weitgehende Zustimmung dürfte folgender Definitionsvorschlag erfahren: Roboter sind digital codierte Systeme, die mehr oder weniger räumlich abgegrenzt selbstständig definierte Aufgaben ausführen. Wenn das Kriterium einer räumlich physischen Abgrenzung keine definitorische Berücksichtigung findet, so werden auch Softwarebots (e. g. social bots) zu den Robotern gerechnet. Von Software gesteuerte Hardwareroboter können mit Maschinenoptik (z. B. Industrieroboter oder Feldroboter), zoomorph oder anthropomorph, d. h. als Hermanoide mit menschenähnlichen Gliedmaßen, Bewegungsabläufen, mit Haut, Mimik, Gestik und Sprachfertigkeiten ausgestaltet werden. Hardwareroboter sind stationär bzw. dreiachsig beweglich oder vollständig mobil. Mobile Roboter gibt es für den Einsatz zu Land oder Luft (Drohnen), im All (Spacerroboter) bzw. unter Wasser. Nanoroboter (Nanobots, Naniden, gr. Nanon – Zwerg) bestehen aus kleinsten Materialstrukturen und sind ein weiterer Beleg für die rasante technologische Entwicklung.

Die Wissenschaft der Robotik wird von manchen bereits als Leitdisziplin des 21. Jahrhunderts bezeichnet. Sie befasst sich mit der technischen Entwicklung von Robotern, deren Entwurf und Gestaltung, Programmierung, Steuerung, Produktion, Betrieb und Anwendungsfeldern. Die Industrie ist mit dem Militärwesen das älteste Einsatzgebiet der Robotik und verfügt auch heute noch über die größte Anzahl und Vielfalt von Robotern. Die VDI-Richtlinie 2860 charakterisiert Industrieroboter als „universell einsetzbare Bewegungsautomaten mit mehreren

[3] Bereits 2005 wurde im Rahmen des Research Atelier on Roboethics in Rom als Desiderat formuliert: „develop a common language among scholars and stakeholders on Roboethics“ (Veruggio, Roboethics, S. 4).

Achsen, deren Bewegungen hinsichtlich der Bewegungsfolgen und Wegen bzw. Winkeln frei (...) programmierbar und ggf. sensorgeführt sind“ (VDI-Richtlinie 2860).[4] Zunehmend werden Industrieroboter auch zur Steuerung und Inspektion eingesetzt und überprüfen die Qualität der Erzeugnisse.

Seit geraumer Zeit verlassen Roboter die Fabrikhallen und finden Einsatz in der Landwirtschaft (Feldroboter). Soziale Roboter begegnen Menschen in Freizeit und Haushalt, z. B. als sich selbst steuernder Staubsauber oder Schwimmbadreiniger. Zu Diagnose-, Therapie-, OP- und Pflegezwecken werden Roboter in Arztpraxen, Krankenhäusern und Altersheimen eingesetzt. In der Justiz dienen sie der Rechtsberatung und Vertragsprüfung. Roboter unterrichten in Schulen und bedienen als Serviceroboter Gäste in (japanischen) Hotels. Roboter retten bei Erdbeben und Schneelawinen verschüttete Menschen. Autofahrer erleben sprachfähige Navigations- und Assistenzsysteme und werden auf eine kommende Generation selbstfahrender Roboterautos vorbereitet. Unterhaltungsroboter spielen mit Kindern und bieten als Sexpartner für jegliche Neigung und Orientierung ihre Dienste an. Drohnen finden als Transportmittel in der Logistikindustrie für privat- wirtschaftliche und militärische Zwecke zunehmende Verwendung, wobei Kampfroboter zu tödlichen Waffen werden.[5]

Angesichts der weitreichenden technologischen Entwicklungsperspektiven wird bereits im Februar 2004 auf der internationalen Robotermesse in Fukuoka, Japan, die *World Robot Declaration* verabschiedet:

„1. Next-generation robots will be partners that coexist with human beings;
2. Next-generation robots will assist human beings both physically and psychologically;
3. Next-generation robots will contribute to the realization of a safe and peaceful society”.[6]

Aus der erfolgsorientierten Perspektive der Robotikindustrie formuliert, mag diese positive Technikfolgenabschätzung naheliegen. Angesichts von Technikeuphorie und Technikrisiken postulieren geisteswissenschaftliche Forscher allerdings die Notwendigkeit einer *‚Roadmap for Roboethics‘* (2005): „The public is already asking questions such as: ‚Could a robot do *good*

[4] Der detaillierte Definitionsgrad von Industrieroboter ist für diese relativ alte Teildisziplin der Robotik kennzeichnend. Das US-amerikanische *Robot Institut* definiert den Industrieroboter als „ein programmierbares Mehrzweck-Handhabungsgerät für das Bewegen von Material, Werkstücken, Werkzeugen oder Spezialgeräten“ (Schnetter, Robotik, S. 20).
[5] Tzafestas, Roboethics, S. 46-62.
[6] Veruggio, Roboethics, S. 3.

and *evil*?', 'Could robots be dangerous for human kind?'"[7] Wie könnte eine solche Roboterethik inhaltlich und konzeptionell ausgestaltet werden?

2. Fragestellungen der Roboterethik

Die oben kurz skizzierte Wirkungsweise von Robotern in unterschiedlichen Anwendungsfeldern hat bereits deren Relevanz für das Leben und Handeln des Menschen in Haushalt, Wirtschaft und Gesellschaft verdeutlicht.

2.1 Zweckrationalität und Ethik

Roboter verändern zunehmend die Gewohnheiten, Sitten und Gebräuche (Ethos) des Menschen. Diese hochtechnologisch getriebene Veränderungsdynamik erfasst die Mikroebene der Individuen, die Mesoebene von Unternehmen und Organisationen sowie die Makroebe der (Welt-)Gesellschaft. Für erfolgsrationale Zwecke werden immer leistungsfähigere Roboter entwickelt. Das Primat der zweckrationalen digitalen Effizienzsteigerung diktiert nicht nur der Ökonomie, sondern zunehmend auch der Politik die Agenda. Aus diesem Sachverhalt der ersten, fachwissenschaftlichen Untersuchungsstufe, dass Robotik (R) die Praxis (P) des Menschen verändert (R - > P), ergibt sich aus philosophischer Sicht die Frage, wie wir als Vernunftwesen auf welche Weise darauf Bezug nehmen. Die Klärung eines solchen Sachverhalts zweiter Stufe kennzeichnet mit Holm Tetens (* 1948) die „Philosophie als Disziplin höherer Ordnung (.. und findet ihren Ausdruck in der, Einschub v. Verf.) Wendung von den Sachverhalten zur Bezugnahme auf Sachverhalte".[8]

Es bedarf einer dritten Untersuchungsstufe, um zu klären, welche Weise für vernünftige Personen angemessen ist, um auf den Sachverhalt (R - > P) Bezug zu nehmen. Vernünftige Personen können Sachverhalte „unter höchst verschiedenen letzten Gesichtspunkten und Zielrichtungen >>rationalisieren<<".[9] Neben der Zweckrationalität, die sich aus der „Geeignetheit der Mittel bei gegebenem Zwecke"[10] (Mittelrationalität) und der angemessenen Wahl der Zwecke angesichts gegebener Mittel und Optimierungsbedingungen (Zielrationalität) ergibt, möchte ich in Weiterführung von Max Weber mit Jürgen Habermas (*1929) Wertrationalität von ethischer Rationalität unterscheiden. Wertrationalität systematisiert die

[7] Veruggio, Roadmap, S. 614.
[8] Tetens, Argumentieren, S. 18 f.
[9] Weber, Kapitalismus, S. 20.
[10] Weber, Objektivität, S. 25.

konkreten materiellen Wertinhalte, die als substantielle Sittlichkeit den Moralhaushalt des Menschen bilden. Bei gegebenen Mitteln und Zielen führen unterschiedliche, oftmals kulturell geprägte Wertpräferenzen zu abweichenden Bewertungen von Sachverhalten wie R - > P. Beispielsweise ist die abendländische (Sciencefiction-)Literatur vom Topos geprägt, dass sich Roboter in einer Rebellion gegen die Menschheit wenden.[11] Die japanische Kultur ist hingegen wesentlich von der Wertvorstellung geprägt, dass intelligente Maschinen dem Menschen nützlich, weil dienstbar und als Erschaffer dankbar sind. Diesen kulturell geprägten moralischen Wertehaushalt, der z. B. mehr oder weniger technikfreundlich oder kapitalismuskritisch bestückt sein kann, kritisch zu hinterfragen ist Aufgabe der ethischen Rationalität: „Denken Philosophen über die Moral nach, nennt man das >> *Ethik* <<. [12]

Robotik intendiert eine Höchstform von Mitteleffizienz. Geschwindigkeit und Präzision der Roboter übertreffen in immer mehr Anwendungsbereichen die altbekannten analogen Instrumente und Verfahren. Welchen Zielen aber dienen diese hocheffizienten Mittel? Müssen wir mit Albert Einstein (1879-1955) zugestehen, „daß wir zwar über perfekte Mittel verfügen, aber doch gleichzeitig nur auf verworrene Ziele zurückgreifen können“?[13] Werden Zielsetzungen, wie ökonomische Effizienz oder militärische Durchschlagskraft von der Robotik ethisch unhinterfragt akzeptiert? Welche Wertvorstellungen fließen z. B. bei der Entwicklung und Gestaltung von menschenähnlichen Pflege- oder Sexrobotern in den Zweckmittelentscheid ein? Wie können individuelle Wertvorstellungen für andere nachvollziehbar begründet werden? Eben diese Reflexion, Kritik und Begründung von Normen und Werten in der Robotik ist die grundlegende Aufgabe und Fragestellung einer Roboterethik (Roboethik). Konkret:[14]

- Wie können roboterethische (Grund-)Normen ethisch begründet werden?

- Wie kann die Anwendung ethisch legitimierter Normen auf den Feldern der Robotik gelingen?

- Welche Zielsetzungen und Wertvorstellungen prägen Menschen, die Roboter entwerfen, produzieren und einsetzen?

[11] Bereits bei Čapek erheben sich die übermächtigen Roboter gegen den Menschen bzw. die Menschheit – ein Thema, das zuletzt Frank Schätzing in seinem neuen Roman „Die Tyrannei des Schmetterlings“ aufgegriffen hat.
[12] Tetens, Argumentieren, S. 139.
[13] Zwierlein, Wirtschaftsethik, S. 79.
[14] Vgl. Tzafestas, Roboethics, S. 65.

- Welche Rolle sollen Roboter zukünftig in den unterschiedlichen Anwendungsfeldern übernehmen?

- Gibt es spezifische Roboter (Applikationen), die aus welchen Gründen nicht erschaffen werden sollten?

- Können Roboter hinsichtlich einer normativen Orientierung programmiert werden?

- Zu welchen Handlungskonsequenzen kann eine normativ-ethische Programmierung von Robotern führen?

- Welche Moralvorstellungen leiten Menschen im Umgang mit Robotern?

Bevor diese grundsätzlichen Fragestellungen einer Roboterethik hinsichtlich unterschiedlicher Robotertypen und Anwendungsfeldern konkretisiert werden, möchte ich kurz auf ein Abschätzungsproblem der rasanten technologischen Veränderungsdynamik der Robotik eingehen.

2.2 Ethische Herausforderung der künstlichen Intelligenz

Wesentlicher Entwicklungstreiber der Robotik ist die Kombination von Künstlicher Intelligenz[15] und exponentiell steigender Rechenleistung. Beispielsweise verfügt unser Smartphone heute über eine größere Rechenkapazität als alle NASA-Computer, die zusammen 1969 die erste Mondlandung ermöglichten.

Die Forschung zur Künstlichen Intelligenz (KI) untersucht kognitive menschliche Fähigkeiten, wie Lernen und Problemlösen, um intelligente künstliche Agenten zu entwickeln. Im Jahre 2016 gewann ein KI-Agent (AlphaGo) das schwierigste Brettspiel der Welt >Go< mit 4:1 gegen Lee Sedol, den weltbesten Spieler. Zur Vorbereitung hat AlphaGo in einem menschlich gesteuerten Lernprozess mehrere tausend Go-Spiele nachvollzogen, um die Regeln und Strategien besser zu verstehen. Für ein derartiges künstliches Expertensystem „muss das Wissen des Experten in

[15] Der Begriff ‚Artificial Intelligence' wurde 1955 zur Vorbereitung der Dartmouth Conference (1956) vom Kyoto-Preisträger (1988) John McCarthy (1927-2011) geprägt.

Regeln gefasst werden, in eine Programmiersprache übersetzt und mit einer Problemlösungsstrategie bearbeitet werden“.[16]

Im Oktober 2017 publizierte *Nature* einen Artikel über AlphaGo Zero.[17] Ohne mit Datensets bisheriger Spielstrategien gefüttert zu werden, kannte AlphaGo Zero nur die Regeln und wurde mit einer Belohnungsfunktion programmiert. In einem von Menschen nicht gesteuerten Lernprozess entdeckte der künstliche Agent völlig unbekannte neue Strategiezüge und gewann nach drei Tagen gegen seine Vorgänger AlphaGo mit 100:0. Leistungsstärkere Algorithmen haben den KI-Agenten in die Lage versetzt, mehr Go-Wissen anzuhäufen als menschliche Go-Spieler in tausenden von Jahren. Die neue Generation intelligenter künstlicher *deep learning* Agenten besteht aus einer lernfähigen Software, die sich eigenständig und mit rasanter Geschwindigkeit fortschreibt. Mit jedem Update ist die Entwicklung für die Programmierer schwerer nachzuvollziehen.[18] „Deep learning agents (..) are not programmed to do a specific task in a set way. Instead they learn to perform a task by attempting it millions of times until they evolve a successful strategy – sometimes one that its human creators had not anticipated and do not understand.” [19] So waren die erfahrenen Go-Programmierer überrascht, welch ‘unmenschliche’ Spielzüge und Strategien, die ihnen bisher völlig unbekannt waren, AlphaGo Zero zum Sieg geführt haben.

Mittels leistungsstarker Algorithmen und dem Zugriff auf Big Data vermag der KI-Agent ein spezifisches Problem besser zu lösen als der Mensch: Verkehrsschilder werden ebenso treffsicherer erkannt, wie gutartige von bösartigen Tumoren unterschieden. Auch wenn diese Fortschritte für die Entwicklung der KI bedeutungsvoll sind, so beschränkt sich doch deren Problemlösungsvermögen auf ein eng gestecktes spezifisches Anwendungsfeld (ANI: Artificial Narrow Intelligence), wie z. B. Go oder Schachspielen.

Zur Überwindung dieser Bereichsgrenzen, unternimmt Deep Mind, die Entwicklungsfirma von AlphaGo, größte Anstrengungen, um einen Masteralgorithmus zu erschaffen, der jedes komplexe Problem quer durch alle Wissens- und Anwendungsbereiche mittels eigenständiger Lernerfahrung zu lösen vermag (AGI: Artificial General Intelligence). Während das Ziel der Entwicklung einer reichhaltigen generellen künstlichen Intelligenz für zahlreiche Mitglieder der scientific community erreichbar erscheint, ist es sehr umstritten, ob auch die Erschaffung einer

[16] Mainzer, Intelligenz, S. 43.
[17] Silver/Schnittwieser/Simonyan, Game, S. 354-359.
[18] Vgl. Weizenbaum, Computer, S. 308 f.
[19] Devlin, Ethics, S. 2.

machtvollen künstlichen Superintelligenz gelingen kann (ASI: Artificial Superintelligence). Ein solcher KI-Agent hätte „an intellect that is much smarter than the best human brains in practically every field including scientific creativity, general wisdom and social skills“.[20]

Die bloße Möglichkeit von ASI führt zu der Singularitätshypothese, die nicht nur in der Sciencefiction-Literatur, sondern laut Stanislaw Ulam (1909-1984) erstmals von John von Neumann (1903-1957) vorgebracht wurde: „die stete Beschleunigung des technischen Fortschritts und der Veränderungen im Lebenswandel (.. haben den Anschein ...) auf eine entscheidende Singularität in der Geschichte der Menschheit hinauszulaufen, nach der die Lebensverhältnisse, so wie wir sie kennen, sich nicht fortsetzen können“.[21]

Der Mathematiker Irving John Good (1916-2009) prägt die heutige Begriffsbedeutung von technologischer Singularität: Wenn der Mensch eine ultraintelligente Maschine baut, die die Fähigkeiten des Menschen bei weitem übertrifft, so wird dies die letzte Erfindung des Menschen sein. Warum? Weil dann ultraintelligente Maschinen noch bessere Maschinen bauen, die zu einer explosionsartigen Entwicklung der Intelligenz führen und den Menschen hinter sich zurück lassen. Die künstlichen Superagenten übertreffen den Menschen und übernehmen die Gestaltung der Welt.[22]

In seiner noch nicht veröffentlichten und der philosophischen Fakultät der Universität Köln vorgelegten Habilitationsschrift *Can we create strong & safe AI?* postuliert Karl Johannes Lierfeld: “Whatever path to artificial superintelligence proofs to be successful – this very pathway will guarantee the party who has chosen it a decisive strategic advantage. The possessors of the world´s first strong AI will be automatically in the ultimative leadership position. It is mandatory that given party has installed sound control mechanisms and represents a moral code that is in concordance with basic humanistic values.
These ethical implications of the control problem refer not to technology but to our very own nature. Only if we can control the egocentric, non-altruistic and sociopathic layers of our nature we can effectively control artificial intelligence”.[23]

[20] Bostrom, Superintelligence, S. 11-30.
[21] John von Neumann (1903-1957) has already “centered on the accelerating process of technology and changes in the mode of human live, which gives the appearance of approaching some essential singularity in the history of the race beyond which human affairs, as we know them, could not continue” (Stanislaw Ulam, John von Neumann, S. 5).
[22] Vgl. Good, Machine, S. 31-88.
[23] Lierfeld, AI, S. III.

3. Roboterethische Problemlösungsansätze

Theaterstücke und Sciencefictionliteratur haben nicht nur die Robotertechnologie beeinflusst, sondern bilden auch den historischen Ursprung der Roboterethik. Zwar wurde der Begriff Roboterethik (Roboethics) erst 2004 von Gianmarco Veruggio geprägt [24]; 1942 hatte jedoch bereits der Biochemiker Isaac Asimov (1920-1992) in seinem Sciencefiction *Runaround* (Herumtreiber) drei ethische Robotergesetze formuliert, die er später (1983) um ein nulltes Gesetz ergänzt hat, das über die Beziehung des Roboters zum Menschen hinaus die Verantwortung für die ganze Menschheit betont:

0. Ein Robot darf die Menschheit nicht verletzten oder durch Nichtstun zulassen, daß die Menschheit zu Schaden kommt.[25]

1. „Ein Robot darf kein menschliches Wesen verletzten oder Untätigkeit gestatten, daß einem menschlichen Wesen Schaden zugefügt wird.“[26]

2. „Ein Robot muß dem ihm von einem Menschen gegebenen Befehl gehorchen, es sei denn, ein solcher Befehl würde mit Regel Eins kollidieren“.[27]

3. „Ein Robot muß seine eigene Existenz beschützen, solange dieser Schutz nicht mit Gesetz Eins oder Zwei kollidiert“.[28]

Asimov betont die unbedingte Pflicht der Robotik, den Menschen und die Menschheit zu schützen. Jegliches Handeln von Robotern unterliegt dieser deontologischen Regel. Auch das Existenzrecht der Roboter ist an die in lexikalischer Ordnung vorausgesetzte Wahrung der Humanität geknüpft. Wie kann nun sichergestellt werden, dass Roboter Menschen keinen Schaden zufügen?

Die Asimovschen Gesetze werfen u. a. die Frage nach dem moralischen Status von Robotern auf. Können Roboter (ggf. in absehbarer Zeit) selbstveranlasst, d. h. autonom aus selbst gesetzten Gründen handeln? Sind Roboter genuine moralische Akteure, die auf Grund

[24] 1st International Symposium on Roboethics, Jan. 2004 (www.roboethics.org).
[25] Vgl. Asimov, Aufbruch.
[26] Asimov, Robotergeschichten, S. 286.
[27] Ebenda.
[28] Ebenda.

systeminterner Faktoren selbstkontrolliert entscheiden können? Die Frage nach der Autonomie und damit nach dem moralischen Status von Robotern ist für die Wahl ethischer Problemlösungsansätze von entscheidender Relevanz.

3.1 Ethische Programmierung von (super-) intelligenten künstlichen Agenten

Die Perspektive, vollständig autonome, superintelligente Roboter zu entwickeln, die dem Menschen in jeder Hinsicht überlegen sind, gehört für viele Forscher auch heute noch in den Bereich von Sciencefiction. Zwar mag eine intelligente Maschine den Turing-Test[29] bestehen und bei einer intensiven Befragung von einem denkenden Menschen nicht zu unterscheiden sein, die intelligente Maschine mag auch regelgerecht chinesisch sprechen, ohne jedoch verstehen zu können, dass sie chinesisch spricht.[30] Noch können künstliche Agenten keinen reflexiven Bezug auf ihr Tun nehmen. Es fehlt ihnen diese philosophische Fähigkeit des Menschen, die ihn als Vernunftwesen charakterisiert. Deshalb streben Informatiker nach einer engen Kooperation mit Gehirnforschern, um die fortschreitende Kenntnis neuronaler Strukturen für die Weiterentwicklung der KI zu nutzen.

So werden in den interdisziplinären Forschungslabors der GAFA (Google, Apple, Facebook, Amazon)- und BAT (Baidu, Alibaba, Tencent)-Industrie Milliardenbudgets in die Entwicklung von Masteralgorithmen investiert, die jegliches Problem eines beliebigen Anwendungsfeldes schneller und präziser lösen sollen als der Mensch es vermag. Dass es bei diesen Forschungsanstrengungen zum oben beschriebenen Kippmoment der Singularität kommen könnte, kann zumindest nicht ausgeschlossen werden.

Die hochtechnologische Entwicklungsdynamik hat bereits im letzten Jahrhundert zu einer neuartigen Gefährdung des Menschen geführt, der nunmehr über eine absolute Zerstörungsgewalt verfügt, mit der er seine eigene Gattung vernichten kann. Auch die möglichen Fernfolgen der KI-Forschung betreffen nicht nur einzelne Menschen, Forscher und Unternehmen, sondern die ganze Menschheit. Diese ungeheure Verfügungsmacht ist eine Herausforderung für den ethischen Verantwortungsbegriff, der den mikroethischen Nahbereich von individuellen Face-to-face-Beziehungen grundsätzlich übersteigt. Angesichts dieser planetarischen makroethischen Dimension formuliert Hans Jonas (1903-1993) mit seinem

[29] Alan Turing entwickelte 1950 einen Test, bei dem der Forscher zwei unbekannten Probanden über ein Terminal Fragen stellt, ohne diese zu sehen oder zu hören. Wenn der menschliche Fragensteller nicht unterscheiden kann, welcher Antwortgeber eine Maschine oder ein Mensch ist, hat die Maschine den Turingtest bestanden.
[30] Vgl. Das Chinese-Room-Argument von John Searle; vgl. Burkholder, Searle, S. 336.

Imperativ eine allgemeine Pflicht zur Zukunftsverantwortung: „Handle so, daß die Wirkungen Deiner Handlungen verträglich sind mit der Permanenz echten menschlichen Lebens auf Erden.“[31] Diese unbedingte Verpflichtung, beständig nach der Permanenz echten menschlichen Lebens zu streben, beinhaltet selbst den Menschenwürdegrundsatz, kann doch *echtes menschliches Leben* expliziert werden als ein Leben mit Anspruch auf Achtung der Menschenwürde und im Sinne der Menschenwürde. Dementsprechend kann der Jonas’sche Imperativ als ein auf der Zeitachse in die Zukunft führender Grundsatz verstanden werden, das Gebot der Menschenwürde immer und überall zu beachten.

Jonas schlägt mehrere Formulierungsvarianten seines Imperativs der Zukunftsverantwortung vor, u. a.: „Gefährde nicht die Bedingungen für den infiniten Fortbestand der Menschheit auf Erden“[32] und entwickelt zur Risikovermeidung eine *Heuristik der Furcht*. Aufgrund der unsicheren Technikfolgenabschätzung ist entsprechend dieser heuristischen Regel „im Zweifelsfall Vorrang für die Unheilsprognose“[33] zu geben. Angesichts der überbordenden Dynamik und Unsicherheiten in komplexen Systemen wie der KI-Forschung würde die Heuristik der Furcht allerdings zu einem „Verlust von Innovations- und Reaktionsfähigkeit führen und könnte sich am Ende als eine Strategie erweisen, die aufgrund der Lähmung von Handlungspotentialen mehr Risiken erzeugt als vermeidet.“ [34] Um die Resourcen zur Chancennutzung und Problembewältigung komplexer Technologien wie der Robotik verantwortungsvoll einzusetzen, bedarf es der *Risikomündigkeit*. „Risikomündigkeit ist die Fähigkeit, unter mehrfacher Unsicherheit – hinsichtlich der Handlungsfolgen, hinsichtlich unterschiedlicher ethischer Maßstäbe der Betroffenen, über die nur ein unvollständiger Konsens gefunden wird, sowie hinsichtlich der eigenen moralischen Rationalität, für die es unter Modernitätsbedingungen keine Letztbegründung und keine vollständige Kohärenz zu geben scheint – begründete Entscheidungen zu treffen.“[35]

Auch wenn bezüglich der Entwicklungsperspektiven superintelligenter Roboter in der Scientific Community und darüber hinaus kein vollständiger Konsens erzielt werden kann, so ist es diskurspragmatisch geboten, danach zu streben. Warum? Weil in der rationalen Argumentation (Diskurs) alle thematisch relevanten Gründe und Gegengründe erörtert werden, um ein Problem

[31] Jonas, Verantwortung, S. 36.
[32] Jonas, Verantwortung, S. 36.
[33] Jonas, Verantwortung, S. 63 f.
[34] Vogt, Nachhaltigkeit, S. 370.
[35] Vogt, Nachhaltigkeit, S. 370.

zu erörtern und ggf. zu lösen. Argumentieren bedeutet dieses gegenseitige Infragestellen von Behauptungen und Anerkennen von rational überzeugenden Begründungen.[36]

In einem solchen Diskurs würde z. B. der Kölner Philosoph Karl Johannes Lierfeld, den ich hier advokatorisch vertreten möchte, gemeinsam mit Nick Bostrom, Dimitry Itskov, Ray Kurzweil, Hans Moravec, Elon Musk und Eliezer Yudkowsky u. a. gewichtige Argumente vorbringen, die es angesichts unterschiedlicher Technikszenarien nahelegen, dass superintelligente Roboter in den nächsten Jahrzehnten entwickelt werden und es eine zukunftsverantwortliche Aufgabe ist, bereits jetzt für diese potentiellen Superagenten eine Roboterethik zu konzipieren. Auch wenn superintelligente Roboter nicht entwickelt würden, so besteht die Aufgabenstellung einer ethischen Programmierung von Robotern bereits heute, z. B. für selbstfahrende Autos, die in Dilemmasituationen geraten.[37]

Ohne kodierte digitale Informationsübertragung wären künstliche intelligente Agenten nicht zu erschaffen. Informationen müssen in einem binären Code (lat. bini – zwei) übertragen werden. Auch eine ethische Programmierung von Robotern ist nur mittels eines binären Codes möglich. Eine solche binär codierte Programmierung würde voraussetzen, dass ethische Prinzipien trotz ihrer Allgemeinheit und der Fallibilität von Situationseinschätzungen in eindeutige (binäre) moralische Handlungsanweisungen zu überführen sind. Wie können Grundnormen der Menschenwürde oder der *Permanenz echten menschlichen Lebens auf Erden* so eindeutig programmiert werden, dass ambivalente Interpretationsspielräume ausgeschlossen werden? Lierfeld zieht eine erste, recht ernüchternde Zwischenbilanz: „The translation of mankind's philanthropic canon into code marks a huge challenge and inherits hazards in its own right. It seems that the philosophical tradition of creating increasingly higher levels of abstraction might have led into the wrong direction in this context. (...) mostly failing at defining definite answers to our most fundamental questions, we seem to be unable to reduce complexity in such a large scale."[38]

Ethische Programmierung setzt klar definierte Regeln, feststehende Zielsetzungen und Ausschlusskriterien voraus. Um den moralischen Gehalt derartiger Ziele und die Ausgestaltung von Verfahren streitet die Ethik seit 2500 Jahren. Im Wissen um die Diskrepanz zwischen dem, was ethische Roboterprogrammierung voraussetzt und eine philosophische Ethik zu leisten

[36] Vgl. Kuhlmann, Letztbegründung, S. 184.
[37] Vgl. Trolley-Problem.
[38] Lierfeld, AI, S. 231.

vermag, resumiert Lierfeld: „If mankind won't define those issues, a superintelligent AI will do it for us."[39]

3.2 Moralbasierte Steuerung von (teil-) autonomen Robotern

Für kommende Generationen selbstfahrender Automobile ist eine moralbasierte Steuerung ebenso relevant, wie für die Entwicklung von unbemannten Kampfrobotern. Allerdings ermöglicht es der aktuelle Entwicklungsstand autonomer Waffen (AWS ‚Autonomous Weapons Systems') trotz großer Fortschritte beim Einsatz unbemannter Kampfeinsätze noch nicht, dass diese ohne menschliche Steuerung Ziele auswählen und tödlich attackieren (LAWS: Lethal Autonomous Weapons Systems).[40]

Bereits heute ersetzen teilautonome Roboter den Menschen zunehmend bei Kampfeinsätzen; damit besteht die Gefahr, dass die Einsatzschwelle abgesenkt wird, da Soldaten der angreifenden Partei weniger gefährdet sind. Ganz anders stellt sich die Gefährdungslage der angegriffenen Seite dar. Die sinkende Einsatzschwelle führt naheliegender Weise zu häufigeren Angriffen mit zerstörerischen Konsequenzen.

Vollautonome Waffensysteme könnten zukünftig ohne relevante menschliche Einflussnahme Ziele auswählen und Attacken ausführen. Roboter träfen dann die Entscheidung über Leben und Tod von Menschen. Aus zweckrationaler Effizienzperspektive mögen unbemannte Drohnen bereits heute Ziele eindeutiger identifizieren und adäquatere Mittel zu deren Eliminierung auswählen. Wenn das Ziel aber nun keine Chemiewaffenfabrik, sondern ein Mensch ist, so stellt sich die ethische Frage, ob dessen algorithmengesteuerter Tod unter welchen Voraussetzungen überhaupt verantwortbar ist.

Roboter sind binärgesteuerte Zweckrationalität. Solange Menschen über den Kampf- und Waffeneinsatz entscheiden, können wertrationale oder ethische Überlegungen Berücksichtigung finden. Wie aber, wenn der Kampfroboter mittels eines binären Codes (>1< oder >0<) über Leben und Tod entscheidet?

---

[39] Lierfeld, AI, S. 230.

[40] Frankreich, Großbritannien, Israel, Russland, Südkorea und die USA forschen und fördern die Entwicklung (teil-) autonomer Killerroboter und verweigern sich eines weltweiten Entwicklungs- und Verbreitungsverbotes bzw. eines Kontrollsystems, vergleichbar der ABC-Waffen. Ein solches preemptives Verbot von Killerroboter wird allerdings von zahlreichen Staaten und Organisationen gefordert.

Für die ethische Programmierung binär codierter digitaler Transaktionen liegt der Einsatz utilitaristischer Optimierungskalküle nahe.[41] Mit dem *größtmöglichen Nutzen* formuliert der Utilitarismus ein Kriterium, das sich an den Folgen einer Handlung orientiert. Zweckrationalität wird als Nutzenstiftung zum teleologischen Moralkriterium aufgewertet. Der ethische Konsequenzialismus konzentriert sich dabei auf den Effekt und unterstellt, dass Handlungsfolgen kohärent bewertet werden können, da Nutzenfunktionen quantitativ messbar und vergleichbar seien.

In kriegerischen Auseinandersetzungen werden Kampfroboter z. B. entsprechend der Zielsetzung programmiert, möglichst viele gegnerische Kombatanten und wenige unbeteiligte Zivilisten zu töten. Um den Erwartungswert zu ermitteln, bedürfte es einer Gewichtung der Alternativen nach Eintrittswahrscheinlichkeiten. Ob diese Optimierungskalkulation und Auswahl von Zielen und Mitteln zweckrational gelingen kann, ist allerdings fraglich, da das Optimierungskalkül komplexe Interdependenzrelationen von unterschiedlichen zur Verfügung stehenden (Kampf-)Mitteln und (Angriffs-)Zielen abbilden und evaluieren können müsste.

Aus moralphilosophischer Sicht ist das Vertrauen auf utilitaristische Optimierungskalküle grundsätzlich problematisch. Warum? Weil eine solche Kalkulation moralisch blind ist. Es ersetzt den Menschen als verantwortungsfähigen Entscheidungsträger und blendet das individuelle Ethos und Mitgefühl für unbeteiligte Zivilisten aus, die bei einem Angriff zu Tode kämen. Die Auswahl von Angriffszielen ist ein moralischer Delibrationsprozess, der zumeist dilemmatisch strukturiert ist. Dafür gibt es keine exakt zu kalkulierenden Optimierungslösungen, mit denen sich wertethische Fragen der individuellen moralischen Schuldübernahme erübrigen. Vielmehr steigt das Risiko kriegerischer Auseinandersetzungen, wenn Menschen sich *kein Gewissen mehr machen müssen* über den Tod anderer, weil optimierte Steuerungsprogramme der Kampfroboter über Einsatzdetails entscheiden und keine wertrationalen Überlegungen einfließen. Der Roboter kennt keine Scham, er hat keine Schuldgefühle. Roboter können weder leiden, noch sterben. Auch wenn sie technisch endlich sind, wissen Sie (noch) nicht um ihre Vergänglichkeit. Es fehlt ihnen die Fähigkeit zur Selbstreflexion, zum abwägenden Urteil sowie zum argumentativen Diskurs. Roboter sind eben keine leibhaftigen Diskurspartner, sondern Maschinen ohne Leibapriori.

[41] Vgl. Julian Nida-Rümelin, Humanismus, S. 143.

Darüber hinaus ist ein gravierendes ethisches Methodenproblem zu thematisieren. Die Fokussierung auf den effizienten Zweck-Mittel-Einsatz führt zu einer Ausblendung von ethischen Gerechtigkeitsfragen, da im utilitaristischen Kalkül z. B. der Tod unschuldiger Zivilisten und das Leid der Angehörigen verrechnet wird mit dem Nutzen getöteter Gegner und zerstörter gegnerischer Kampfmittel. Eine solche Verrechnung von Menschenleben verstößt gegen das unbedingte Prinzip der Würde des Menschen. Menschenleben sind als Selbstzweck von höchstem Wert und nicht verrechenbar. Diese Nichtverrechenbarkeit des Menschen ist das fundamentale Prinzip unserer humanen Zivilgesellschaft.

Auch bei Zulassungsverfahren selbstfahrender Autos wird derzeit geprüft, wie Fahrroboter für Dilemmasituationen moralisch programmiert werden sollen, ohne gegen das deontologische Prinzip der unverfügbaren Menschenwürde und der daraus folgenden Nichtverrechenbarkeit von Menschenleben zu verstoßen. In diesem Zusammenhang fragt die Online-Untersuchung ‚Moral Machine' (moralmachine.mit.edu) des MIT (Massachusetts Institute of Technology) an der seit 2016 nunmehr 40.000 User teilgenommen haben, nach den moralischen Intuitionen von Menschen in virtuellen Dilemmasituationen im Straßenverkehr.[42] Beispielsweise versagen bei einem SUV, der mit einer dreiköpfigen Familie besetzt ist, die Bremsen. Der Fahrer und Familienvater hat nur die Alternative, mit dem Auto bei höchster Geschwindigkeit gegen eine Mauer oder in eine Fußgängergruppe zu fahren. Entweder sterben die drei Familienmitglieder oder fünf Fußgänger. Intuitiv entscheidet sich die Mehrzahl der Befragten dafür, das eigene Leben und das der mitfahrenden Familienmitglieder zu schützen. Auch legen Marktstudien nahe, dass selbstfahrende Autos unverkäuflich wären, wenn diese in einer solchen Dilemmasituation nicht grundsätzlich die Rettung der Fahrzeuginsassen bevorzugten. Eine utilitaristische Nutzenkalkulation, die den Tod von drei Insassen angesichts der Alternative des Todes von fünf Passanten moralisch als vorzugswürdig empfiehlt, widerspricht demnach nicht nur dem deontologischen Nichtverrechenbarkeitsgebot, sondern auch der menschlichen Intuition.

Wie aber könnte ein moralisches Steuerungsprogramm für derartige Entscheidungssituationen ausgestaltet werden? Offensichtlich wäre ein Programmbefehl: >Du sollst nicht töten< für Kampfroboter selbstwidersprüchlich und für Autoroboter unterkomplex, da es im Straßenverkehr nun mal zu derartigen extremen Dilemmasituationen mit tödlichem Ausgang kommen kann. Um Lösungsstrategien für derartige komplexe Problemstellungen zu entwickeln

[42] Lacroix, Maschinen, vgl. S. 27 f.

bedarf es einer moralischen Urteilsfähigkeit, die „ein allgemeines Welt- und Hintergrundwissen“[43] voraussetzt, das (noch) nicht algorithmisch programmierbar ist.

Da es bis heute keine valide Möglichkeit gibt, (teil-) autonome Roboter moralisch so zu programmieren, dass diese ohne menschliche Einflussnahme humanitäre Grundsätze der Menschenwürde bei ihren Handlungsentscheidungen einfließen lassen können, ist es m. E. ethisch geboten, bis auf weiteres keine vollautonomen Roboter herzustellen und einzusetzen.

Allerdings gibt es ein gewichtiges utilitaristisches Gegenargument: Experten sind sich sicher, dass der Einsatz selbstfahrender Autos zu einer drastischen Verringerung tödlicher Unfallzahlen um bis zu 90 Prozent führen würde. Wäre dann nicht in Kauf zu nehmen, dass der tödliche Unfallverlauf der verbleibenden 10 Prozent auf einem ethisch problematischen Algorithmus beruht, zumal der Mensch in extremen Stresssituationen, wie bei Kriegshandlungen oder Autounfällen, sein ethisches Reflexionsniveau, zu dem er in Lehnstuhlsituationen befähigt ist, oftmals unterbietet. Wäre es demnach nicht vorzugswürdig, stressresistente Roboter autonom entscheiden zu lassen?

Meines Erachtens lässt sich diesem Einwand das Argument entgegenhalten, dass Roboter grundsätzlich hinter das ethische Reflexions- und Verantwortungsniveau des Menschen zurückfallen, da bisher kein ethisch valides Moralprogramm für Roboter entwickelt werden konnte. Dies vorausgesetzt, ist die (Mit-)steuerung von Kampf- und Fahrzeugrobotern durch den Menschen bis auf weiteres moralisch geboten.

Roboter werden auf absehbare Zeit kein zureichendes ethisches Reflexionsniveau erlangen. Robotern fehlt die menschliche Urteilskraft. „The idea that robots would some day become philosophers or ethicists – being able to critically reflect on who they are – belongs to the field of science fiction.“[44] Roboterethik sollte deshalb zunächst den Menschen befähigen, auf hochkomplexe Maschinen zu reflektieren und (noch) nicht die Befähigung von Robotern zur selbstständigen ethischen Reflektion unterstellen. Da Roboter (bis auf weiteres) zu keiner ethischen Reflektion, Kritik und Begründung von Normen und Handlungen befähigt sind, wäre es sinnlos, diese Maschinen als moralische Agenten für irgendetwas zur Verantwortung zu ziehen. Umso wichtiger ist eine intensive ethische Diskussion über den sinnvollen Gebrauch

[43] Mainzer, Intelligenz, S. 53.
[44] Capurro, Roboethics, S. 4.

und Missbrauch von Robotern durch den Menschen in unserer Gesellschaft. Es bedarf einer Ethik für Roboter, die selber nicht zum moralischen Urteil fähig sind.

3.3 Normative Orientierung für die soziale Interaktion von Menschen und Robotern mit geringer Autonomie

Zunehmend entwerfen Menschen Roboter für Menschen. Roboter verlassen die geschlossene, vollständig definierte Umgebung der Produktionshallen von Industrieunternehmen und begegnen Menschen im Haushalt und in Pflegeeinrichtungen. Derartige soziale Roboter werden für die Interaktion mit Menschen erschaffen. Je nach Design können sie weinen und lachen, Gefühle unterschiedlicher Art simulieren, informieren und helfen, sprechen, singen und umarmen. Das Design der Roboter wird auf den jeweiligen Anwendungsbereich abgestimmt und „ist das Bindeglied zwischen der Konstruktion und der Anwendung der Roboter. Es stellt sich vor die Technik. Die Nutzerin kommt nur mit den Funktionen in Berührung, die das Design zur Verfügung stellt. Das Design ist somit der Möglichkeitsraum für den Nutzer, die Technik zu verwenden.“[45] Menschen reagieren auf Roboter bzw. ihr Design je nach persönlichen Erfahrungen und Wertepräferenzen unterschiedlich. Ähnelt der Roboter dem Menschen zu sehr wirkt dies oftmals unheimlich. Der soziale Lebensraum des Menschen ist für Roboter ein komplexes Einsatzfeld mit wertrationalen Aspekten, die je nach Kulturraum differieren: „Culturally Europeans recognize robots as machines for labor, while Japanese and Koreans consider them as friends“.[46]

Wie ist diese komplexe Interaktion von Mensch und Roboter auszugestalten? Welche ethischen Erwägungen sind zu berücksichtigen? Dabei geht es sowohl um die Frage, wie Roboter kommunizieren und interagieren als auch darum, wie der Mensch Roboter in seine Lebenswirklichkeit integriert.

Eine Ethik der Robotik ist diesseits der zuvor behandelten Zukunftsperspektiven künstlicher (Super-)intelligenz und (teil-)autonomer Roboter zuallererst eine Ethik für Menschen, die Roboter entwerfen und produzieren bzw. einsetzen und nutzen. Oftmals sind die Produzenten der Roboter von den Nutzern zu unterscheiden. Produzenten und Marktanbieter, z. B. von Service-Roboter, erstreben eine Verkaufsrendite, die Anwender einen Nutzenvorteil. Kann dieses Spiel von Anbietern und Nachfragern dem Effizienzmechanismus der

[45] Koolwaay, Roboter, S. 206.
[46] Han et al., Robots, S. 101.

marktwirtschaftlichen Allokations- und Distributionsfunktionen überlassen werden oder bedarf es einer moralphilosophischen Orientierung für den Umgang des Menschen mit Robotern in Haushalt, Wirtschaft und Gesellschaft?

Eine propädeutische Aufgabe der Roboterethik ist die Klassifizierung von Gefährdungen des Menschen durch Roboter. Immer wieder kommen Anwender durch Roboter zu Tode. Allerdings kann eine falsche Handhabung von einfachen Geräten, wie Messer, Beil und Säge, oder komplexen Produktionsmaschinen ebenso fatale bzw. lethale Konsequenzen haben wie der missbräuchliche oder regelwidrige Einsatz von Robotern.

Fragen der technischen Sicherheit sind ebenso zu erwägen wie Probleme des Datenschutzes sowie psychische Probleme, die sich z. B. aus einer Abhängigkeit des Menschen von (Sex-) Robotern ergeben können.[47] Pragmatisch geht es um die dreistellige Relation des handelnden Menschen (1), des agierenden Roboters (2) und des Interaktionsprozesses von Mensch und Roboter (3).[48] „Humans and robots must be able to coordinate their actions so that they interact productively with each other. It is not appropriate (or even necessary) to make the robot as socially competent as possible. Rather, it is more important that the robot be compatible with the human's needs, that it matches application requirements; that it be understandable and believable, and that it provides the interactional support the human expect."[49]

Die Robotersoziologie untersucht, wie der Einsatz von Robotern die gesellschaftliche Wirklichkeit in den unterschiedlichen Anwendungsfeldern verändert. Ingenieure entwerfen mit Roboter nicht nur hochkomplexe Maschinen, sondern eine Spezie, die das menschliche Miteinander verändert. Der soziale Charakter von Robotern erschließt sich nur, wenn man über die digitale Infrastruktur hinaus die Wirkungsweise und Erscheinung von Robotern im Verhältnis zum Menschen untersucht.[50] Wie verändert der Einsatz von Robotern unsere menschlichen Verhaltensweisen?[51] Wie beeinflussen KI-gestützte Assistenzsysteme unsere Ernährung, Fitness und Stressbewältigung? Was erwartet der Mensch in einer konkreten Situation von Robotern, wie verändern Roboter die menschliche Erwartungshaltung? Wie

[47] Meister, Interaktivität, o. S.
[48] Je nach Autonomiegrad eines Roboters versuchen Autoren, diesem moralische Rechte und Pflichten zuzuschreiben und finden Analogien in der Tierethik (vgl. Birnbacher, Robotik, Abschnitt 3; vgl. Coeckelbergh, Animals, S. 197-204).
[49] Fong u.a., Robots, S. 160 f, zitiert in: Meister, Interaktivität.
[50] Vgl. Meister, Robot, S. 129.
[51] Vgl. Pfadenbauer, Robots, S. 149 f.

können Roboter zum menschlichen Wohlbefinden beitragen? „ Can human good appear in human-robot interaction (or relationships), or only in human-human interaction (and relationships)? Can human-robot interaction (relationships) contribute to human flourishing and happiness? Can such interactions constitute friendship, love, or relationships at all? Can they co-shape a flourishing community?"[52] (Wie) kann es Robotern gelingen, intuitiv und gefühlvoll mit Menschen zu interagieren?

Die Interaktion von Mensch und Roboter ist wie zwischenmenschliche Interaktionen kontextbezogen und für Fehlinterpretationen und Missverständnisse anfällig. Interaktionen können scheitern oder gelingen. [53] Dies gilt sowohl für die Interaktion des Menschen mit einem mechanischen Roboter als auch für das Verhältnis des Menschen zu einem virtuellen *Socialbot*. Sogenannte Softwareroboter (Bots) existieren ohne mechanische Glieder, die sie auf den ersten Blick als Roboter erkennbar machen würden. Bots werden u. a. für soziale Interaktionen im Internet erschaffen. Die fehlende Hardware-Konfiguration ist für die soziale Interaktion mit Menschen allerdings kein Effiziensnachteil – ganz im Gegenteil: Als virtuelle Softwarespezie gelingt es dem Bot, in Chatrooms so zu kommunizieren, als wäre er ein Mensch. In sozialen Medien sind Menschen von digitalen Bots nur schwer bis gar nicht zu unterscheiden. Beide Spezies kommunizieren mittels digitaler Codierung; spezifische analoge Fähigkeiten des Menschen sind irrelevant. Menschen und Bots kommunizieren entsprechend der Regeln digitaler sozialer Plattformen und streben nach Aufbau von Sozialkapital in Form von Likes und Friends bzw. Freundschaftsanfragen. Eine große Resonanz bei Instagram, z. B. gemessen an Followern, steigert den Marktwert sogenannter Influencer. Die Beeinflussung von Kauf- oder Wahlentscheidungen durch die Multiplikation von Meinungen führt zu einer Nutzensteigerung der Plattformbesitzer sowie der Influencer, Unternehmen bzw. Parteien. Da Sozialbots in der digitalen Kommunikation schneller sind als der Mensch, keine Erholungsphasen benötigen und nicht krank werden, können sie Meinungen effizienter multiplizieren als Menschen. Da der Mensch sich den von ‚bits' und ‚bytes' geprägten Regeln der OSN (Online Social Networks) anpasst, auf argumentativen Tiefgang und feine Zwischentöne verzichtet, kann es den Socialbots gelingen, den Menschen in dieser virtuellen Welt zu übertreffen.

„The *anthropos* in online social networks has shed off a lot of its critical advantages over the robot. That is why socialbots have good chances to be successful in presenting themselves as

[52] Coeckelberg, Roboethics, S. 220.
[53] Vgl. Companga, Muhl, Interaktion, S. 27.

humans." [54] Der Effizienzerfolg der Socialbots liegt nicht nur in den Codierungsfähigkeiten der Softwareingenieure begründet, sondern im Verzicht des Menschen, seine analogen Fähigkeiten einzusetzen, wenn er den digitalen Raum der OSN betritt.[55] Beispielsweise entspricht die sprachliche und argumentative Struktur von menschlichen Twitter-Botschaften oftmals einfachsten Schemata, die von Socialbots mühelos kopiert werden können. So gelingt es Socialbots wie dem sogenannten James M. Titus, mit der Versendung von Katzenphotos und Kurzberichten, unterstützt durch digital verbundene Bots-Netzwerkfreunde, in kurzer Zeit mehr Likes und Friends zu gewinnen als menschliche Kommunikationsteilnehmer.[56] „Many acts of human behavior on Twitter are so bit-like that the socialbot has no trouble blending seamlessly into the stream of platformed sociality."[57]

Die Regeln der OSN ermöglichen es Softwareroboter, mit Menschen zu interagieren, ohne dass diese es merken. Dabei führt die zunehmende Anzahl von Socialbots, die als solche unerkannt in hocheffizienter menschenähnlicher Weise im OSN kommunizieren, zu einer gesellschaftlich relevanten meinungsbildenden Einflussnahme. So prägen Socialbots bereits heute das öffentliche Meinungsklima, lenken Wahlentscheidungen und Konsumtrends und verbreiten Propaganda und (Miss-)informationen. Studien zeigen ebenso wie die Praxis, dass die Infiltration von OSN durch Socialbots in kurzer Zeit gelingen kann, private Daten menschlicher Nutzer weitgehend ungeschützt sind und die Sicherheitssysteme sozialer Netzwerke nicht ausreichen, um Individuen, Unternehmen und Parlamente vor Missbrauch zu bewahren.[58]

3.4 Moralentwicklung durch Dialogfähigkeit

Dialogbezogenes Begründen setzt den Logosgrundsatz voraus, mit dem Sokrates seine Maxime für das rechte Handeln formuliert: „Denn nicht jetzt nur, sondern schon immer habe ich das an mir, daß ich nichts anderem von mir gehorche als dem Satze, der sich mir bei der Untersuchung als der beste zeigt."[59] Die argumentative Untersuchung des Logos (von Cicero mit lat. ratio übersetzt) führt Diskurspartner vom *Meinen* zum *Wissen*. Eine Meinung kann erst

[54] Bakardjieva, Rationalizing, S. 248.
[55] Vgl. Gehl, Software, S. 25.
[56] Beispielhaft ist auch die Resonanz, die Lil Miquela in OSN erzeugt. Mit 1,4 Millionen Followern ist sie bei Instagram eine der bedeutendsten Influencer. Miquela ist eine virtuelle Kunstfigur im Cyberspace (Avater). Sie gibt vor, 19 Jahre alt zu sein und in Los Angeles zu leben. Sie habe spanisch-brasilianische Wurzeln, engagiert sich für die Rechte schwarzer Frauen und trägt bevorzugt Chanel und Prada.
[57] Bakardjieva, Rationalizing, S. 248.
[58] Vgl. Boshmaf, Muslukhov u. a., Botnet, S. 1-19.
[59] Platon, Gorgias, 482 c.

dann Geltung beanspruchen, wenn sie mit der Vernunftrolle des Denkenden im Einklang steht. [60] Sokrates formuliert in Platons Gorgias diese sinnkritische Forderung der Übereinstimmung mit sich selbst: „Und ich wenigstens, du bester, bin der Meinung (...), daß eher die meisten Menschen nicht mit mir übereinstimmen, sondern mir widersprechen mögen, als daß ich allein mit mir selbst nicht zusammenstimmen, sondern mir widersprechen müßte.“[61] Der Diskurs verpflichtet, auf die Argumente des anderen einzugehen, diese in Rede und Gegenrede zu prüfen und kritisch zu hinterfragen, um im Gegenzuge die eigenen Behauptungen kritisieren zu lassen und das bessere Argument der Dialogpartner anzuerkennen. Der Diskurs ist das Verfahren, um Geltungsansprüche aus unterschiedlichsten Rationalitätsperspektiven zu prüfen. Dialogizität, so Raya A. Jones (*1953) ist der Schlüssel, um die Sozialität von Robotern zu (re-) konzeptualisieren: „A robot will become truly social only if and when it autonomously and inescapably partakes in dialogical action. Irrespect of whether this criterion can serve as an achievable benchmark for near-future robotics, its absence in the technology-oriented discourses is significant“.[62]

Das Kriterium der Dialogizität als Voraussetzung für wirkliche Sozialität setzt voraus, dass alle möglichen Anspruchs- und Argumentationssubjekte, unabhängig von ihrer sozialen Stellung als gleichberechtigte Dialogpartner, anerkannt werden. Die soziale Interaktion von Mensch und Roboter ist jedoch immer mehr oder weniger asymmetrisch. Es fehlt dem Roboter an kognitiver und sprachlicher Kompetenz. Auf Grund des defizitären Reziprozitätsverhältnisses von Mensch und Roboter ist zu prüfen, unter welchen Voraussetzungen die >Gegenseitigkeit< im Diskurs ein kontrafaktisches roboterethisches Kriterium bilden könnte.

Die menschliche Moral- und Sozialgeschichte kennt als wichtigsten Maßstab die Entfaltung der Reziprozität, d. h. das zunehmende Gegenseitigkeitsverhältnis von Alter und Ego. Mittels des Rollentausches vermag es der Mensch, den Standpunkt des Anderen einzunehmen. Ich lerne, die Welt aus der Perspektive des Anderen zu verstehen. In dem Maße, wie sich dieses ethische Reflexionsvermögen entwickelt, wird der soziale Akteur fähig, vom konkreten Anderen zu abstrahieren und sich selber im Verhältnis zum allgemeinen Anderen zu denken. Wie die Untersuchungen von Jean Piaget (1896-1980) und Lawrence Kohlberg (1927-1987) verdeutlichen, ist die Moralentwicklung des Menschen, gemessen am Kriterium der verallgemeinerbaren Gegenseitigkeit, sehr unterschiedlich ausgeprägt.

[60] Vgl. Böhler, Verbindlichkeit, S. 223 f.
[61] Platon, Gorgias 482 c.
[62] Jones, Robot, S. 22 f.

Können soziale Roboter bzw. intelligente künstliche Agenten auf einer Stufenleiter der Moralentwicklung verortet werden? Auf der ersten Stufe der Nutzenorientierung orientieren sich Akteure an den unmittelbaren Handlungskonsequenzen, d. h. an Lob und Strafe. Wie der AlphaGo-Agent reagiert der Mensch auf unmittelbare Belohnungsanreize. Je intensiver eine Handlung positiv unterstützt wird, desto häufiger wird diese ausgeführt, um so den unmittelbaren Nutzen zu maximieren. Auf der nächsten Stufe wird der Andere als Instrument der eigenen Nutzenmaximierung ins Spiel der Interessensverfolgung gebracht. Im weiteren Verlauf der Moralentwicklung entdecken soziale Akteure, dass Tauschakte durch Vertrauen und Regelgehorsam erleichtert werden. Zumindest das Handeln in klar definierten Regelsystemen ist für soziale Roboter geübt, auch wenn ihnen das Bewusstsein und die Reflexionsfähigkeit fehlen, das regelkonforme Verhalten als Rechtsgehorsam zu qualifizieren.

Roboter können (noch) nicht auf ihr eigenes Tun der intentio directa Bezug nehmen. Ihnen fehlt die kritische Reflexionsperspektive der intentio obliqua. Deshalb ist nicht abzusehen, ob und wann künstliche Superagenten mit dem Menschen einen Sozialvertrag schließen können oder sich gar auf der höchsten Stufe der Moralentwicklung am universalen ethischen Prinzip der verallgemeinerbaren Gegenseitigkeit orientieren werden.

Wie wir oben gesehen haben, ist sich die Scientific Community uneins bzgl. der technologischen Entwicklungschancen vollkommen autonomer Roboter. Einigkeit scheint jedoch in der Notwendigkeit zu bestehen, bereits für Roboter mit geringer Autonomie das Sozialverhalten zu optimieren. M. E. bietet das kurz skizzierte Moralentwicklungsschema von Piaget/Kohlberg eine Roadmap für die Roboterethik. Für das Hochschreiten auf der Stufenleiter bedarf der soziale Akteur einer zunehmenden Dialogkompetenz. Nur in der sinnvollen Argumentation mit allen Betroffenen und Beteiligten können die Geltungsansprüche aus den relevanten Rationalitätsperspektiven geprüft werden.

Maria Bakordjieva postuliert, dass durch eine kommunikative Rationalisierung der Roboterethik die zweckrationale Outputorientierung der Robotik überwunden werden könnte.[63] Aus diskursethischer Perspektive bedarf es dazu jedoch nicht nur des deontologischen Prinzips nach einem argumentativen Konsens aller sinnvoll Argumentierenden zu streben. Vielmehr ist dieser Universalisierungsgrundsatz durch das teleologische Regulativprinzip zu ergänzen, die

[63] Vgl. Bakardjieva, Rationalizing, S. 254.

realen Kommunikationsverhältnisse im Blick auf die Verhältnisse einer idealen Kommunikationsgemeinschaft zu verbessern.

Damit qualifiziert die diskursethische Architektur für einen hybriden Ansatz der Roboterethik, der die *Top-down*-Perspektive mit einer Bottom-up-Reflexion kombiniert.[64] Top-down gilt die Verpflichtung zur rationalen Argumentation unbedingt, sowohl für den Menschen, der Roboter gestaltet, produziert und anwendet, als auch für zukünftige mehr oder weniger autonome und moralfähige künstliche Agenten.

Je nach Kontext sind die Möglichkeiten, mit allen Betroffenen und Beteiligten in eine sinnvolle Kommunikation zu treten, unterschiedlich gegeben. Deshalb gilt es, Bottom-up Voraussetzungen dafür zu schaffen, dass die Top-down Regel soweit es geht befolgt werden kann.

Roboterethiker befürchten, dass die Würde des Menschen durch die Entwicklung autonomer Agenten mit starker künstlicher Intelligenz zukünftig gefährdet würde. Menschenwürde und Zukunftsverantwortung bilden den substanziellen Kern der Diskursmoral. Es wäre deshalb zu erwägen, ob und wie Roboter entsprechend dieses ethischen Regulativs programmiert und gesteuert werden können.

[64] Wallach, Allen, Moral, S. 83-124; vgl. Loh, Roboterethik, S. 5-7.

LITERATUR

Asimov, Isaac (Aufbruch): Der Aufbruch zu den (Sternen), München1984

Asimov, Isaac (Robotergeschichten): Alle Robotergeschichten, Köln 1987

Bakardjieva, Maria (Rationalizing): Rationalizing Sociality, an unfinished Script for Socialbots. In: The Information Society, Nr. 31, 2015, S. 244-256

Birnbacher, Dieter (Robotik): Ethik und Robotik, Vortragsfolien, Heinrich-Heine-Universität Düsseldorf

Böhler, Dietrich (Verbindlichkeit): Verbindlichkeit aus dem Diskurs, Freiburg, München 2013

Boshmaf, Yazan/Muslukhov, Ildar/Beznosov, Konstantin/Ripeanu, Matei (Social Botnet): Design and Analysis of a Social Botnet. In: Elsevier Journal of Computer Network, No. 9, July 2012

Bostrom, Nick (Superintelligence): How long before Superintelligence. In: Linguistic and Philosophical Investigations, Vol. 5, No. 1, 2006, S. 11-30

Burkholder, Leslie (Searle): John Searle und das chinesische Zimmer. In: Bruce, Michael/Barbone, Steven (Hg.): Die 100 wichtigsten philosophischen Argumente, S. 335-337.

Čapek, Karel (Robots): R.U. R.: Rossum's universal robots, trans. P. Selver and N. Playfair. http://preprints.readingroo.ms/RUR/rur.pdf

Capurro, Rafael (Roboethics): Quest for Roboethics. In: Robohub, 14.02.2017, S. 1-11

Coeckelbergh, Mark (Animals): Humans, Animals and Robots: A Phenomenological Approach to Human-Robot Relations. In: International Journal of Social Robot, Nr. 3, 2011, S. 197-204

Coeckelberg, Mark (Roboethics): Personal Robots, Appearance, and Human Good: A Methodological Reflection on Roboethics. In: International Journal of Social Robot, No. 1, 2009, S. 217-221

Companga, Diego/Muhl, Claudia (Interaktion): Mensch-Roboter Interaktion, S. 19-34. In: Stubbe, Julian/Töppel, Mandy (Hg.): Muster und Verläufe der Mensch-Technik-Interaktivität, Bd. zum gleichnamigen Workshop am 17./18. Juni 2011, TUTS-WP-2012 Berlin

Devlin, Hannah (Ethics): Do no harm, don't discriminate: official guidance issued on robot ethics. In: The Guardian, 18.09.2016, S. 1-4

Fong, Terrence/Nourbakhsh, Illah/Dautenhahn, Kerstin (Robots): A Survey of Socially Interactive Robots. In: Robotics and Autonomous Systems, Nr. 42, 2003, S. 143-166

Gehl, Robert W. (Software): Reverse Engineering Social Media: Software, Culture, and Political Economy in New Media Capitalism, Philadelphia 2014

Good, Irvin John (Machine): Speculations Concerning the First Ultraintelligent Machine. In: Alt, F. L./Rubinoff, M. (Hg.): Advance in Computers, Academic Press, 6, 1965, S. 31-88

Han, Jeonhye/Hym, Eunja et. Al (Robots): The Cross – Cultural Acceptance of Tutoring Robots with Argumental Reality Services. In: International Journal of Digital Content Technology and its Application 3(2), S. 95-102

Jonas, Hans (Verantwortung): Das Prinzip Verantwortung, Frankfurt/Main 1979

Jones, Raya A. (Robot): What makes a robot ‚social'? To be published in: Social Studies of Science

Koolwaay, Jens (Roboter): Die soziale Welt der Roboter, Bielefeld 2018

Kuhlmann, Wolfgang (Letztbegründung): Reflexive Letztbegründung, Freiburg, München 1985

Lacroix, Alexandre (Maschinen): Sind Maschinen moralischer als wir? In: Philosophie Magazin Nr. 04, 2018, S. 27-31

Lierfeld, Karl Johannes (AI): Can we create strong & safe AI?, Habilitation thesis, Universität Köln, erscheint in 2018

Loh, Janina (Roboterethik): Roboterethik. In: Information Philosophie, Heft 1, 2017, S. 1-11

Mainzer, Klaus (Intelligenz): Künstliche Intelligenz – Wann übernehmen die Maschinen?, Berlin, Heidelberg 2016

Meister, Martin (Interaktivität): Mensch-Technik-Interaktivität mit Servicerobotern. In: KIT(Hg): Technikfolgenabschätzung – Theorie und Praxis, 20. Jg., Heft 1, April 2011, S. 46-52

Nida-Rümelin, Julian (Humanismus): Digitaler Humanismus. In: Auf die Zukunft, Das Magazin zum Innovationstag 2017, FAZ-Beilage, 05.10.2017

Pfadenbauer, Michaela (Robots): „On the Sociality of Social Robots", in: Science, Technology & Innovation Studies, Vol. 10 No. 1, Jan. 2014, S. 135-153

Platon (Gorgias): Gorgias, Hamburg 1981

Schätzing, Frank: Die Tyrannei des Schmetterlings, Köln 2018

Schnetter, Martin (Robotik): Robotik und ihre Regulierung, Berlin 2016

Silver, David/Schnittwieser, Julian/Simonyan, Karen et al (Game): Mastering the Game of Go without Human Knowledge, Nature, 550 (7676), 2017, S. 354-359

Tetens, Holm (Argumentieren): Philosophisches Argumentieren, 4. Aufl., München 2014

Tzafestas, Spyros G. (Roboethics): Roboethics, Heidelberg u. a. 2016

Ulam, Stanislaw (John von Neumann): Tribute to John von Neumann. In: Bulletin of the American Mathematical Society, 64.3, part 2, May, 1958, S. 1-49

Veruggio, Gianmarco (Roadmap): The EURON Roboethics Roadmap, Workshoppaper, Euron Roboethics Atelier, Humanoids '06, Genua 2006, S. 612-617

Veruggio, Gianmarco (Roboethics): The Birth of Roboethics, Workshoppaper, ICRA 2005, IEEE International Conference on Robotics and Automation, Barcelona, 18. April 2005

Vogt, Markus (Nachhaltigkeit): Prinzip Nachhaltigkeit, 3. Aufl., München 2013

Wallach, Wendell/Allen, Colin (Moral): Moral Machines, Oxford 2010

Weber, Max (Kapitalismus): Die protestantische Ethik und der Geist des Kapitalismus. In: Winckelmann, J. (Hg.): Die protestantische Ethik I, 9. Aufl., Gütersloh 2000

Weber, Max (Objektivität): Die >>Objektivität<< sozialwissenschaftlicher und sozialpolitischer Erkenntnis. In: Sukale, M. (Hg.): *Max Weber*, Schriften zur Wissenschaftslehre, S. 21-101, Stuttgart 2002

Weizenbaum, Joseph (Computer): Die Macht der Computer und die Ohnmacht der Vernunft, Frankfurt/M. 1977

Zwierlein, Eduard (Wirtschaftsethik): Wirtschaftsethik und der Ausgleich von Ökonomie und Ökologie. In: Allgemeine Zeitschrift für Philosophie, Nr. 19.2, S. 79 ff., 1994

ABKÜRZUNGEN

EURON	European Robotics Research Network
KIT	Karlsruher Institut für Technologie
MIT	Massachusetts Institute of Technology
OSN	Online Social Network