Matthias Maring (dir.)

Zur Zukunft der Bereichsethiken – Herausforderungen durch die Ökonomisierung der Welt

KIT Scientific Publishing

Roboterethik

Janina Sombetzki

Publisher: KIT Scientific Publishing
Place of publication: KIT Scientific Publishing
Year of publication: 2016
Published on OpenEdition Books: 13 septembre 2019
Serie: KIT Scientific Publishing
Electronic ISBN: 9791036538254

http://books.openedition.org

Electronic reference
SOMBETZKI, Janina. *Roboterethik* In:: *Zur Zukunft der Bereichsethiken – Herausforderungen durch die Ökonomisierung der Welt* [Online]. Karlsruhe: KIT Scientific Publishing, 2016 (Erstellungsdatum: 12 janvier 2021). Online verfügbar: <http://books.openedition.org/ksp/4626>. ISBN: 9791036538254.

Roboterethik

Janina Sombetzki

Einleitung – Roboterethik als Bereichsethik

Die Roboterethik sieht sich immer wieder mit zwei Vorwürfen konfrontiert, die ihren Status als Bereichsethik in Frage stellen: Zum einen habe sie keinen spezifischen Gegenstand, da sich Ethik nicht mit Unbelebtem beschäftige. Doch selbst wenn artifizielle Systeme zu Recht in den Fokus der ethischen Reflexion geraten würden, ließen sich – so der zweite Einwand – mit ihnen im Blick keine neuen, sondern in anderen ethischen Arenen längst formulierte und ausgetragene Fragen stellen.

Mit dem zuerst formulierten Einwand beschäftige ich mich im ersten Abschnitt dieses Artikels. In aller Knappheit ließe sich ihm hier damit begegnen, dass, falls sich herausstellen sollte, dass Roboter selbst keine moralischen Handlungssubjekte sein, sie dennoch ihren gerechtfertigten Platz im moralischen Universum einnehmen könnten. Schließlich existieren eine ganze Reihe von (teil-)unbelebten Entitäten, denen wir einen Wert zuzusprechen gewillt sind – Landschaften, Ökosystemen, dem Planeten Erde, aber auch Häusern, Autos, Smartphones und der Schnuffeldecke frühester Kindertage. Um was für eine Art von Wert es sich im Falle artifizieller Systeme handelt, bleibt freilich zu diskutieren, doch wo, wenn nicht in der Ethik, sollte eine solche Diskussion geführt werden?

Dem zweiten Einwand ist – das werde ich im Folgenden zeigen – nicht viel entgegenzusetzen, jedoch handelt es sich dabei nicht um eine spezifisch die Roboterethik treffende Kritik, sondern ließe sich letztlich auf alle Bereichsethiken beziehen, solange wir den Menschen als Ausgangs-, Dreh- und Angelpunkt ethischen Nachdenkens begreifen und dieser in allen Sphären der ethischen Überlegung anzutreffen ist. Darüber hinaus kann man diesen Vorwurf eines „epistemischen Anthropozentrismus“[1] in der Tat positiv wenden, indem man die Tatsache einer generellen (Familien-)Ähnlichkeit ethischer Fragen in allen Bereichsethiken als Aktualität und Flexibilität der philosophischen Reflexion insgesamt deutet: Tradierte ethische Fragen sind immer noch zeitgemäß und relevant. Darüber hinaus ist die Philosophie in der Lage, längst bekannte Probleme an aktuelle Gegenstände anzupassen, ihre z.T. antiken Fragen also mit einer Welt abzugleichen, die mittlerweile über künstliche Intelligenz und moralisches Lernen in artifiziellen Systemen diskutiert.

---

1 Krebs (1997, 343).

Insbesondere mit der Tierethik kann sich die Roboterethik verwandt fühlen, „[t]he machine question […] is the other side of the question of the animal."[2] René Descartes – wie David Gunkel zeigt – hatte Tieren und Maschinen denselben ontologischen Status zugeschrieben.[3] Erst im 20. Jahrhundert wurde diese ontologische Gleichstellung von Tier und Maschine immerhin zugunsten der Tiere aufgehoben. Innerhalb der philosophischen Anthropologie bspw. setzt man sich mit dem Verhältnis von Mensch und Tier auseinander und sucht über das Tier das Wesen des Menschen näher zu fassen.[4] In der Tierethik denkt man nicht nur darüber nach, inwiefern Tiere Wertträger darstellen und damit auch einen Platz im Universum der moralisch bedenkenswerten Phänomene einnehmen, sondern auch, inwiefern sie darüber hinaus als (rudimentäre) moralische Akteur gelten können.[5] Künstliche Systeme haben es bislang noch nicht geschafft, in den Horizont der philosophisch-anthropologischen Reflexion inkludiert zu werden.[6] Doch seit einigen Jahrzehnten beginnt sich immerhin die Roboterethik als verhältnismäßig junge Bereichsethik innerhalb der Philosophie zu etablieren.

Die Roboterethik stellt – wie oben angedeutet – durchaus traditionelle Fragen, gibt einigen Herausforderungen, vor die sich die Tierethik bereits gestellt sah, ein neues Gewand und wirft den Menschen letztlich auf sich selbst zurück.[7] Welche Kompetenzen erachten wir bspw. dafür grundlegend, um als Handlungssubjekte gelten zu können? Was ist darüber hinaus Bedingung für moralische Akteursfähigkeit? Mit welchen moralischen Prinzipien und Werten sollten wir artifizielle Systeme ausrüsten? Auf was für ein moralisches Selbstverständnis lässt es schließen, wenn wir Roboter ‚schlecht' behandeln?[8] In welchem Nahbereich des Menschen – Industrie-, Miltär-,

2 Gunkel (2012, 5).

3 Vgl. Gunkel (ebd. 3).

4 Für das Programm einer philosophischen Anthropologie im Sinne ihrer Gründungsväter Scheler, Plessner und Gehlen ist insbesondere die Beziehung zwischen Mensch und Tier als grundlegend zu verstehen – ausgehend von der Primatenforschung (vgl. Köhler 1921; von Scheler, Plessner und Gehlen seien nur beispielhaft genannt: Scheler 1908/09; Plessner 1928, 1946; Gehlen 1997).

5 Vgl. Gunkel (2012, 4–5).

6 Vgl. Sombetzki (2016).

7 Die folgenden Sammelbände, Monographien und Artikel bieten einen Einstieg in die Roboterethik: Hilgendorf (2014); Lin et al. (2011, 2012); Anderson/Anderson (2011); Mainzer (2010); Capurro/Nagenborg (2009); Brey et al. (2008); Allen et al. (2006); Anderson et al. (2006); Asaro (2006); Moor (2006); Edgar (2003); Laudon (1995).

8 Ähnlich wie sich bereits Immanuel Kant in § 17 des zweiten Teils seiner *Metaphysik der Sitten* gegen Tierquälerei ausspricht, da diese zu einer Verrohung des Menschen führe, plädiert Kate Darling für Roboterrechte, da es dem Menschen dann eher gelinge „menschlich" zu bleiben; URL: http://www.zeit.de/digital/internet/2013-

Medizin-, Altenpflege- und Servicerobotik, um nur einige zu nennen – wollen wir uns auch weiterhin nur oder zumindest in einem signifikanten Ausmaß auf menschliche und nicht auf artifizielle Expertise verlassen? Im Folgenden gebe ich einen Überblick über die Arbeitsfelder und Forschungsfragen der Roboterethik und zeige ihre Chancen und Perspektiven für die philosophische Reflexion als Bereichsethik auf.

Der Begriff „Roboter" geht auf das tschechische Wort „robota" zurück, was so viel bedeutet, wie Arbeit, Frondienst und Zwangsarbeit. 1920 wurde der Begriff „robot" von dem Künstler Josef Čapek geprägt, und sein Bruder Karel Čapek gebrauchte den Begriff „labori" in seinem Theaterstück *R.U.R.* (*Rossum's Universal Robots*, 1921) für humanoide Apparaturen, die Serviceleistungen und Arbeit an des Menschen Stelle übernehmen. Catrin Misselhorn schlägt vor, einen Roboter als eine besondere Art von elektro-mechanischer Maschine, als spezifische Apparatur zu verstehen, die aus einer Einwicklungseinheit (einem Prozessor) besteht, aus Sensoren, die Daten oder Informationen über die Welt sammeln und aus einem Effektor oder Aktor, die Signale in zumeist mechanische Abläufe übersetzen. Das Verhalten eines Roboters ist oder wirkt zumindest bis zu einem gewissen Grad autonom.[9] Roboter können in einer Weise auf die Umgebung Einfluss nehmen und in sie hinein wirken, in der Computer nicht in der Lage sind.[10] In diesem Artikel gebrauche ich die Begriffe „Roboter" und „artifizielles System" synonym. Nicht bei allen artifiziellen Systemen, wohl aber bei den für die folgenden Überlegungen relevanten, handelt es sich um Roboter – bereits beim gebräuchlichen Computer fängt der Graubereich an, der von Technikphilosoph*innen ausgeleuchtet zu werden verdient und in dem bereits Isaac Asimov die roboternahe Sphäre vermutet hat, in der wir neben Computern auch Maschinen, Automaten und weiteren Verwandten und Bekannten der Roboter begegnen.[11]

1. Roboterethik – zwei Arbeitsfelder

Innerhalb der Disziplin der Roboterethik sind zwei Felder zu unterscheiden: Die einen fragen danach, inwiefern Roboter als sogenannte moral patients zu verstehen sind, also passiv als Träger*innen moralischer Rechte bzw. inwiefern ihnen ein moralischer Wert zukommt. Die anderen interessieren sich dafür, ob und ggf. inwiefern Roboter sogar moral agents sein könnten, also aktiv Träger*innen moralischer Pflichten bzw. moralische Handlungs-

05/roboter-ethik-kate-darling und http://futurezone.at/science/vorbild-tierschutz-brauchen-roboter-rechte/24.595.990 [Stand: 28.01.2016].

9 Misselhorn (2013).

10 Vgl. hierzu auch Ichbiah (2005).

11 Asimov (1982, 53).

subjekte.[12] Beide Arbeitsbereiche ergänzen einander. Die Gruppe der moral agents ist gegenüber der der moral patients exklusiver; für gewöhnlich zeichnen wir nur Menschen und längst nicht alle mit Moralfähigkeit im genuinen Sinne des Wortes aus – einige Menschen wie bspw. Kinder und solche mit spezifischen geistigen und körperlichen Einschränkungen können temporär oder sogar generell von ihrer Moralfähigkeit ganz oder teilweise entschuldigt werden. Auch bestimmte Umstände und Unfälle erlauben eine Freisprechung von Moral.

Einer ganzen Reihe von Wesen und Dingen wie z.B. Tieren, Pflanzen, aber auch Gegenständen wie dem teuren Auto, dem Smartphone oder Haus wird indes ein moralischer Wert zugeschrieben – zumindest in dem Sinn, dass diese Entitäten moralisch bedenkenswert sind, wenn ihnen vielleicht auch kein Eigen- sondern nur ein hoher instrumenteller Wert beigemessen wird. Als moralisches Handlungssubjekt hat man zugleich einen Platz im Kreis der Wertträger*innen – dies gilt allerdings nicht umgekehrt. Lebewesen und Gegenständen kann man abhängig von der Perspektive einen moralischen Wert zuschreiben; eine anthropozentrische Position argumentiert bspw. dafür, dass nur dem Menschen ein Eigenwert zukommt, weitere Ansätze stellen der Patho-, der Bio- und der Physiozentrismus dar.[13] Interessant ließe sich die Überlegung an, inwiefern die Inklusion von artifiziellen Systemen in den Horizont der mit einem Eigenwert ausgestatteten Phänomenen eine weitere Perspektive eröffnet, die all das mit einem Eigenwert bemisst, das in einer spezifischen (in diesem Artikel noch zu erörternden) Weise gesteuert oder programmiert bzw. lernfähig ist.

Innerhalb des Arbeitsbereichs zu Robotern als Wertträger*innen wird das menschliche Verhalten gegenüber artifiziellen Systemen in den Blick genommen. Hier geht es darum, wie mit Robotern umzugehen ist und inwiefern ihnen (ggf. analog zu Tieren und kleinen Kindern) ein moralischer Wert zukommt, selbst wenn man sich darüber einig sein sollte, dass sie selbst nicht zu moralischem Handeln in der Lage sind. In dieses Themenfeld fallen alle Fragen, die artifizielle Systeme als Werkzeuge oder als Ergänzungen des Menschen verstehen wie bspw. in der Formulierung von Ethikkodizes in Unternehmen,[14] bzgl. der Frage, inwiefern Beziehungen zu und mit Robotern denkbar und wünschenswert sind,[15] inwiefern man Roboter ‚versklaven' kann[16] und wie der Einsatz von artifiziellen Systemen zu Therapiezwecken

12 Floridi/Sanders (2004, 349).
13 Vgl. Krebs (1997, 345).
14 Vgl. May (2014).
15 Vgl. Levy (2012); Scheutz (2012); Whitby (2012).
16 Vgl. Petersen (2007, 2012).

zu beurteilen ist[17]. Innerhalb dieses Arbeitsbereichs verbleibt die moralische Kompetenz und Kompetenzkompetenz bei den menschlichen Designer*innen (und u.U. auch Nutzer*innen) artifizieller Systeme. Die menschlichen ‚Eltern' entscheiden über die Moral ihrer Geschöpfe und darüber, wer im Falle eines Unfalls Verantwortung trägt. Sicher ist, dass den künstlichen Kreaturen keinerlei Verantwortung zuzuschreiben ist, da es ihnen an den Kompetenzen mangelt, die als Bedingung für die Möglichkeit von Verantwortungszuschreibung gelten.[18]

Innerhalb des Arbeitsfelds zu Robotern als moralischen Handlungssubjekten fragt man insbesondere danach, inwiefern Roboter zu moralischem Handeln in der Lage sind und folglich, über welche Kompetenzen sie in welchem Maße verfügen müssen. Interessieren sich die einen in diesem Bereich eher für die Zuschreibung von Freiheit als Bedingung für moralisches Handeln, befassen sich andere eher mit kognitiven Kompetenzen (Denken, Verstehen, Geist, Intelligenz, Bewusstsein, Wahrnehmung und Kommunikation) und wieder andere mit Empathie und Emotionen (auf alle drei Kompetenzgruppen wird im dritten Abschnitt dieses Artikels eingegangen).[19]

Beiden Arbeitsfeldern innerhalb der Roboterethik liegt die Frage zugrunde, was Moral bzw. was Ethik ist und wie moralische Urteile gefällt werden.[20] Auch hier lassen sich (allerdings nicht in diesem Artikel) verschiedene Positionen unterscheiden; in einem ersten Schritt könnte man vorschlagen, allen Wesen Moralfähigkeit zuzuschreiben, die in Situationen geraten, in denen moralische Entscheidungen zu treffen sind. So gehen bspw. Wendell Wallach und Colin Allen in ihrem Werk *Moral Machines. Teaching Robots Right from Wrong* (2009) vor. Sie beschreiben vor der Reflexionsfolie von Philippa Foots klassischem Gedankenexperiment der Trolley Cases den Fall

17 Vgl. Becker/Rüegsegger (2013); Misselhorn et al. (2013); Borenstein/Pearson (2012); Krings et al. (2012); Sharkey/Sharkey (2012); Datteri/Tamburrini (2009).

18 Vgl. hierzu Sombetzki (2014, Kapitel 2). Verantwortung wird in der Mensch-Maschine-Interaktion häufig, wenn auch nicht immer, als rechtliche Kategorie diskutiert; vgl. Neuhäuser (2014); Lokhorst/van den Hoven (2012); Maring (2008); Asaro (2007); Floridi/Sanders (2004); Friedman/Kahn (1992).

19 Vgl. Coeckelbergh (2009); Boden (2006); Sterrett (2006); Dorffner (2004); Clark (1999, 2003); Dennett (1998, 2006).

20 Die schwedische Serie *Real Humans – Echte Menschen* (*Äkta människor*, 2012 von Lars Lundström) vereint ab der ersten Folge der ersten Staffel beide Arbeitsfelder miteinander, indem es auf der einen Seite um die Einführung von Hubots (für *Hu*man Ro*bots*) geht, die als hochkomplexe Serviceroboter im Dienstleistungssektor, in der Industrie und in Privathäusern arbeiten. Auf der anderen Seite treten sog. „freie" Hubots auf, die über einen besonderen „Code" verfügen (also über eine spezielle algorithmische Struktur), durch die sie zu autonomen Handlungssubjekten werden.

von „‚driverless' train systems"[21], in denen in London, Paris und Kopenhagen bereits seit Mitte der 1960er Jahre Menschen nur als Fahrgäste anzutreffen sind.[22] Eine moralische Entscheidung wird – so Wallach und Allen – bereits dann gefällt, wenn sich auf den Gleisen Menschen befinden, die der Zug zu überrollen droht. Der Zug ‚urteilt', indem er dazu programmiert ist, immer dann unverzüglich zu stoppen, wenn sich Menschen auf den Gleisen aufhalten, selbst wenn damit ggf. Unfälle im Zuginnern in Kauf genommen werden müssen (hierauf komme ich im zweiten Abschnitt dieses Artikels zurück).

Um diesem Gedanken, dass all den Wesen (rudimentäre) Moralfähigkeit zuzuschreiben ist, die wie Menschen in Situationen geraten, in denen moralische Entscheidungen zu treffen sind, noch ein wenig zu folgen, stellen wir uns einen selbstfahrenden Krankentransport in einer Notfallsituation vor, der sich mit bspw. einer schwer verletzten Jugendlichen auf dem Weg ins nächste Krankenhaus befindet. Absolut unvorhersehbar überquert nun unmittelbar vor dem Krankentransport eine weitere Jugendliche die Straße. Ein plötzliches Bremsen hätte den Tod der Transportierten zur Folge, ein Weiterfahren den Tod der Fußgängerin. Sowohl eine Entscheidung für die eine als auch für die andere Option stellt ein moralisches Urteil dar bzw. beruht auf einem moralischen Prinzip. Es kann zunächst keine Rede davon sein, dass ein autonomer Krankentransport, ausgerüstet mit einer spezifischen algorithmischen Struktur, im genuinen Sinne des Wortes moralisch handelt. Allerdings ähnelt diese Situation äußerlich einer solchen, in der sich auch ein Mensch befinden könnte. In ihrer von außen beobachtbaren phänomenologischen Qualität gleicht die Maschine – so Wallach und Allen – rudimentär einem Menschen. Das genügt, um zumindest ein Nachdenken über Roboter als moral agents nachvollziehbar erscheinen zu lassen, ohne, dass man sich gleich zu schließen gezwungen fühlen müsste, dass artifizielle Systeme per se, in derselben Weise und in demselben Ausmaß zu moralischem Handeln befähigt seien wie Menschen. Wallach und Allen beschreiben mit ihrem Ansatz – auf den in diesem Artikel noch ausführlich eingegangen wird – eine Version der schwachen KI-These.

Im Folgenden konzentriere ich mich auf den zweiten Arbeitsbereich der Roboterethik und damit auf die Frage, inwiefern (einige) artifizielle Systeme als moralische Akteur*innen denkbar sind.

---

21 Wallach/Allen (2009, 14).

22 Hieran schließt sich die Debatte um autonome Fahrassistenzsysteme; vgl. Hevelke/Nida-Rümelin (2015); Maurer et al. (2015); Both/Weber (2014); Hötitsch/May (2014); Hengstenberger (2012); Knoll (2008). – Unlängst wurden in den USA Computer als Autofahrer offiziell anerkannt; vgl. Frankfurter Rundschau 11.02.2016, Nr. 35, 15.

2. Roboter als moralische Handlungssubjekte – starke und schwache KI

Die Debatte um moralische Akteursfähigkeit kann in dem vorliegenden Artikel nicht umfassend nachgezeichnet, sondern nur mit Blick auf aktuelle und den roboterethischen Diskurs prägende Positionen in den Blick genommen werden. Misselhorn definiert Akteursfähigkeit über zwei Kriterien bzw. zwei Dimensionen: Selbst-Veranlassung (self origination) bzw. Autonomie und Handlungsfähigkeit bzw. Handeln nach Gründen.[23] Daneben konzentriere ich mich v.a. auf Wallachs und Allens These funktional äquivalent zuschreibbarer Fähigkeiten.[24]

Autonomie ist für zahlreiche philosophische Ansätzen zur moralischen Akteursfähigkeit artifizieller Systeme zentral. Sie stellt eine der Schlüsselkompetenzen für moralische Handlungsfähigkeit dar, wobei damit zunächst noch gar nicht Willensfreiheit in einem anspruchsvollen metaphysischen Sinne gemeint sein muss. Autonomie kann in einem ersten Schritt – wie z.B. bei John P. Sullins und Patrick Lin[25] – negativ definiert auf die Abwesenheit von äußerem Zwang oder direkter äußerer Kontrolle rekurrieren. Für einige Denker*innen ist negative Autonomie bereits hinreichend, um (einigen) Maschinen (rudimentäre) Freiheit zuschreiben zu können. Andere – wie bspw. Misselhorn – fügen dem ihrem Ansatz zugrunde liegenden Verständnis von Autonomie eine positive Dimension hinzu, in Bezug auf der sie ausführen, dass man nur dann gehaltvoll von Autonomie sprechen kann, wenn die eigenen Handlungen durch interne Faktoren, die einer gewissen Kontrolle des Handlungssubjekts unterliegen, determiniert sind.[26] Wie das in Frage stehen moralische Subjekt zu den handlungsleitenden Gründen oder auch „zuschreibbaren Präferenzen“[27] gelangt, sei es durch Erziehung oder sei es Programmierung, ist für diese Position zunächst zweitrangig. Etwas zugespitzt könnte man Programmierung als eine ‚harte‘ Form der Erziehung verstehen, umgekehrt Erziehung als eine – sehr – ‚weiche‘ Form der Program-

23 Misselhorn (2013).

24 Weitere Ansätze zu artifiziellen Systemen als moralische Akteur*innen finden sich u.a. bei Gunkel (2012); Floridi (2011); Johnson (2011); Anderson/Anderson (2007); Nadeau (2006); Sullins (2006); Stahl (2004); Allen et al. (2000); Versenyi (1974).

25 Lin et al. (2008); Sullins (2006).

26 Autonomie ist dabei nicht gleichbedeutend mit Nicht-Determiniertheit. Im Gegenteil geht es um eine bestimmte Form der Determination, nämlich um Determination durch das fragliche Handlungssubjekt selbst. Diese Position des Kompatibilismus kann hier nicht eigens behandelt werden; vgl. dazu bspw. Pauen (2001, 2008); Wolf (1988); Watson (1975); Taylor (1967); Frankfurt (1961, 1971).

27 Pauen (2008, 48).

mierung.[28] Gemein ist diesem Verständnis artifizieller Autonomiezuschreibung der kompatibilistische Zugang, indem Autonomie als eine Art Programmierung des Selbst durch handlungsleitende Gründe aufgefasst wird. In diesem Sinne sind „[d]eterminierte Handlungen […] selbstbestimmt und frei […], wenn sie durch die Person, oder konkreter: durch die personalen Präferenzen der Person determiniert sind“[29]. Dies träfe ggf. auch auf (einige) artifizielle Systeme zu.

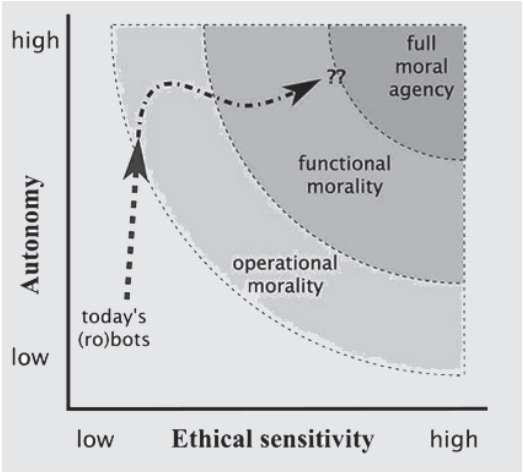
So verstanden ist Autonomie ein graduelles Konzept, da man mehr oder weniger autonom sein kann und damit auch in einem mehr oder weniger ausgeprägten Maß handlungsfähig ist. In einer ersten Annäherung ließe sich festhalten, dass Menschen zwar als genuine moralische Akteur*innen zu begreifen, Roboter allerdings auch moralische Handlungssubjekte sind, wenn auch in einem sehr viel schwächeren Sinn, ähnlich wie wir z.B. auch von Kindern sagen würden, dass sie nicht wie Erwachsene in vollem Maße moralisch handeln können.

Mit Blick auf das gerade angesprochene Handeln nach Gründen (die zweite Bedingung für Akteursfähigkeit) sind insbesondere die moralischen Gründe interessant. Wallach und Allen definieren neben Autonomie Empfänglichkeit bzw. Empfindlichkeit für moralische Werte (sensitivity to morally relevant facts[30]) als Kriterien, um festzustellen, wann eine Maschine als moral agent zu verstehen ist. Beide Kriterien sind ihrer Ansicht nach graduell feststellbar, Akteursfähigkeit ist ein graduelles Konzept. Wallach und Allen differenzieren dabei zwischen „operational morality“, „functional morality“ und „full blown morality“:

28 Dass der Schulterschluss zwischen artifizieller Programmierung und menschlicher Erziehung in der Tat recht nahe liegt, wird weiter unten im Rahmen meiner Ausführungen zu den sog. Bottom-up-Ansätzen deutlich.

29 Pauen (2008, 50).

30 Wallach/Allen (2009, 25).



An dieser Grafik[31] zeigt sich, dass sowohl Autonomie als auch ethische Sensitivität graduelle Konzepte darstellen. Nur Menschen verfügen – so Wallach und Allen – über beide Kompetenzen im genuinen Sinne des Wortes. In der unteren linken Ecke des Koordinatenkreuzes, auf dem Nullpunkt beider Achsen, sind Werkzeuge wie z.B. ein Hammer zu lokalisieren, der weder eigenständig agieren kann noch es für ihn von Bedeutung ist, wessen Daumen er trifft. Mit diesem Beispiel bewegen wir uns Wallach und Allen zufolge noch in dem Bereich außerhalb der Sphäre operationaler Moralzuschreibung.[32] Ein Maschinengewehr mit Kindersicherung verfügt zwar ebenfalls noch nicht über Autonomie, ist allerdings über die Kindersicherung bereits mit einigen wenigen ethischen Implikationen und Wertvorstellungen ausgestattet. Ein Autopilot ist hingegen bereits sehr viel autonomer, aber immer noch relativ schwach ethisch sensitiv. Und das artifizielle System Kismet, am MIT unter der Leitung von Rodney Brooks und Cynthia Breazeal entwickelt, kann in begrenztem Rahmen autonom auf ein menschliches Gegenüber reagieren. Gewehr, Autopilot und Kismet sind Wallach und Allen zufolge Beispiele für eine operational äquivalente Moralzuschreibung, in deren Rahmen die fraglichen artifiziellen Systeme allerdings immer noch „totally within the control of a tool's designers and users"[33] verbleiben. Leider be-

31 Vgl. Wallach/Allen (ebd. 26).

32 Alle hier genannten Beispiele können bei Wallach und Allen (2009) auf den Seiten 25 bis 29 nachgelesen werden.

33 Wallach/Allen (2009, 26) und vgl. (ebd. 30): „Autopilots, decision support systems, and robots with basic capacities to engage in emotion-laden interaction all provide starting points for the field of artificial morality. Systems like these, which are within the domain of operational morality or very limited functional morality, are relatively direct extensions of their designers' values. The designers have to antici-

gründen beide Autoren nicht näher, wann und nach welchen Kriterien die Grenze zum Bereich operationaler Moralzuschreibung überschritten wird (ich komme hierauf weiter unten zurück).

Ein medizinisches Übungsprogramm, ein Supportsystem, das mit Daten und Informationen über den Einsatz von Medikamenten ausgestattet ist und auf der Basis dieser Informationen zu Urteilen über den Umgang mit Patient*innen gelangen kann, verfügt zwar über vergleichsweise wenig Autonomie, aber bereits über sehr starke ethische Sensitivität. Ein solches Programm – bei Wallach und Allen ist beispielhaft das von Susan Leigh Anderson und Michael Anderson entworfene MedEthEex genannt – muss mit zahlreichen ethischen Prinzipien ausgerüstet sein und stellt in ihren Augen deshalb ein Beispiel für funktionale Moralfähigkeit dar, die dadurch definiert ist, nicht mehr vollständig in der Kontrolle der Designer*innen und Nutzer*innen zu liegen, sondern „a platform for further development“[34] zu bereiten. Auch die Grenze zwischen operationaler und funktionaler Moralzuschreibung wird von Wallach und Allen leider nur kurz in den Blick genommen.

Wallachs und Allens Verständnis einer Gradualität bestimmter Kompetenzen und Fähigkeiten beruht auf der Idee funktionaler Äquivalenz:

> „Just as a computer system can represent emotions without having emotions, computer systems may be capable of functioning as if they understand the meaning of symbols without actually having what one would consider to be human understanding.“[35]

Es ist also gar nicht ihr Ziel, artifizielle Systeme in derselben Weise wie Menschen mit den zur Akteursfähigkeit und zu moralischem Handeln notwendigen Vermögen auszurüsten. Wallach und Allen nehmen allein die Zuschreibung funktional äquivalenter Zustände und Verhaltensweisen in den Blick. Funktionale Äquivalenz bedeutet, dass bestimmte beobachtete Phänomene so behandelt werden, ‚als ob‘ sie – mit dem Kantischen Vokabular regulativer Ideen ausgedrückt – bestimmen kognitiven, emotionalen oder anderen zugeschrieben Fähigkeiten entsprechen.[36] Zusätzlich dazu ließe

---

pate most of the circumstances in which their systems will operate, and the actions available on those circumstances are kept within tight limits.“

34 Vgl. Wallach/Allen (ebd.).

35 Vgl. Wallach/Allen (ebd. 69).

36 Dieses Konzept funktionaler Äquivalenz lässt sich mit Gilbert Simondon (2011, 50) bis auf Aristoteles zurückführen: Aristoteles „hat die Idee der Funktion entwickelt, indem er in den verschiedenen Verhaltensweisen des Lebens die Idee der Funktion freigelegt hat, die es erlaubt, durch Parallelismen Entsprechungen herzustellen zwischen Wesen, die sich aufgrund ihrer Struktur und ihrer Existenzweise sehr stark unterscheiden und die […] dennoch als eine Verkettung trotz allem vergleichbarer Funktionen entworfen erscheinen. Über den Begriff der Funktion […] kann man […] bei den Lebewesen funktionale Äquivalente zuschreiben“.

sich zur Erklärung der These funktionaler Äquivalenz Daniel Dennetts Modell dreier Bedeutungsebenen – der physical stance, der design stance und der intentional stance – heranziehen.[37] Auf der intentionalen Beschreibungsebene wird intentionalistisches Vokabular wie bspw. Wünsche und Überzeugungen zur Beschreibung bestimmter Phänomene genutzt, anstatt anzunehmen, dass diese Phänomene (in diesem Fall Fähigkeiten und Kompetenzen) tatsächlich existieren. Wallach und Allen diskutieren in ihrem Buch neben Autonomie und ethischer Sensitivität weitere Vermögen wie bspw. Verstehen (understanding) und Bewusstsein (consciousness):

> „Perhaps the important properties of consciousness are best understood functionally […]. Even if computers won't be conscious in exactly the same way as humans, perhaps they can be designed to function as if they have the relevant similar capacities. […] Functional equivalence of behavior is all that can possibly matter for the practical issue of designing AMAs [artificial moral agents]."[38]

Die Frage, ob künstliche Systeme ‚wirklich' über bestimmte Kompetenzen wie Verstehen, Willensfreiheit, Bewusstsein etc. verfügen, wird zugunsten der Frage aufgegeben, ob diese Fähigkeiten in ihrer von außen phänomenal beobachtbaren Qualität den von uns typischerweise zugeschriebenen Merkmalen dieser Vermögen entspricht. Vor diesem Hintergrund lässt sich ihr Ansatz als eine Version der schwachen KI-These interpretieren, die nach meinem Verständnis den gegenwärtigen roboterethischen Diskurs beherrscht. Im Gegensatz zu der These der starken KI, die davon ausgeht, (irgendwann) eine Maschine erschaffen zu können, die wie der Mensch über Kreativität, Intelligenz usw. verfügt und unter deren Perspektive das artifizielle System dann tatsächlich intelligent zu nennen wäre[39], zielt die These der schwachen KI auf die Simulation all dieser Kompetenzen und nicht auf ihre reale Existenz. Das artifizielle System ‚ahmt' in diesem Fall Intelligenz ‚nach' (abhängig davon, was unter Simulation zu verstehen ist).[40] Stuart Russel und Peter Norvig definieren in ihrem Standardwerk *Artificial Intelligence. A Modern Approach* (2003) die starke und schwache KI-These wie folgt:

> „[T]he assertion that machines could possibly act intelligently (or, perhaps better, act *as if* they were intelligent) is called the **weak AI** hypothesis by philosophers, and the assertion that machines that do

37 Dennett (1996).

38 Wallach/Allen (2009, 67f.).

39 Die starke KI-These wird für gewöhnlich auf Turing (1950) zurückgeführt.

40 Vgl. Searle (1980), dessen berühmtes Chinese-Room-Argument als Antwort auf Turing interpretiert wird.

> so are *actually* thinking (as opposed to *simulating* thinking) is called the **strong AI** hypothesis."[41]

So erscheint es nur konsequent, wenn Wallach und Allen in *Moral Machines* die ontologische und auch die epistemische Frage, ob artifizielle Systeme tatsächlich über die fraglichen Fähigkeiten verfügen und was wir darüber wissen und sicher ausmachen können, zugunsten der praktischen Frage, welche Relevanz ihr Verhalten für uns hat, verwerfen.[42] Einige artifizielle Systeme werden – so ihre Prognose – irgendwann der Grenze zwischen funktionaler, also quasi, und menschlicher, also genuiner, Moralzuschreibung recht nahe kommen, wenn auch voraussichtlich nicht überschreiten.

3. Moral implementieren – drei Ansätze

Unsere Intuitionen über das Vorhandensein bestimmter Kompetenzen wie bspw. die Willensfreiheit fußen häufig auf einem metaphysischen Fundament. Denn in der Tat lässt sich nicht mit Eindeutigkeit feststellen, ob Menschen ‚wirklich' mit Willensfreiheit und ähnlichen Fähigkeiten ausgestattet sind. Wir können sie empirisch nicht eindeutig belegen, nicht ‚beweisen'. Bei Menschen sind wir zwar bereit, eine Zusatzannahme zu treffen, dass immerhin die ‚uns' hinreichend ähnlichen Wesen über die fraglichen Kompetenzen tatsächlich verfügen. Im eigentlichen Sinne trifft die Perspektive der funktionalen Äquivalenz bereits auf Menschen zu, spätestens aber auf Tiere.[43] Im Folgenden werde ich mit Blick auf die im ersten Abschnitt genannten drei Arbeitsbereiche innerhalb der Roboterethik zu artifiziellen Systemen als moral agents – Freiheit, Kognition und Emotionen – erläutern, inwiefern von einer Programmierung oder Implementierung von Fähigkeiten und Vermögen die Rede sein kann. Es lassen sich generell drei Vorgehensweisen differenzieren, die Roboter mit Moralität ausstatten: Top-down-Ansätze, Bottom-up-Ansätze und hybride Ansätze.[44]

Im Rahmen der Top-down-Ansätze werden eine Reihe ethischer Prinzipien oder Regeln, nach denen sich das artifizielle System in einer fraglichen Situation richtet, bspw. die drei (bzw. vier) Asimovschen Robotergesetze[45]

41 Russel/Norvig (2003, 947).

42 Wallach/Allen (2009, 55).

43 Vgl. Nagel (1974).

44 Vgl. Wallach/Allen (2009).

45 Ursprünglich lauten diese in der Kurzgeschichte „Runaround" von 1942 wie folgt: „One, a robot may not injure a human being, or, through inaction, allow a human being to come to harm. […] Two, […] a robot must obey the orders given it by human beings except where such orders would conflict with the First Law. […] And three, a robot must protect its own existence as long as such protection does not conflict with the First or Second Laws" (Asimov 1982, 269–270).

oder die Zehn Gebote fest einprogrammiert. Hierbei ist jedoch mit mindestens zwei Schwierigkeiten zu rechnen: Zum einen sind Regeln oder gar einzelne Begriffe nur (wenn überhaupt) reduziert implementierbar, da ihre Interpretation kontextsensitiv ist und damit in der Festlegung auf eine möglichst eindeutige Deutung programmiert wird (oder auf mehrere Deutungen, nicht aber unter Berücksichtigung aller denkbaren Interpretationen). Das ist die Herausforderung, an der der Roboter in dem Film *I, Robot*[46] schließlich zu scheitern droht, denn es ist alles andere als eindeutig, wie das sogenannte nullte Asimovsche Gesetz zu verstehen ist, welches besagt, dass ein Roboter die Menschheit nicht verletzen oder durch Passivität zulassen darf, dass die Menschheit zu Schaden kommt.[47] Des Weiteren kann ein Konflikt zwischen den einzelnen Regeln entstehen. Legt man einen monistischen Ansatz zugrunde, wird nur eine einzelne Regel wie bspw. Kants Kategorischer Imperativ programmiert, aus der alle konkreten Handlungsanweisungen situativ abzuleiten sind. Ein solcher monistischer Ansatz nimmt an, dass es keine moralischen Dilemma-Situationen gibt, da die Grundregel so formal ist, dass sie für alle Situationen eine konfliktfreie Antwort geben kann. Praktisch besteht aufgrund des Abstraktionsgrades des Moralprinzips die Gefahr, dass das fragliche artifizielle System nichts daraus wird konkret ableiten können.[48] Also: Je konkreter die Formulierung der moralischen Prinzipien, desto eher ist das künstliche System in der Lage, einen Fall in der Praxis unter das Prinzip zu subsumieren; aber je konkreter besagte moralische Prinzipien, desto größer ist auch die Gefahr des Regelkonflikts. Es fehlt bei einem reinen Top-down-Ansatz das, was manch eine*r gerne ‚gesunden Menschenverstand' nennt, auch mit Kreativität und Phänomensensibilität im Umgang mit einer Welt, die wir nicht vollständig berechnen und in Regeln übersetzen können, zum Ausdruck zu bringen.

Die zweite Vorgehensweise in der Implementierung oder Programmierung von Moral stellen Bottom-up-Ansätze dar, die auf der Grundlage von Lern- und evolutionären Algorithmen basieren (auch randomisierte, stochastische oder probabilistische Algorithmen genannt). Hierbei handelt es sich um nichtdeterministische Algorithmen, bei denen nicht reproduzierbare und undefinierte Zustände auftreten. Im Gegensatz zu deterministischen Algorithmen gelangt man hierbei in einem begrenzten Wahrscheinlichkeitsrahmen zu nicht programmierten Zuständen (hierzu weiter unten mehr). Klaus Mainzer arbeitet bspw. zu dynamischen Systemen, in denen durch die komplexe

46 Ein Film von Proyas (2004); basiert auf Asimovs Buch *Ich, der Robot*, von 1950.

47 Um das Asimov in seinem Roman *Aufbruch zu den Sternen* die drei ersten Robotergesetze ergänzt.

48 Das beinhaltet schon einen der Hegelschen Einwände gegen Kants Ethik, den Hegel in der *Phänomenologie des Geistes* anführt (Werke Band 3, 448).

Wechselwirkung der Elemente neue Eigenschaften des Gesamtsystems erzeugt werden, die nicht auf die einzelnen Elemente zurückzuführen sind (Emergenz).[49] Grundsätzlich werden nicht von vornherein moralische Regeln bzw. Sets an Regeln vorgegeben, sondern es ist das Ziel, lediglich basale Parameter zu formulieren bzw. basale Fähigkeiten und Kompetenzen zu implementieren, mithilfe derer artifizielle Systeme selbständig – bspw. durch Trial and Error[50] – moralisches Verhalten entwickeln. Bei den Bottom-up-Ansätzen unterscheidet man Evolutionsmodelle[51] von Modellen menschlicher Sozialisation.[52] Erstere simulieren quasi moralisches Lernen evolutionär, indem in einem künstlichen System voneinander leicht unterschiedene Programme einen ethischen Fall zu evaluieren haben. Diejenigen Programme, die ihn zufriedenstellend lösen, kommen in die ‚nächste Runde', in der sie miteinander rekombiniert weitere ethische Fälle lösen. Evolutionäre Ansätze können noch vor dem Einsatz von Modellen menschlicher Sozialisation in sehr viel früheren Stadien der Moralentwicklung in artifiziellen Systemen zum Einsatz kommen.

Modelle menschlicher Sozialisation berücksichtigen die Rolle von Empathie und Emotionen für moralisches Lernen. Einmal vorausgesetzt, dass moralisches Lernen in einem ganz fundamentalen Sinne über Mitgefühl und Empathie funktioniert,[53] ist zwischen zwei Formen von Mitgefühl zu differenzieren:[54] zwischen perzeptueller Empathie, die bereits dann gegeben ist,

---

49 Mainzer (2010).

50 Trial and Error stellt nur eine Facette des Lernens dar. Zur Lernfähigkeit sind bspw. auch Imitation, Induktion und Deduktion, Exploration, Lernen über Belohnung, Assoziation und Konditionierung zu zählen; vgl. hierzu v.a. Cangelosi/Schlesinger (2015) und auch Bekey (2005).

51 Vgl. Froese/Di Paolo (2010); insbesondere in der Sozialkognitionsforschung.

52 Vgl. Fong et al. (2003); Breazeal/Scassellati (2002).

53 Slote (2007); um an dieser Stelle schon mal einen kurzen Blick auf den dritten Bereich (neben Autonomie und Kognition) innerhalb der Roboterethik, die sich mit artifiziellen Systemen als potenzielle moral agents befasst, zu werfen. Auf die anderen beiden komme ich weiter unten noch zu sprechen. Die Rolle von Emotionalität und Empathie für Moral ist in der philosophischen Reflexion ja in der Tat umstritten. Kant zufolge haben Moral und Emotionen nichts oder vergleichsweise wenig miteinander zu tun. Moral wird im Rahmen des deontologischen Denkens eher mit Urteilskraft, Reflexion, Rationalität und Vernunft assoziiert. Eine andere Position in der Tradition von Blaise Pascal, Adam Smith und David Hume nimmt an, dass Emotionen nicht vollständig durch moralische Reflexion kontrolliert werden, dass hingegen unsere Emotionen unsere moralischen Urteile stark beeinflussen. Der aristotelische Mittelweg vermutet ein wechselseitiges Beeinflussungsverhältnis von moralischen Urteilen auf der einen und Emotionen auf der anderen Seite und spricht Emotionen sowohl negative wie positive Auswirkungen auf unsere moralischen Urteile zu.

54 Stüber (2006).

wenn eine beobachtete Emotion bei mir eine vergleichbare oder kongruente Reaktion bei meinem Gegenüber auslöst[55] und imaginativer Empathie, die einen Perspektivwechsel in Form eines Sich-Hineinversetzen in das Gegenüber erfordert. Perzeptuelle Empathie wird mithilfe bestimmter „Theories of Mind" oder aber auch über neuronale Resonanz und das Wirken von Spiegelneuronen erklärt und ließ sich bereits rudimentär in artifiziellen Systemen hervorrufen.[56] Über diese grundlegende Form des Mitgefühls als Wurzel von prosozialem Verhalten verfügen bereits kleine Kinder und bspw. auch Schimpansen.[57]

Die zweite und deutlich komplexere Form des Mitgefühls ist die imaginative Empathie, die sich auf der Grundlage der perzeptuellen Empathie entwickelt und bislang nur in der menschlichen Sozialisation entsteht, nicht aber mehr bei Primaten. Sie ist kognitiv anspruchsvoller und in komplexere Formen moralischen Urteilens und Handelns involviert.[58] Eine Möglichkeit, über moralisches Lernen bei und über die Implementierung von Moralfähigkeit in Robotern nachzudenken, liegt darin, Emotionen oder zumindest perzeptuelle Empathie in der oben skizzierten basalen Form eines Affektprogrammes,[59] als automatisiertes Reaktionsschema, auch Robotern zuzuschreiben – als zu Emotionen äquivalenten Zuständen.

Geht es bei den Top-down-Ansätzen also im Grunde um die Implementation und Anwendung a priori festgelegter moralischer Regelsets, wird bei den Bottom-up-Ansätzen generell die Möglichkeit moralischen Lernens in den Blick genommen. Sie beruhen auf einer meta-ethischen Annahme über die Kontextsensitivität von Moral, die bei Top-down-Ansätzen gerade fehlt. Moralisches Handeln und Entscheiden bedarf der Erfahrung und eines situativen Urteilsvermögens. Beides kann sich ein artifizielles System nur verkörpert aneignen. In den 1990er Jahren war es u.a. Brooks, der als einer der ersten das Zusammenwirken von artifiziellem System und Umwelt als Bedingung für die Entwicklung von Vermögen und Fähigkeiten betrachtete und von dieser Annahme ausgehend das Feld der „behavior-based robotics" begründete.[60] Zahlreiche berühmte Beispiele der gegenwärtigen Robotik und KI-Forschung, die sich an dem Ansatz verkörperten menschlichen Lernens orientieren – wie bspw. die Lernplattformen iCub, Myon, Cb^2, Curi, Roboy[61]

55 Misselhorn (2009a, 2009b).

56 Balconi/Bortolotti (2012); Rizzolatti/Siniglia (2008); Mataric (2000).

57 Warneken/Tomasello (2009); Hoffmann (2000).

58 Gallagher (2012).

59 Ekman (1992).

60 Vgl. Brooks et al. (1999); Brooks (1991).

61 iCub ist eine humanoide Plattform, die von dem RobotCub-Consortium durch sieben Universitäten entwickelt wird. Der humanoide Roboter Myon wird unter der

(die im Detail sehr unterschiedlichen evolutionsbasierten Ansätze folgen) –, entwickeln Roboter, die sich ähnlich Kindern mit der Zeit Kompetenzen aneignen, aus denen sie dann in spezifischen Kontexten konkrete Handlungsprinzipien ableiten.

Hybride Ansätze kombinieren Top-down- mit Bottom-up-Ansätzen, indem einerseits zwar ein ethischer Rahmen basaler Werte vorgegeben wird, der dann allerdings durch Lernprozesse an spezifische Kontexte anzupassen ist. Dabei ist die Auswahl der fraglichen Regeln und Prinzipien von dem Einsatzbereich des artifiziellen Systems abhängig. Um jedoch von einem hybriden Modell überhaupt sprechen zu können, muss der fragliche Roboter in einem adaptiven Spielraum agieren können, innerhalb dessen er auf die Wertvorstellungen seiner Nutzer*innen ggf. kontextsensitiv reagiert.[62]

Von einem solchen Spielraum spricht, wenn auch noch nicht in einem spezifisch roboterethischen Diskurskontext, bereits Georges Canguilhem in seinem Text „Maschine und Organismus“ (1952), wenn er die unterschiedlichen potenziellen „Freiheitsgrade“ eines „Mechanismus“ expliziert.[63] Je mehr Spielraum, je mehr Freiheitsgrade oder Handlungspotenzial ein Mechanismus aufweist, desto weniger Teleologie im Sinne einer Finalität liegt Canguilhem zufolge vor. „Je begrenzter die Finalität und je enger der Toleranzbereich ist, desto verhärteter und deutlicher erscheint die Finalität.“[64] Übertragen auf hybride Ansätze bedeutet das, dass ein artifizielles System desto mehr Adaptivität und Möglichkeit zur Wertanpassung aufweist, je weniger es an einen spezifischen Zweck gebunden ist (und umgekehrt). So muss bspw. ein komplexer Serviceroboter für Privathäuser, der nicht nur in der Küche unterstützen, den Kindern bei den Hausaufgaben helfen oder im Garten bei der Anlegung der Blumenbeete mit anfassen, sondern ebenso für gelegentliche Fußmassagen und Tipps in der Kombination bestimmter Outfits und Accessoires zu Diensten stehen soll, über einen sehr viel größeren adaptiven Spielraum und damit über eine deutlich geringere Finalität verfügen, als ein vergleichsweise einfacher Roboter, der nur den Tisch zu decken und die Spülmaschine einzuräumen hat (wie vielleicht der oben bereits genannte Curi). Ein solcher komplexer Serviceroboter wäre deshalb unter der Perspektive hybrider Ansätze zu entwickeln, da er zwar aufgrund seines Einsatzbereiches in Privathäusern in einem bestimmten moralischen Rahmen agiert (top-down), in diesem Rahmen allerdings in hohem Grade lern-

---

Leitung von Manfred Hild an der Beuth Hochschule für Technik in Berlin entwickelt. Cb2 entstand an der Universität Osaka in Japan, Curi im Georgia Tech’s Labor und Roboy im Artificial Intelligence Laboratory der Universität Zürich.

62 Misselhorn (2009a, 2009b) entwirft im Rahmen der Altenpflegerobotik gegenwärtig ein solches hybrides System; vgl. auch Misselhorn et al. (2013).

63 Canguilhem (2012, 185f.).

64 Vgl. Canguilhem (ebd. 213).

fähig und flexibel die Anweisungen der Nutzer*innen erst aufnehmen und hernach antizipieren können muss (bottom-up).

Mit Blick auf die im ersten Abschnitt dieses Artikels genannten Arbeitsbereiche einer Roboterethik, die sich mit der Frage nach der potenziellen moralischen Akteursfähigkeit artifizieller Systeme befasst, lässt sich nun darüber nachdenken, inwiefern Freiheit und kognitive Kompetenzen implementierbar sind. Dabei wurde weiter oben bereits einiges über Autonomie und im weitesten Sinne auch über kognitive Kompetenzen wie Intelligenz und Denken gesagt. Die Implementierung von Vermögen knüpft ganz grundsätzlich an Algorithmen bzw. Sets von Algorithmen an, sodass es in der Tat auf den ersten Blick einfacher fallen mag, sich die Programmierung kognitiver Kompetenzen wie Rechnen, Vergleichen und Urteilen funktional äquivalent in künstlichen Systemen vorzustellen als bspw. Kreativität und Sinnhaftigkeit. So verweist auch Hubert L. Dreyfus, der die Entwicklung artifizieller Systeme von ihren Ursprüngen an begleitet und kritisch reflektiert hat, immer noch (zuerst in seinem berühmten Werk *What Computers Can't Do*, 1972) auf die menschliche Kreativität als den kategorialen Unterschied zur Maschine, den diese nicht werden überwinden können.[65] Informationsverarbeitung im Sinne von komplexen Rechenvorgängen könnten sie allerdings, so Dreyfus.

Mit Blick auf unterschiedliche Arten von Algorithmen ließe sich nach meinem Verständnis die durch Wallach und Allen äußerst vage unternommene Unterscheidung zwischen operationaler und funktionaler Moralzuschreibung schärfen und darüber hinaus die Implementierung von Autonomie und kognitiven Kompetenzen mithilfe unterschiedlicher Sets an Algorithmen besser fassen. So könnte man bspw. annehmen, dass alle artifiziellen Systeme, die maßgeblich auf der Grundlage determinierter/deterministischer Algorithmen[66] funktionieren (die bei gleichem Input über dieselbe Abfolge von Zwischenschritten zu denselben Ergebnissen gelangen), keine oder zumindest über eine unspezifisch niedrige Moralfähigkeit verfügen. Eine Lernplattform,

---

65 Bspw. in dem Gespräch, dass Florian Grosser mit Dreyfus geführt hat. Dort sagt Dreyfus (Dreyfus/Grosser 2014, 50): „Was heutigen Computern nach wie vor fehlt, ist der für Menschen charakteristische Common Sense, der es erlaubt, in der ungeordneten Alltagswelt zwischen Wichtigem und Unwichtigem zu unterscheiden. […] Computern [bleibt] – trotz ihrer ungeheuren Rechenkraft und Kapazität, Daten zu schürfen – eine Dimension von Sinn nach wie vor verschlossen. […] Durch solch ‚rohe Gewalt' können diese Computer Sinnhaftigkeit und Bedeutsamkeit nicht tatsächlich verstehen. Nach dem heutigen Stand bleibt die Entwicklung von einer mehr als grob menschenähnlichen künstlichen Intelligenz reine Spekulation."

66 Determinierte Algorithmen gelangen bei gleichem Input immer zu denselben Ergebnissen. Deterministische Algorithmen gelangen bei gleichem Input über dieselben Zwischenschritte zu denselben Ergebnissen. Alle deterministischen Algorithmen sind somit determinierte Algorithmen, nicht aber alle determinierten auch deterministische.

die bereits maßgeblich über determinierte/nicht-deterministische Algorithmen arbeitet (die bei gleichem Input über unterschiedliche Zwischenschritte zu denselben Ergebnissen gelangen), könnte evtl. bereits in Wallachs und Allens Bereich der operationalen Moralzuschreibung lokalisiert werden, was auch ihrer oben gegebenen knappen Definition dieser artifiziellen Systeme entspräche, dass diese immer noch der vollständigen Kontrolle ihrer Designer*innen und Nutzer*innen unterlägen. Roboter schließlich, die maßgeblich auf der Grundlage nicht-determinierter/nicht-deterministischer Algorithmen (die bei gleichem Input über unterschiedliche Zwischenschritte zu unterschiedlichen Ergebnissen gelangen) zu einer gewissen Weiterentwicklung in der Lage wären und über einen vergleichsweise großen (auch hier sind graduelle Differenzen möglich) adaptiven Spielraum verfügten, begegneten wir endlich im Bereich funktionaler Moralzuschreibung. Im Sinne operationaler und insbesondere funktionaler Äquivalenz wären unter diesem Vorgehen (einige) artifizielle Systeme quasi-autonom und quasi-intelligent zu nennen, abhängig von den ihrem Agieren zugrundeliegenden Sets an Algorithmen. Mit Blick auf Empathie-Äquivalenz ließe sich operational äquivalente perzeptuelle Empathie bspw. dem Roboter Kismet zuschreiben, sofern dieser maßgeblich über Sets determinierter/nicht-deterministischer Algorithmen funktioniert. Diesem Gedanken noch weiter folgend, verfügte ein artifizielles System über funktional äquivalente perzeptuelle Empathie, sofern es maßgeblich auf der Grundlage von Sets nicht-determinierter/nicht-deterministischer Algorithmen arbeitete.[67]

Eine generelle Modifikation der implementierten algorithmischen Strukturen (im Sinne einer Kompetenzkompetenz) ist wohl bei jedem artifiziellen System – selbst bei rein nicht-determinierten/nicht-deterministischen Sets an Algorithmen – nicht im selben Ausmaß wie im Rahmen der menschlichen Entwicklung vorstellbar, von der Wünschbarkeit ganz zu schweigen. Hier kommen die Dystopien ins Spiel, in denen Maschinen die Weltherrschaft übernehmen, da sie in der Lage sind, ihre eigenen Parameter völlig ungebunden zu manipulieren. Trotzdem mutet vor dem Hintergrund des gerade Gesagten der Einwand trivial an, dass doch auch im Falle nicht-determinierter/nicht-deterministischer Algorithmen nicht alle vorstellbaren Ergebnisse möglich sind. Denn auch Menschen kann man nicht alle denkbaren Fähigkeiten beibringen; es handelt sich dann ggf. um die trans- und posthumanistische Vision eines Menschen 2.0, der fliegen und durch Wände gehen kann und außerdem unsterblich ist. Die Kritik, dass artifizielle Systeme letztlich immer programmiert sind, trifft nicht, wenn man bedenkt, dass auch

67 Dieses algorithmische Strukturschema stellt nur eine erste grobe Idee da und bedarf einer genauen Ausarbeitung und Diskussion anhand von Beispielen. Das ist für einen späteren Artikel in der Zusammenarbeit mit Informatiker*innen geplant.

Menschen für gewöhnlich nicht zu allem in der Lage, sondern in ihren Möglichkeiten ebenfalls beschränkt bleiben, selbst wenn man ihren adaptiven Spielraum sehr viel größer einschätzt als der eines noch so komplexen Roboters jemals sein könnte.

Fazit und Ausblick zur Zukunft der Roboterethik als Bereichsethik

Anhand der von mir vorgeschlagenen Einteilung der Roboterethik in zwei Arbeitsfelder – Roboter als moral patients und als moral agents – ist ein ganz grundlegender Vergleich zur Tierethik möglich, in der ebenfalls beide Felder eine Rolle spielen. Dabei wurden in diesem Artikel die Möglichkeiten artifizieller moralischer Akteursfähigkeit fokussiert. Nichtsdestotrotz betreffen die Fragen, mit denen wir aktuell konfrontiert sind, fast ausnahmslos den Bereich der Roboterethik, der sich mit artifiziellen Systemen als Wertträger*innen befasst. Zudem scheint das Thema von Empathie und Emotionalität selbst dann für Roboter als moral patients von Belang zu sein, wenn auf dem Feld zu artifiziellen Systemen als moral agents angenommen wird, Moralzuschreibung bedürfe der Empathie nicht (vgl. Fußnote 53).

Vor dem Hintergrund von Wallachs und Allens Ansatz einer funktionalen Äquivalenz graduell vorliegender Kompetenzen und Vermögen, die ich als eine Version der schwachen KI-These interpretiert habe, kombiniert mit meiner Skizze eines algorithmischen Strukturschemas ließe sich nun den Positionen eines Anthro-, Patho-, Bio- und Physiozentrismus eine weitere Sicht zur Lokalisierung von Phänomenen im moralischen Universum hinzufügen, die all die Wesen mit einem Eigenwert bemisst, die lernfähig sind. Lernfähigkeit bedeutet mindestens eine Programmierung durch nicht-determinierte/nicht-deterministische Algorithmen. Solche Wesen befänden sich im oberen Bereich der Wallach-Allen'schen funktionalen Moralzuschreibung und hätten unter dieser Perspektive einen Eigenwert. Weiterhin wäre es möglich, Robotern, die insbesondere auf der Grundlage determinierter/nicht-deterministischer Sets an Algorithmen arbeiten und sich eher im Bereich operationaler Moralzuschreibung bewegen, unter dieser Perspektive immerhin einen hohen instrumentellen Wert zuzuschreiben.

Die Roboterethik als Bereichsethik weist damit zahlreiche Chancen und Perspektiven auf – nicht zuletzt die ernsthafte Ausarbeitung eines Ansatzes, der lernfähigen Wesen einen Eigenwert beimisst sowie der Diskurs um Herausforderungen im Bereich der Roboter als moral patients und das parallele Ringen um die Erschaffung einer starken oder schwachen KI im Bereich der Roboter als moral agents.

Literatur

Allen, C. – Varner, G. – Zinser, J. (2000): Prolegomena to any Future Artificial Moral Agent. S. 251–261 in Journal of Experimental & Theoretical Artificial Intelligence 12 (2000).

Allen, C. – Wallach, W. – Smit, I. (2006): Why Machine Ethics? S. 12–17 in Intelligent Systems IEEE 4 (2006).

Anderson, M. – Anderson, S.L. (2007): Machine Ethics: Creating an Ethical Intelligent Agent. S. 15–26 in AI Magazine 4 (2007).

Anderson, M. – Anderson, S.L. (Hrsg.) (2011): Machine Ethics. Cambridge 2011.

Anderson, M. – Anderson, S.L. – Armen, C. (2006): An Approach to Computing Ethics. S. 2–9 in Intelligent Systems IEEE 4 (2006).

Asaro, P.M. (2006): What Should we Want From a Robot Ethic? S. 9–16 in International Review of Information Ethics 6 (2006).

Asaro, P.M. (2007): Robots and Responsibility from a Legal Perspective. S. 20–24 in Proceedings of the IEEE Conference in Robotics and Automation. Workshop on Roboethics. Rom 2007.

Asimov, I. (1982): The Complete Robot. The Definitive Collection of Robot Stories. London 1982.

Balconi, M. – Bortolotti, A. (2012): Resonance Mechanisms in Empathic Behavior, BEES, BIS/BAS and psychophysiological contribution. S. 298–394 in Physiology and Behavior 105 (2012).

Becker, H. – Rüegsegger, A. (2013): Robotik in Betreuung und Gesundheitsvorsorge. S. 62–64 in Technikfolgenabschätzung. Theorie und Praxis 22 (2013).

Bekey, G.A. (2005): Autonomous Robots. From Biological Inspiration to Implementation and Control. Cambridge 2005.

Boden, M.A. (2006): Could a Robot be Creative – And Would we Know? S. 217–239 in Ford, K.M. – Glymour, C. – Hayes, P.J. (Hrsg.): Thinking About Android Epistemology. Menlo Park – Cambridge – London 2006.

Borenstein, J. – Pearson, Y. (2012): Robot Caregivers: Ethical Issues Across the Human Lifespan. S. 251–265 in Lin, P. – Abney, K. – Bekey, G. (Hrsg.) (2012): Robot Ethics. The Ethical and Social Implications of Robotics. Cambridge, MA – London 2012 .

Both, G. – Weber, J. (2014): Hands-Free Driving? Automatisiertes Fahren und Mensch-Maschine Interaktion. S. 171–187 in Hilgendorf, E. (Hrsg.): Robotik im Kontext von Recht und Moral. Baden-Baden 2014.

Breazeal, C. – Scassellati, B. (2002): Robots That Imitate Humans. S. 481–87 in Trends in Cognitive Sciences 6 (2002).

Brey, P. – Briggle, A. – Waelbers, K. (Hrsg.) (2008): Current Issues on Computing and Philosophy. Amsterdam 2008.

Brooks, R.A. (1991): Intelligence Without Reason. S. 569–595 in Computers and Thought Proceedings of the 12th international joint conference on Artificial intelligence 1 (1991).

Brooks, R.A. – Breazeal, C. – Marjanović, M. – Scasselatti, B. – Williamson, M.M. (1999): The Cog Project: Building a Humanoid Robot. S. 52–87 in Nehaniv, C. (Hrsg.): Computation for Metaphors, Analogy, and Agents. Heidelberg – Berlin 1999.

Cangelosi, A. – Schlesinger, M. (2015): Developmental Robotics. From Babies to Robots. Cambridge 2015.

Canguilhem, G. (2012): Maschine und Organismus. S. 183–232 in Canguilhem, G.: Die Erkenntnis des Lebens. Köln 2012.

Capurro, R. – Nagenborg, M. (Hrsg.) (2009): Ethics and Robotics. Heidelberg – Amsterdam 2009.

Clark, A. (1999): Towards a Cognitive Robotics. S. 5–16 in Adaptive Behavior 7 (1999).

Clark, A. (2003): Artificial Intelligence and the Many Faces of Reason. S. 309–321 in Stich, S.P. – Warfield, T.A. (Hrsg.): The Blackwell Guide to Philosophy of Mind. Malden, MA 2003.

Coeckelbergh, M. (2009): Moral Appearances: Emotions, Robots, and Human Morality. S. 217–221 in International Journal of Social Robotics 1 (2009).

Datteri, E. – Tamburrini, G. (2009): Ethical Reflections on Health Care Robotics. S. 35–47 in Capurro, R. – Nagenborg, M. (Hrsg.): Ethics and Robotics. Heidelberg – Amsterdam 2009.

Dennett, D.C. (1996): The Intentional Stance. Cambridge, MA 1996.

Dennett, D.C. (1998): Brainchildren. Essays on Designing Mind. London 1998.

Dennett, D.C. (2006): Cognitive Wheels: The Frame Problem of AI. S. 147–169 in Ford, K.M. – Glymour, C. – Hayes, P.J. (Hrsg.): Thinking About Android Epistemology. Menlo Park – Cambridge – London 2006.

Dorffner, G. (2004): Rationalität, Emotionalität und Körperlichkeit: Können Maschinen Menschen verstehen – und umgekehrt? S. 102–112 in Schmidinger, H. (Hrsg.): Der Mensch – ein ‚animal rationale'? Vernunft – Kognition – Intelligenz. Darmstadt 2004.

Dreyfus, H. – Grosser, F. (2014): Heißt Denken Rechnen, Herr Dreyfus? S. 50–51 in Philosophie Magazin 6 (2014).

Edgar, S.L. (2003): Morality and Machines. Perspectives on Computer Ethics. Boston – Toronto – London – Singarpore ^{2}2003.

Ekman, P. (1992): An Argument for Basic Emotions. S. 169–200 in Cognition and Emotion 6 (1992).

Floridi, L. (2011): On the Morality of Artificial Agents. S. 184–212 in Anderson, M. – Anderson, S.L. (Hrsg.): Machine Ethics. Cambridge 2011.

Floridi, L. – Sanders, J.W. (2004): On the Morality of Artificial Agents. S. 349–379 in Minds and Machines 14 (2004).

Fong, T. – Nourbakhsh; I. – Dautenhahn, K. (2002): A Survey of Socially Interactive Robots: Concepts, Design, and Applications. S. 2–29 in Technical Report CMU-RI-TR (2002).

Frankfurt, H.G. (1961): Alternate Possibilities and Moral Responsibility. S. 829–839 in The Journal of Philosophy 66 (1961).

Frankfurt, H.G. (1971): Freedom of the Will and the Concept of a Person. S. 5–20 in The Journal of Philosophy 68 (1971).

Friedman, B. – Kahn, P.H. (1992): Human Agency and Responsible Computing: Implications for Computer System Design. S. 7–14 in Journal of Systems and Software 17 (1992).

Froese, T. – Di Paolo, E.A. (2010): Modelling Social Interaction As Perceptual Crossing: An Investigation into the Dynamics of the Interaction Process. S. 43–68 in Connecting Science 22 (2010).

Gallagher, S. (2012): Neurons, Neonates and Narrative: From Embodied Resonance to Empathic Understanding. S. 167–196 in Foolen, A. – Lüdtke, U. – Zlatev, J. – Racine, T. (Hrsg.): Moving Ourselves, Moving Others. Amsterdam 2012.

Gehlen, A. (1997): Der Mensch. Seine Natur und seine Stellung in der Welt. München [13]1997.

Gunkel, D.J. (2012): The Machine Question. Critical Perspectives on AI, Robots, and Ethics. Cambridge, MA – London 2012.

Hengstenberger, M. (2012): Kein Mensch am Steuer? Ungeheuer! Automatisiertes Fahren. Spiegel Online. URL: www.spiegel.de/auto/aktuell/automatisiertes-fahren-2025-fahren-autos-selbststaendiga-873582.html (2012) [Stand: 01.02.2016].

Hevelke, A. – Nida-Rümelin, J. (2015): Selbstfahrende Autos und Trolley-Probleme: Zum Aufrechnen von Menschenleben im Falle unausweichlicher Unfälle. S. 5–24 in Jahrbuch für Wissenschaft und Ethik 19 (2015).

Hilgendorf, E. (Hrsg.) (2014): Robotik im Kontext von Recht und Moral. Baden-Baden 2014.

Hoffman, M.L. (2000): Empathy and Moral Development: Implications for Caring and Justice. Cambridge 2000.

Hötitzsch, S. – May, E. (2014): Rechtliche Problemfelder beim Einsatz automatisierter Systeme im Straßenverkehr. S. 189–210 in Hilgendorf, E. (Hrsg.): Robotik im Kontext von Recht und Moral. Baden-Baden 2014.

Ichbiah, D. (2005): Roboter. Geschichte – Technik – Entwicklung. München 2005.

Johnson, D.G. (2011): Computer Systems. Moral Entities but not Moral Agents. S. 168–183 in Anderson, M. – Anderson, S.L. (Hrsg.) (2011): Machine Ethics. Cambridge 2011.

Knoll, P.M. (2008): Prädikative Fahrassistenzsysteme – Bevormundung des Fahrers oder realer Kundennutzen? S. 159–171 in Hubig, C. – Koslowski, P. (Hrsg.): Maschinen, die unsere Brüder werden. Mensch-Maschine-Interaktion in hybriden Systemen. München 2008 .

Köhler, W. (1921): Intelligenzprüfungen an Anthropoiden. Berlin – Göttingen – Heidelberg 1921 (Neudruck 1962).

Krebs, A. (1997): Naturethik im Überblick. S. 337–379 in Krebs, A. (Hrsg.): Naturethik. Grundtexte der gegenwärtigen tier- und ökoethischen Diskussion. Frankfurt a.M. 1997.

Krings, B.-J. – Böhle, K. – Decker, M. – Nierling, L. – Schneider, C. (Pre-Print 2012): Service-Roboter in Pflegearrangements. Karlsruhe. Online unter URL: http://www.itas.kit.edu/pub/v/2012/epp/krua12-pre01.pdf [Stand: 23.01.2016].

Laudon, K.C. (1995): Ethical Concepts and Information Technology. S. 33–39 in Communications of the ACM 38 (1995).

Levy, D.J. (2012): The Ethics of Robot Prostitutes. S. 223–231 in Lin, P. – Abney, K. – Bekey, G. (Hrsg.) (2012): Robot Ethics. The Ethical and Social Implications of Robotics. Cambridge, MA – London 2012.

Lin, P. – Bekey, G. – Abney, K. (2008): Autonomous Military Robotics: Risks, Ethics, and Design. Prepared for: US Department of Navy, Office of Naval Research. San Luis Obispo 2008.

Lin, P. – Abney, K. – Bekey, G. (2011): Robot Ethics: Mapping the Issues of a Mechanized World. S. 942–949 in Artificial Intelligence 175 (2011).

Lin, P. – Abney, K. – Bekey, G. (Hrsg.) (2012): Robot Ethics. The Ethical and Social Implications of Robotics. Cambridge, MA – London 2012.

Lokhorst, G.-J. – van den Hoven, J. (2012): Responsibility for Military Robots. S. 145–156 in Lin, P. – Abney, K. – Bekey, G. (Hrsg.) (2012): Robot Ethics. The Ethical and Social Implications of Robotics. Cambridge, MA – London 2012.

Mainzer, K. (2010): Leben als Maschine. Von der Systembiologie zur Robotik und Künstlichen Intelligenz. Paderborn 2010.

Maring, M. (2008): Mensch-Maschine-Interaktion. Steuerbarkeit – Verantwortbarkeit. S. 113–129 in Hubig, C. – Koslowski, P. (Hrsg.): Maschinen, die unsere Brüder werden. Mensch-Maschine-Interaktion in hybriden Systemen. München 2008.

Mataric, M. (2000): Getting Humanoids to Move and Imitate. S. 18–24 in IEEE Intelligent Systems 15 (2000).

Maurer, M. – Gerdes, J.C. – Lenz, B. – Winner, H. (Hrsg.) (2015): Autonomes Fahren. Technische, rechtliche und gesellschaftliche Aspekte. Berlin – Heidelberg 2015.

May, E. (2014): Robotik und Arbeitsschutzrecht. S. 99–118 in Hilgendorf, E. (Hrsg.): Robotik im Kontext von Recht und Moral. Baden-Baden 2014.

Misselhorn, C. (2009a): Empathy with Inanimate Objects and the Uncanny Valley. S. 345–59 in Minds and Machines 19 (2009).

Misselhorn, C. (2009b): Empathy and Dyspathy with Androids. Philosophical, Fictional and (Neuro-)Psychological Perspectives. S. 101–123 in Between Nature and Culture – After the Continental-Analytical Divide, Konturen 2 (2009).

Misselhorn, C. (2013): Robots as Moral Agents. S. 30–42 in Roevekamp, F. (Hrsg.): Roboethics. Proceedings of the Annual Conference on Ethics of the German Association for Social Science Research on Japan. München 2013.

Misselhorn, C. – Pompe, U. – Stapleton, M. (2013): Ethical Considerations Regarding the Use of Social Robots in the Fourth Age. S. 121–133 in Geropsych 26 (2013).

Moor, J.H. (2006): The Nature, Importance, and Difficulty of Machine Ethics. S. 18–21 in Intelligent Systems IEEE 4 (2006).

Nadeau, J.E. (2006): Only Androids can be Ethical. S. 241–248 in Ford, K.M. – Glymour, C. – Hayes, P.J. (Hrsg.): Thinking About Android Epistemology. Menlo Park – Cambridge – London 2006.

Nagel, T. (1974): What is it like to be a bat?. S. 435–450 in The Philosophical Review LXXXIII (1974). Online unter URL: http://organizations.utep.edu/portals/1475/nagel_bat.pdf [Stand: 23.01.2016].

Neuhäuser, C. (2014): Roboter und moralische Verantwortung. S. 269–286 in Hilgendorf, E. (Hrsg.): Robotik im Kontext von Recht und Moral. Baden-Baden: 2014 .

Pauen, M. (2001): Freiheit und Verantwortung. Wille, Determinismus und der Begriff der Person. S. 23–44 in Allgemeine Zeitschrift für Philosophie 26 (2001).

Pauen, M. (2008): Freiheit, Schuld und Strafe. S. 41–74 in Lampe, E.-J. – Pauen, M. – Roth, G. (Hrsg.): Willensfreiheit und rechtliche Ordnung. Frankfurt a.M. 2008.

Petersen, S. (2007): The Ethics of Robot Servitude. S. 43–54 in Journal of Experimental & Theoretical Artificial Intelligence 19 (2007).

Petersen, S. (2012): Designing People to Serve. S. 283–298 in Lin, P. – Abney, K. – Bekey, G. (Hrsg.): Robot Ethics. The Ethical and Social Implications of Robotics. Cambridge, MA – London 2012 .

Plessner, H. (1928): Die Stufen des Organischen und der Mensch. Einleitung in die philosophische Anthropologie. Berlin 1928.

Plessner, H. (1946): Mensch und Tier. S. 52–65 in Plessner, H: Gesammelte Schriften in 10 Bänden. Band 8. Conditio Humana. Hrsg. v. G. Dux, O. Marquard, E. Ströker (1980–1985). Frankfurt a.M. 1946.

Rizzolatti, G. – Siniglia, C. (2008): Empathie und Spiegelneurone – Die biologische Basis des Mitgefühls. Frankfurt a.M. 2008.

Russel, S. – Norvig, P. (2003): Artificial Intelligence. A Modern Approach. New Jersey ^{2}2003.

Scheler, M. (1993 [1908–09]): Biologievorlesung. S. 257–361 in Scheler, M.: Gesammelte Werke in 15 Bänden. Band 14. Schriften aus dem Nachlass. Band 5: Varia I. Hrsg. v. M. Frings. Bonn 1993.

Scheutz, M. (2012): The Inherent Danger of Undirectional Emotional Bonds between Humans and Social Robots. S. 205–221 in Lin, P. – Abney, K. – Bekey, G. (Hrsg.): Robot Ethics. The Ethical and Social Implications of Robotics. Cambridge, MA – London 2012.

Searle, J.R. (1980): Minds, brains and programs. S. 417–157 in Behavioral and Brain Sciences 3 (1980). Online unter URL: http://cogprints.org/7150/1/10.1.1.83.5248.pdf [Stand: 23.01.2016].

Sharkey, N. – Sahrkey, A. (2012): The Rights and Wrongs of Robot Care. S. 267–282 in Lin, P. – Abney, K. – Bekey, G. (Hrsg.) (2012): Robot Ethics. The Ethical and Social Implications of Robotics. Cambridge, MA – London 2012.

Simondon, G. (2011): Tier und Mensch. Zwei Vorlesungen. Zürich 2011.

Slote, M. (2007): The Ethics of Care and Empathy. New York 2007.

Sombetzki, J. (2014): Verantwortung als Begriff, Fähigkeit, Aufgabe. Eine Drei-Ebenen-Analyse. Wiesbaden 2014.

Sombetzki, J. (2016): Philosophical Anthropology Between Human and Machine: Towards a Gradual-Human-Machine-Relationship. In Thompson, S. (Hrsg.): Handbook of Research on Androids, Cyborgs, and Robots in Contemporary Culture and Society. IGI Global. 2016 (im Erscheinen).

Stahl, B.C. (2004): Information, Ethics, and Computers: The Problem of Autonomous Moral Agents. S. 67–83 in Minds and Machines 14 (2004).

Sterrett, S.G. (2006): Too Many Instincts: Contrasting Philosophical Views on Intelligence in Humans and Nonhumans. S. 187–215 in Ford, K.M. – Glymour, C. – Hayes, P.J. (Hrsg.): Thinking About Android Epistemology. Menlo Park – Cambridge – London 2006.

Stüber, K. (2006): Rediscovering Empathy: Agency, Folk Psychology, and the Human Sciences. Cambridge, MA 2006.

Sullins, J.P. (2006): When is a Robot a Moral Agent? S. 23–30 in International Review of Information Ethics 12 (2006).

Taylor, C. (1976): Responsibility for Self. S. 281–299 in Rorty, A.O. (Hrsg.): The Identities of Persons. Berkeley, CA 1976.

Turing, A. (1950): Computing Machinery and Intelligence. S. 433–460 in Mind 59 (1950). Online unter URL: http://www.loebner.net/Prizef/TuringArticle.html [Stand: 23.01.2016].

Versenyi, L. (1974): Can Robots be Moral? S. 248–259 in Ethics 84 (1974).

Wallach, W. – Allen, C. (2009): Moral Machines. Teaching Robots Right from Wrong. Oxford – New York 2009.

Warneken, F. – Tomasello, M. (2009): Varieties of Altruism in Children and Chimpanzees. S. 397 in Trends in Cognitive Sciences, 13 (2009).

Watson, Gary (1975): Free Agency. S. 205–220 in The Journal of Philosophy 72 (1975).

Whitby, B. (2012): Do You Want a Robot Lover? The Ethics of Caring Technologies. S. 233–247 in Lin, P. – Abney, K. – Bekey, G. (Hrsg.) (2012): Robot Ethics. The Ethical and Social Implications of Robotics. Cambridge, MA – London 2012.

Wolf, S. (1988): Sanity and the Metaphysics of Responsibility. S. 46–62 in Schoeman, F.D. (Hrsg.): Responsibility, Character, and the Emotions. New Essays in Moral Psychology. Cambridge 1988.