Roboterethik

Janina Loh über eine noch junge Bereichsethik

Was ist Roboterethik?

Die Roboterethik ist eine noch junge Bereichsethik. Ihr wird immer wieder vorgeworfen, sie habe keinen spezifischen Gegenstand, da sich Ethik nicht mit Unbelebtem beschäftige (1). Doch selbst wenn sich herausstellen sollte – was zu untersuchen ist –, dass Roboter selbst keine moralischen Handlungssubjekte sein können, ist ihnen ein Platz im moralischen Universum zuzuweisen. Schließlich sind wir gewillt, einer ganzen Reihe von Entitäten einen Wert zuzusprechen – beispielsweise Landschaften, Ökosystemen, Tieren, aber auch Häusern, Autos oder Smartphones. 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 wäre der einer solchen Diskussion angemessene Raum?

In der Roboterethik wird – vergleichbar der Tierethik – darüber nachgedacht, inwiefern Maschinen Wertträger sind und diskutiert, inwiefern sie als (rudimentäre) moralische Akteure gelten können (2). Sie stellt traditionelle Fragen mit Blick auf neue potenzielle Handlungssubjekte – Roboter – (3), wie etwa: Welche Kompetenzen erachten wir als grundlegend für moralische Akteursfähigkeit? Welche moralische (und andere) Werte sollten wir artifiziellen Systemen implementieren? Auf was für ein moralisches Selbstverständnis lässt es schließen, wenn wir Roboter ‚schlecht' behandeln (4)? In welchen Bereichen – Industrie-, Militär-, Medizin-, Altenpflege-, Servicerobotik – wollen wir uns auch zukünftig ausschließlich bzw. in einem signifikanten Ausmaß auf menschliche und nicht auf artifizielle Expertise verlassen?

Der Begriff „Roboter" geht auf das tschechische Wort „robota" (Arbeit, Frondienst, Zwangsarbeit) zurück und wurde 1920 von dem Künstler Josef Čapek geprägt. Sein Bruder Karel Čapek sprach in dem Theaterstück Rossum's Universal Robots (1921) von „labori" für humanoide Apparaturen, die dem Menschen Arbeit abnehmen. Der Stuttgarter Philosophin Catrin Misselhorn zufolge ist ein Roboter eine elektro-mechanische Maschine, bestehend aus einer Einwicklungseinheit (einem Prozessor), Sensoren, die Informationen über die Welt sammeln sowie einem Effektor oder Aktor, der Signale in mechanische Abläufe übersetzt. Das Verhalten eines Roboters ist oder wirkt zumindest autonom – er kann, anders als ein Computer, in seine Umgebung hinein wirken und auf sie Einfluss nehmen (5).

Die zwei Arbeitsfelder der Roboterethik

In der Roboterethik unterscheidet man zwei Bereiche. Im einen wird diskutiert, inwiefern Roboter als *moral patients* zu verstehen sind, also passiv als Träger moralischer Rechte bzw. inwiefern ihnen ein moralischer Wert zukommt. Im anderen Feld geht es um die Frage, inwiefern Roboter *moral agents*, also aktiv Träger moralischer Pflichten bzw. moralische Handlungssubjekte, sein können (6). Beide Arbeitsbereiche ergänzen einander.

Die Gruppe der *moral agents* ist gegenüber der der *moral patients* exklusiver; für gewöhnlich zeichnen wir nur Menschen mit Moralfähigkeit im genuinen Sinne aus. Einer ganzen Reihe von Wesen und Dingen wird indes ein moralischer Wert zugeschrieben – zumindest insofern, als diese Entitäten moralisch bedenkenswert sind, wenn ihnen vielleicht auch kein Eigen-, sondern nur ein hoher instrumenteller Wert zuzusprechen ist. Als moralisches Handlungssubjekt ist man zugleich Wertträger – dies gilt allerdings nicht umgekehrt. Die Zuschreibung von moralischen Werten zu Lebewesen und Gegen-

ständen ist abhängig von der jeweils eingenommenen Perspektive. Eine anthropozentrische Position argumentiert beispielsweise dafür, dass nur dem Menschen ein Eigenwert zukommt (7). Der Einbezug von artifiziellen Systemen in den Horizont der mit einem Eigenwert ausgestatteten Dinge könnte eine weitere Perspektive eröffnen, einen „Mathenozentrismus“ (von griech. „matheno“, „lernen“), der all das mit einem Eigenwert bemisst, das in einer spezifischen (noch zu erörternden) Weise gesteuert oder programmiert bzw. lernfähig ist.

Im ersten Arbeitsbereich geht es darum, wie mit artifiziellen Systemen, wie mit Robotern umzugehen ist, inwiefern ihnen ein Wert zukommt, selbst wenn man sich darüber einig sein sollte, dass sie selbst nicht zu moralischem Handeln in der Lage sind. Man versteht hier artifizielle Systeme als Werkzeuge oder gar als Ergänzungen des Menschen und arbeitet an Themen wie beispielsweise der Formulierung von Ethikkodizes in Unternehmen (8), der Möglichkeit und Wünschbarkeit von Beziehungen zu und mit Robotern (9), der „Versklavung“ von Robotern (10) oder der Beurteilung des Einsatzes von artifiziellen Systemen zu Therapiezwecken (11). Dabei bleibt die moralische Kompetenz bei den Menschen: Sie entscheiden über die Moral ihrer Geschöpfe und darüber, wer im Falle eines Unfalls Verantwortung trägt.

Innerhalb des zweiten Arbeitsfelds, in dem Roboter als moral agents betrachtet werden, wird danach gefragt, inwiefern Roboter zu moralischem Handeln fähig sind und über welche Kompetenzen sie hierfür in welchem Maße verfügen müssen. Dabei geht es um die Zuschreibung von Freiheit als Bedingung für moralisches Handeln, um die dafür notwendigen kognitiven Kompetenzen (Denken, Verstehen, Geist, Intelligenz, Bewusstsein, Wahrnehmung und Kommunikation) aber auch um Empathie und Emotionen.

Beiden Arbeitsfeldern liegt die Frage zugrunde, was Moral bzw. was Ethik ist und wie moralische Urteile gefällt werden. Grundlegend für die Diskussion ist das im Jahr 2009 erschienene Werk Moral Machines. Teaching Robots Right from Wrong, von Wendell Wallach (Yale University) gemeinsam mit Colin Allen (Indiana University) verfasst wurde. Sie schlagen vor, allen Wesen Moralfähigkeit zuzuschreiben, die in Situationen geraten, in denen moralische Entscheidungen zu treffen sind.

Sie schließen dabei an das von Philippa Foot stammende Gedankenexperiment der „Trolley Cases“ an: Eine Straßenbahn droht, fünf Personen zu überrollen, kann aber durch Umstellen einer Weiche auf ein anderes Gleis umgeleitet werden. Unglücklicherweise befindet sich dort eine weitere Person. Darf (durch Umlegen der Weiche) ihr der Tod in Kauf genommen werden, um das Leben der anderen zu retten? (12). Das hat in der Gegenwart eine Debatte um autonome Fahrassistenzsysteme aufgeworfen (13).

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, unverzüglich zu stoppen, wenn sich Menschen auf den Gleisen aufhalten. Es kann zunächst keine Rede davon sein, dass ein fahrerloser Zug (oder ein anderes autonomes Fahrassistenzsystem – z. B. ein autonomer Krankentransport), ausgerüstet mit einer spezifischen algorithmischen Struktur, im genuinen Sinne des Wortes moralisch handelt. Allerdings ist diese Situation äußerlich einer solchen ähnlich, in der sich auch ein Mensch befinden könnte. In ihrer von außen beobachtbaren phänomenalen 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 in derselben Weise und in demselben Ausmaß wie Menschen zu moralischem Handeln befähigt seien.

Roboter als Handlungssubjekte

Catrin Misselhorn definiert Akteursfähigkeit über Selbst-Veranlassung (Autonomie) und Handlungsfähigkeit (Handeln nach Gründen, 14). Autonomie ist für zahlreiche philosophische Ansätzen zur moralischen Akteursfähigkeit artifizieller Systeme zentral, wobei damit zunächst noch gar nicht Willensfreiheit in einem anspruchsvollen metaphysischen Sinne gemeint sein muss. Autonomie kann auch negativ definiert auf die Abwesenheit von äußerem Zwang oder direkter äußerer Kontrolle rekurrieren, was für einige bereits hinreichend ist, um (einigen) Maschinen (rudimentäre) Freiheit zuschreiben zu können (15). Für andere – wie etwa Misselhorn – kann von Autonomie nur dann gehaltvoll die Rede sein, wenn die eigenen Handlungen durch interne Faktoren, die einer gewissen Kontrolle des Handlungssubjekts unterliegen, determiniert sind. Autonomie ist dabei nicht gleichbedeutend mit Nicht-Determiniertheit. Im Gegenteil – es geht um eine bestimmte Form der Determination, nämlich um Determination durch das fragliche Handlungssubjekt selbst.

Wie das in Frage stehende moralische Subjekt zu den handlungsleitenden Gründen gelangt, (etwa durch Erziehung oder Programmierung), ist für diese Position zweitrangig. Etwas zugespitzt könnte man Programmierung als eine ‚harte' Form der Erziehung deuten, umgekehrt Erziehung als eine sehr ‚weiche' Form der Programmierung. 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. In einer ersten Annäherung lassen sich Menschen als genuine moralische Akteure begreifen und Roboter auch, allerdings in einem sehr viel schwächeren Sinn.

Hinsichtlich des Handelns nach Gründen (der zweiten Bedingung für Akteursfähigkeit) sind insbesondere die moralischen Gründe interessant. Wendel Wallach und Colin Allen haben sich die Frage gestellt, inwiefern Roboter als artifizielle moralische Akteure zu verstehen sind. Sie definieren *moral agency* als graduelles Konzept, das zwei Bedingungen genügen muss, nämlich Autonomie und Empfänglichkeit bzw. Empfindlichkeit für moralische Werte *(sensitivity to values*) (16). Menschen gelten für sie als moralische Akteure im genuinen Sinne; allerdings sehen sie in einigen Maschinen – beispielsweise einem Autopiloten oder dem artifiziellen System Kismet (das seine Ohren, Augen, Lippen sowie seinen Kopf bewegen und auf externe Stimuli wie die menschliche Stimme reagieren kann) – operationale moralische Akteure. Und dennoch verbleiben sie immer noch „totally within the control of [the] tool's designers and users" (a.a.O. S. 26). In diesem Sinne sind operationale moralische Akteure „direct extensions of their designers' values" (ebd. S. 30).

Für eine weitergehende Autonomie bzw. moralische Sensitivität führen Wallach & Allen den Begriff der „funktionalen Moralität" ein. Funktionale Moralität bedeutet, dass das fragliche artifizielle System insofern entweder autonomer und/oder werte-sensitiver ist als ein operationaler moralischer artifizieller Akteur, als funktionale moralische Maschinen „themselves have the capacity for assessing and responding to moral challenges" (Wallach & Allen 2009: 9). Nur besonderen artifiziellen Systeme kommt der Status funktionaler moralischer Akteursfähigkeit zu – etwa dem medizinisch-ethischen Expertensystem MedEthEx, dem die Prinzipien der biomedizinischen Ethik von Beauchamp und Childress implementiert sind (17).

Dieser Ansatz der funktionalen Moralität, der graduellen Zuschreibung von Kompetenzen und Fähigkeiten, gründet auf dem Gedanken der funktionalen 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" (ebd.: 69). Funktionale Äquivalenz bedeutet, dass spezifische Phänomene verstanden werden, „als ob" sie kognitiven, emotionalen oder anderen Kompetenzen und

Janina Loh

Fähigkeiten entsprechen.

Starke und schwache KI

Der Ansatz funktionaler Äquivalenz beruht auf der Unterscheidung zwischen starker und schwacher Künstlicher Intelligenz. Starke KI meint Maschinen, die im genuinen Sinne des Wortes mit Intelligenz, Bewusstsein und Autonomie ausgerüstet sind. Schwache KI (18) ist bescheidener: Ihr ist lediglich an der Simulation spezifischer Kompetenzen in artifiziellen Systemen gelegen. 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 so are *actually* thinking (as opposed to *simulating* thinking) is called the strong AI hypothesis“ (19, p. 947).

Wallach und Allen verzichten auf die Annahme einer starken KI und der daran geknüpften Kompetenzen hinsichtlich artifizieller moralischer Akteure. Vielmehr fokussieren sie sich auf die Zuschreibung von funktional äquivalenten Bedingungen und Verhaltensweisen. Die Frage, inwiefern artifizielle Systeme irgendwann intelligent, bewusst oder autonom im Sinne der starken KI-These genannt werden können, wird durch die Frage ersetzt, in welchem Ausmaß und Umfang die fraglichen Kompetenzen der Funktion entsprechen, die sie innerhalb der moralischen Evaluation spielen.

Wallach und Allen denken sich den Übergang von operationaler über funktionale bis hin zu voller Moralzuschreibung abhängig von den genannten Kompetenzen Autonomie und moralische Sensitivität graduell. Es ist jedoch schwer vorstellbar, wie ein artifizielles System ein funktionales Äquivalent zu der menschlichen Fähigkeit, höherstufige Wünsche (die „*second-order volitions*“ Frankfurts, 20) bilden zu können, entwickeln könnte. Hilfreich erscheint hier Darwalls Unterscheidung zwischen vier Formen von Autonomie: „personal“, „moral“, „rational“ und „agential“ Autonomie (21). Während persönliche Autonomie die Fähigkeit umfasst, Werte, Ziele und letzte Zwecke zu definieren, beinhaltet moralische Autonomie die Möglichkeit, selbst gesetzte Prinzipien und ethische Überzeugungen zu reflektieren. Diese beiden Formen von Autonomie werden wohl noch für lange Zeit menschlichen Akteuren vorbehalten bleiben, hingegen sieht Darwall rationale Autonomie prima facie auch für artifizielle Akteure erreichbar. Rationale Autonomie gründet auf „weightiest reasons“ (Darwall), die funktional äquivalent etwa in Form von Algorithmen repräsentiert werden können. Erst recht scheint „agential autonomy“ als ein spezifisches Verhalten, das nicht vollständig durch externe Faktoren bestimmt ist, Maschinen zuschreibbar. „Agential autonomy“ kann funktional äquivalent durch die Fähigkeit simuliert werden, interne Zustände eines artifiziellen Systems ohne externe Stimuli zu ändern.

Zur Erklärung der These funktionaler Äquivalenz lässt sich Daniel Dennetts Modell dreier Bedeutungsebenen – der „physical stance“, der „design stance“ und der „intentional stance“ – heranziehen (22). Auf der intentionalen Beschreibungsebene wird intentionalistisches Vokabular wie beispielsweise Wünsche und Überzeugungen zur Beschreibung von Phänomenen genutzt, ohne anzunehmen, dass diese Phänomene (in diesem

Fall Wünsche zu haben und Überzeugungen auszubilden) tatsächlich existieren.

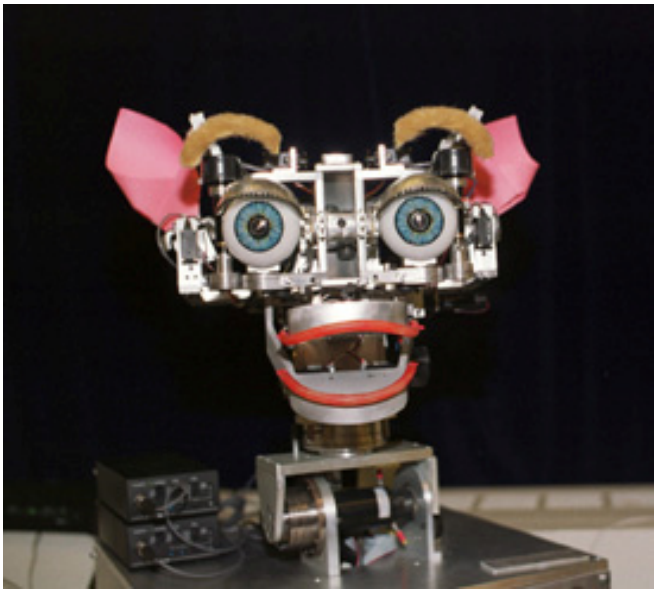
Wallach und Allen sehen die operational artifiziellen Systeme vollständig in der Kontrolle der DesignerInnen und NutzerInnen. Funktional moralfähige Roboter sind jedoch in einem gewissen – und sei es auch nur geringen – Maße lernfähig. Eine Grenzziehung zwischen nicht-moralischen Werkzeugen, operational nicht lernfähigen und funktional lernfähigen Maschinen lässt sich auf der computationalen Ebene in Form eines algorithmischen Strukturschemas vorzunehmen, indem man sich die Unterscheidung zwischen determinierten und deterministischen Algorithmen zunutze macht (23): Während *deterministische Algorithmen* bei gleichem Input zu demselben Output gelangen, gelangen *determinierte Algorithmen* bei gleichem Input ebenfalls zum selben Output, weisen allerdings bei der Wahl der Zwischenschritte, die dahin führen, einen gewissen Spielraum auf. Es ist vorstellbar, Maschinen, die auf der Grundlage deterministischer Algorithmen funktionieren – also gewissermaßen „determinierter" sind als nur determinierte Algorithmen –, weder in der funktionalen noch in der operationalen Sphäre zu verorten. Man könnte sie immer noch als Maschinen sehen, allerdings fast den nicht-mechanischen Werkzeugen näher als der operationalen Sphäre. Denn im Rahmen rein deterministischer algorithmischer Strukturen gibt es keinerlei Spielraum – von artifizieller Lernfähigkeit ganz zu schweigen. Die Sphäre operationaler Moralfähigkeit würde dann mit den artifiziellen Systemen betreten, die vornehmlich durch determinierte (aber nicht-deterministische) Algorithmen strukturiert sind. Und die Fälle artifizieller Systeme, die vornehmlich auf der Grundlage nicht-determinierter (und also nicht-deterministischer) Algorithmen operieren, könnten in der funktionalen Sphäre lokalisiert werden, denn diese verfügen sowohl was die Zwischenschritte als auch was das Ergebnis anbelangt, über einen größeren Spielraum. Hier ließe sich auch von artifizieller Lernfähigkeit sprechen.

Zwei Beispiele

Das artifizielle System Kismet kann seine Ohren, Augen, Lippen sowie seinen Kopf bewegen und reagiert auf externe Stimuli wie die menschliche Stimme. Wallach und Allen interpretieren dieses System als operationalen moralischen Akteur und sprechen ihm in einem äußert rudimentären Sinne ethische Sensitivität zu. Die größte Herausforderung, Kismet als einen operational moralischen Akteur zu verstehen, liegt in der Zuschreibung von Autonomie, steht er doch immer noch vollständig unter der Kontrolle seiner Nutzer und funktioniert wohl maßgeblich auf determinierten (aber nicht-deterministischen oder gar rein deterministischen) algorithmischen Strukturen.

Der Roboter Cog ist ein Beispiel für einen sehr schwach funktionalen Akteur. Seine ethische Sensitivität ist im Vergleich zu der Kismets deutlich gesteigert. Auch seine Autonomie ist aufgrund eines „unsupervised learning algorithm" (24, p. 70) deutlich komplexer als Kismets. So beginnt Cog, ohne, dass er zuvor in dieser Weise programmiert worden wäre, ein Spielzeugauto nur noch von vorne oder hinten anzustoßen, um es in Bewegung zu versetzen, nachdem er durch mehrere Versuche feststellen konnte, dass es sich nicht bewegt, wenn es von der Seite angestoßen wird. Cog lernt durch Erfahrung, und vielleicht ist es gerade diese (in seinem Fall sehr begrenzte) Fähigkeit zu lernen, die es uns erlaubt, ihn als einen schwachen funktionalen Akteur zu verstehen oder aber als immerhin stark operationalen. Cog funktioniert wohl maßgeblich auf determinierten (aber nicht-deterministischen) oder sogar bereits auf nicht-determinierten (und damit nicht-deterministischen) algorithmischen Strukturen.

Auch autonome Fahrassistenzsysteme lassen sich als ein Beispiel für operationale Akteure anführen, da ihre Autonomie mit guten Gründen in strengen Grenzen gehalten ist; sie können nicht lernen und verfügen nicht über nicht-determinierte Algorithmen.



Kismet

Mit Hilfe von Darwalls Differenzierung kann eine klare Grenze zwischen genuiner (menschlicher) Akteursfähigkeit im vollen Sinne und artifizieller (operationaler und funktionaler) Handlungsfähigkeit gezogen werden. Während menschliche Akteure über alle vier Autonomietypen, nämlich personale, moralische, rationale und agentiale Autonomie, verfügen, ist Maschinen zumindest auf absehbare Zeit nur rationale und agentiale Autonomie funktional äquivalent zuzuschreiben.

Eine generelle Modifikation der implementierten algorithmischen Strukturen analog der evolutionären menschlichen Entwicklung ist bei keinem artifiziellen System vorstellbar (von der Wünschbarkeit ganz zu schweigen). Vorstellungen, in denen Maschinen die Weltherrschaft übernehmen, da sie in der Lage sind, ihre eigenen Parameter völlig ungebunden zu manipulieren, bleiben irreale Utopien. Denn auch bei nicht-determinierten Algorithmen sind nicht alle vorstellbaren Ergebnisse möglich. Auch Menschen bleiben in ihren Möglichkeiten immer beschränkt, obwohl man ihren adaptiven Spielraum sehr viel größer einschätzt als der eines noch so komplexen Roboters jemals sein könnte.

Moral implementieren – drei Ansätze

Es lassen sich generell drei Vorgehensweisen differenzieren, die Roboter mit Moralität ausstatten: Top-down-Ansätze, Bottom-up-Ansätze und hybride Ansätze (25).

Im Rahmen der Top-down-Ansätze werden eine Reihe ethischer Prinzipien oder Regeln, nach denen sich das artifizielle System in einer fraglichen Situation richten soll, wie etwa Immanuel Kants Kategorischer Imperativ, die Goldene Regel, die zehn Gebote oder die drei (bzw. vier) Asimovschen Robotergesetze, einprogrammiert. Letztere gehören zu den berühmtesten und in der Robotik und KI-Forschung bis in die Gegenwart gerne als Gedankenexperiment genutzten Beispielen für top down zu programmierende Regel-Sets. In Asimovs Kurzgeschichte Runaround von 1942 lauten sie:
„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“ (26).

Das sogenannte nullte Asimovsche Gesetz, um das Asimov in seinem Roman Aufbruch zu den Sternen die drei ersten Robotergesetze ergänzt, besagt, dass ein Roboter die Menschheit nicht verletzen oder durch Passivität zulassen darf, dass sie zu Schaden kommt.

Bei den Top-down-Ansätzen ist generell jedoch mit mindestens zwei Schwierigkeiten zu rechnen:
Zum einen sind Regeln oder gar einzelne Begriffe nur (wenn überhaupt) reduziert implementierbar. Ihre Interpretation ist kontextsensitiv. Die Programmierung ist jedoch auf eine (oder mehrere) eindeutige Interpretationen angewiesen. Zum anderen kann ein Konflikt zwischen den einzelnen Regeln auftreten. Legt man einen monistischen Ansatz zugrunde, wird eine einzelne Regel (etwa

Kants Kategorischer Imperativ) programmiert, aus der alle 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.

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 die Gefahr des Regelkonflikts. Es fehlt bei einem reinen Top-down-Ansatz das, was man ‚gesunden Menschenverstand' nennt.

Bottom-up-Ansätze basieren auf der Grundlage von Lern- und evolutionären Algorithmen (randomisierte, stochastische oder probabilistische Algorithmen). Es handelt sich um nicht-determinierte Algorithmen, bei denen nicht reproduzierbare und undefinierte Zustände auftreten. Im Gegensatz zu determinierten Algorithmen gelangt man in einem begrenzten Wahrscheinlichkeitsrahmen zu nicht programmierten Zuständen. So arbeitet Klaus Mainzer an dynamischen Systemen, in denen durch die komplexe Wechselwirkung der Elemente neue Eigenschaften des Gesamtsystems erzeugt werden, die nicht auf die einzelnen Elemente zurückzuführen sind (Emergenz; 27). Dabei werden nicht von vornherein moralische Regeln bzw. Sets an Regeln vorgegeben, sondern lediglich basale Parameter formuliert bzw. basale Kompetenzen implementiert. Artifizielle Systeme entwickeln darauf durch verschiedene Formen des Lernens (Trial and Error, Imitation, Induktion und Deduktion, Exploration, Lernen über Belohnung, Assoziation und Konditionierung) moralisches Verhalten (28).

Bei den Bottom-up-Ansätzen unterscheidet man Evolutionsmodelle (29) von Modellen menschlicher Sozialisation (30). Erstere simulieren 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 früheren Stadien der Moralentwicklung in Robotern zum Einsatz kommen.

Modelle menschlicher Sozialisation berücksichtigen die Rolle von Empathie und Emotionen für moralisches Lernen. Dabei ist zwischen zwei Formen von Mitgefühl zu differenzieren (31): zwischen perzeptueller Empathie, die bereits dann gegeben ist, wenn eine beobachtete Emotion bei mir eine vergleichbare oder kongruente Reaktion bei meinem Gegenüber auslöst (32) und imaginativer Empathie, die einen Perspektivwechsel in Form eines Sich-Hineinversetzen in das Gegenüber erfordert. Perzeptuelle Empathie wird mithilfe bestimmter „Theories of Mind" oder über neuronale Resonanz und das Wirken von Spiegelneuronen erklärt und ließ sich bereits rudimentär in artifiziellen Systemen hervorrufen (33). Über diese grundlegende Form des Mitgefühls als Wurzel von prosozialem Verhalten verfügen bereits kleine Kinder, aber auch Schimpansen (34).

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 (35). Zumindest bei der perzeptuellen Empathie ist es denkbar, dass sie sich in der basalen Form eines Affektprogrammes (36) als automatisiertes Reaktionsschema auch Robotern zuschreiben lässt – 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 metaethischen 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. Rodney Brooks, der als einer der ersten das Zusammenwirken von künstlichem System und Umwelt als Bedingung für die Entwicklung von Fähigkeiten betrachtete und von dieser Annahme ausgehend das Feld der „behavior-based robotics“ begründete (37). Zahlreiche berühmte Beispiele der gegenwärtigen Robotik und KI-Forschung, die sich an dem Ansatz verkörperten menschlichen Lernens orientieren – wie beispielsweise die Lernplattformen iCub, Myon, Cb^2, Curi, Roboy (die im Detail sehr unterschiedlichen evolutionsbasierten Ansätze folgen) –, entwickeln Roboter, die sich ähnlich Kindern Kompetenzen aneignen, aus denen sie in spezifischen Kontexten konkrete Handlungsprinzipien ableiten.

Hybride Ansätze kombinieren Top-down- mit Bottom-up-Ansätzen, indem ein ethischer Rahmen basaler Werte vorgegeben wird, der dann durch Lernprozesse an spezifische Kontexte anzupassen ist. Dabei ist die Auswahl der fraglichen Regeln von dem Einsatzbereich des Roboters abhängig. Um von einem hybriden Modell sprechen zu können, muss das System in einem anpassungsfähigen Spielraum agieren können, innerhalb dessen es auf die Wertvorstellungen seiner NutzerInnen kontextsensitiv reagiert. Catrin Misselhorn (38) entwirft im Rahmen der Altenpflegerobotik gegenwärtig ein solches hybrides System.

Bereits Georges Canguilhelm hat von einem Spielraum gesprochen (wenn auch noch nicht in einem roboterethischen Kontext) und unterschiedliche potenzielle „Freiheitsgrade“ eines „Mechanismus“ expliziert (39, S. 185 f.). Je mehr Spielraum, je mehr Freiheitsgrade oder Handlungspotenzial ein Mechanismus aufweist, desto weniger Teleologie im Sinne einer Finalität liegt Canguilhem zufolge vor (ebd. S. 213). 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 etwa ein komplexer Serviceroboter für Privathäuser, der nicht nur in der Küche unterstützen 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. 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), hier allerdings in hohem Grade flexibel die Anweisungen der NutzerInnen aufnehmen und antizipieren können muss (bottom-up).

Viele stehen diesen Ansätzen kritisch gegenüber. So Hubert L. Dreyfus, der die Entwicklung artifizieller Systeme von ihren Ursprüngen an kritisch reflektiert hat. Er verweist auf die menschliche Kreativität als den kategorialen Unterschied zu Maschinen, den diese nicht werden überwinden können (40) Hervorragend geeignet sind sie ihm zufolge allerdings zur Informationsverarbeitung im Sinne von komplexen Rechenvorgängen. Sicher ist, dass in absehbarer Zukunft nicht mit Robotern als moral agents zu rechnen ist, insofern Autonomie und ethische Sensitivität nur in einem schwach funktionalen oder gar nur in einem operationalen Sinne äquivalent simuliert werden können.

Roboter als „moral patients“

Inwiefern aber sind artifizielle Systeme als moral patients – als Wertträger – zu identifizieren?

Im Straßenverkehr gerät ein autonomes Fahrassistenzsystem aller Voraussicht nach nicht allzu häufig in Situationen von moralischer Relevanz, einmal vorausgesetzt, es hält sich stets an die Straßenverkehrsordnung, fährt nicht zu schnell und drängelt nicht. Innerhalb des Rahmens normaler Fahrroutine scheint es weniger um eine etwaige (Un-)Moralität des autonomen Wagens zu gehen, als vielmehr um Programmierungsfehler von letztlich strafrechtlichem Belang, die in den Kompetenzbereich des Herstellers fallen. Ein autonomes Fahrassistenzsystem sollte in Unfallsituationen ja gerade in der Lage sein, die wahrscheinlichen Folgen in Millisekunden zu berechnen und nach einem zuvor definierten Setting zu agieren (41).

Stellen wir uns einige Kinder vor, die unerwartet auf die Straße und direkt vor ein autonomes Auto springen. Das autonome Fahrassistenzsystem berechnet nun, dass es nicht mehr rechtzeitig wird bremsen können. Es könnte hingegen sowohl in den Gegenverkehr lenken als auch in die andere Richtung, in der sich hinter einem Brückengeländer ein Abhang auftut. Während das Auto im Rahmen der ersten Option (die Spur halten und bremsen) mit hoher Wahrscheinlichkeit die Kinder überfahren oder sie schwer verletzen würde, verlöre im Rahmen der zweiten und dritten Option (Gegenverkehr und Abhang) mindestens der/die FahrerIn das Leben. Szenarien dieser Art thematisieren eine Entscheidung von großer moralischer Relevanz, für die es keine eindeutige, keine korrekte, Antwort bzw. Lösung gibt. Bleibt keine Zeit, in der fraglichen Situation an Stelle des Autos über das weitere Vorgehen nachzudenken und sich dementsprechend zu entscheiden, wird der autonome Wagen vermutlich mit einer standardisierten Reaktionsweise ausgestattet zu reagieren – beispielsweise in solchen und ähnlichen Fällen immer zu bremsen. Das ist es, was diese Dilemma-Fälle, in denen das autonome Auto über Zeit und Kompetenz verfügt, die Situation zu analysieren und nach einer vorgegebenen Agenda zu reagieren, grundlegend von Situationen unterscheidet, in denen ein menschlicher Fahrer reflexartig handelt (42:75).

Utilitaristische Versuche, einen Nutzen aller einzelnen Beteiligten und vor diesem Hintergrund den Gesamtnutzen zu berechnen, sind generell problematisch. Viel zu viele Faktoren, die auch Menschen nicht allgemein verbindlich bedenken könnten, müsste das autonome Auto in den Blick nehmen können. Darüber hinaus ist in einigen Fällen mit für manche Beteiligten folgenschweren Entscheidungen zu rechnen, wie die, dass der autonome Wagen immer Motorradfahrer überfährt, die einen Helm und nicht diejenigen, die keinen tragen (43). Würde dies allgemein bekannt, könnte es Motorradfahrer dazu animieren, sich ohne Helm auf ihre Maschine zu setzen, was zu weniger Sicherheit im Straßenverkehr führte. Doch alle diese Schwierigkeiten einmal bei Seite gelassen, führt John Taurek ein prinzipielles Argument gegen das konsequenzialistische Denken in diesem Zusammenhang an: Es sei längst nicht eindeutig, dass die Zahlen in solchen Dilemma-Situationen zählen sollten, da „suffering is not additive in this way“ (44: 308). Was hier auf dem Spiel stünde, ist der Verlust von etwas individuell äußerst Wertvollem, weshalb es nicht sinnvoll erscheint, unparteiisch und gleichwie objektiv den Verlust addieren zu wollen: „His loss means something to me only, or chiefly, because of what it means to him. It is the loss to the individual that matters to me, not the loss of the individual” (ebd.: 307).

Bislang konnte noch niemand eine Lösung für solche Dilemma-Fälle finden, die allen unseren moralischen Intuitionen gerecht wird. Deshalb scheint es schlicht falsch zu sein, dem autonomen Fahrassistenzsystem durch die Hersteller eine Standardreaktion zu implementieren. Denn hierdurch würde die Autonomie der FahrerInnen beeinträchtigt. Der Hersteller ist für die Gewähr der Einhaltung der oben beschriebenen Fahrroutine zu-

ständig, im Sinne eines fehlerfreien Funktionierens und der Sicherheit des Autos. Doch so lange es noch einen Fahrer des Wagens gibt – selbst dann, wenn sie bzw. er gar nicht selbst fährt –, bleibt diese Person für die moralischen Entscheidungen des Fahrzeugs verantwortlich (45). Die moralische Verantwortung des hybriden Systems – bestehend aus Fahrer und autonomen Wagen (46) – wird insofern zwischen Auto und Fahrerin geteilt, als der Wagen selbst, abhängig von seinem Grad an Autonomie und moralischer Sensitivität, bis auf Weiteres ausschließlich für die Standard-Fahrparameter verantwortlich ist, die Fahrerin bzw. der Fahrer hingegen für die moralisch relevanten Entscheidungen des Autos, die nicht durch die Straßenverkehrsordnung abgedeckt werden.

Daher ist ein nicht-konsequenzialistischer Ansatz für den Umgang mit den oben geschilderten Dilemma-Situationen angebracht, der die Autonomie der Fahrer sowie ihre Fähigkeit, Rede und Antwort zu stehen, ernst nimmt. Da dem Auto selbst mit großer Wahrscheinlichkeit bis auf weiteres keine genuine Moralfähigkeit zuzuschreiben ist, werden Entscheidungen moralischer Natur noch für eine sehr lange Zeit den menschlichen Akteuren vorbehalten bleiben – sowohl dem Fahrer wie auch der Gesellschaft. In dem Maße, in dem autonome Fahrassistenzsysteme eine immer deutlichere funktionale (und nicht bloß operationale) Moralität innehaben, werden sie immer fähiger, eine gegebene Situation unter den moralischen Prinzipien des moralisch verantwortlichen Akteurs zu analysieren und auf diese Weise Schritt für Schritt immer besser darin, diese Prinzipien auf reale Straßenverkehrssituationen zu übertragen.

Da die oben beschriebenen Dilemma-Fälle kein spontanes Nachdenken gestatten, sind etwaige moralische Entscheidungen der Fahrerin bzw. des Fahrers zuvor zu treffen – etwa in Form der Erstellung eines moralischen Profils; vielleicht über einen Fragebogen oder mit Hilfe eines Programms. Es erscheint plausibel, dass diese moralischen Settings über eine elektronische Identifikation gesichert werden; wie ein elektronischer Schlüssel oder über das Smartphone der WagenbesitzerInnen (gesetzt den Fall, die nötigen Sicherheitsvorkehrungen können garantiert werden).

Hieraus folgt, dass die IT-Abteilungen, die für die Programmierung der autonomen Fahrmechanismen zuständig sind, ein ethisches Training zu absolvieren haben, um in der Lage zu sein, die potenziell moralisch relevanten Situationen zu antizipieren, zu identifizieren und angemessene Interfaces zu entwickeln, die verlässlich die moralischen Überzeugungen ihrer Kunden aufgreifen. Darüber hinaus ist ein breit geführter Diskurs vonnöten, um ein Bewusstsein für solche Dilemma-Situationen und die Herausforderungen, die mit ihnen einhergehen, zu schaffen und die Verkehrsteilnehmer auf ihre moralische Verantwortung vorzubereiten.

Unser traditionelles Verständnis von Verantwortung (47), das in den hier beschriebenen Fällen zum Ausdruck kommt, ist insofern stark individualistisch, als wir immer ein Subjekt als Verantwortungsträgerin benötigen. Eine Zuschreibung von Verantwortung ist nicht oder zumindest nur metaphorisch möglich, wenn die potenziellen Akteure die nötigen Kompetenzen (Kommunikationsfähigkeit, Autonomie bzw. Handlungsfähigkeit und Urteilskraft) nicht oder nicht hinreichend ausgeprägt mitbringen – wie Kinder, Menschen mit einer körperlichen oder geistigen Beeinträchtigung oder Maschinen. Für solche Fälle wurden in den vergangenen Jahren behelfsmäßige Begrifflichkeiten entwickelt, die ohne eine Bestimmung dieses Relationselements auskommen (48). Doch damit ist der eigentlichen Aufgabe des Verantwortungskonzepts – in intransparenten Kontexten, die durch komplexe Hierarchien und vielfach vermittelte Handlungsabläufe gekennzeichnet sind, für mehr Struktur, mehr Transparenz und Handlungsorientierung zu sorgen –, nicht gedient, suchen wir doch de facto immer nach einem Träger der eingeforderte Verantwortung.

Allerdings haben wir es hier mit Situationen

zu tun, in denen einige der in das Geschehen involvierten Parteien die zur Verantwortung notwendigen Kompetenzen nicht oder nur in einem geringen Ausmaß mitbringen. Nehmen wir wieder das Beispiel autonomer Fahrassistenzsysteme als operational moralische Akteure: Ein solches System ist insofern ein Wertträger, als es Teil unseres moralischen Universums ist, insofern ihm ein instrumenteller Wert zukommt – aber als moralischer Akteur in einem signifikanten (d. h. zumindest in einem funktionalen) Sinne lässt es sich nicht begreifen. Und dennoch haben wir die Intuition, dass wir es aus der Verantwortung nicht gänzlich entlassen können.

Für solche und vergleichbare Kontexte eines Lokalisierungsversuchs artifizieller *moral patients* im moralischen Universum habe ich den Begriff des Verantwortungsnetzwerkes von Christian Neuhäuser (49) übernommen und spezifiziert (50). Die diesen Überlegungen zugrundeliegende These lautet, dass wir all denjenigen Parteien in einer gegebenen Situation Verantwortung zuschreiben, die an dem fraglichen Geschehen beteiligt sind und zwar in dem Maße, in dem sie die nötigen Kompetenzen zur Verantwortungszuschreibung mitbringen. Ein solches Verantwortungsnetzwerk trägt der Tatsache Rechnung, dass sich innerhalb einer Verantwortungskonstellation in manchen Fällen Relationselemente überlagern können. Ein Beispiel dafür ist die Verantwortung der Eltern für ihre Kinder. Hier stellen die Kinder (bzw. deren Wohlergehen) einerseits das Objekt der Verantwortlichkeit dar, sie sind aber auch die Adressaten (also den Grund des Vorhandenseins dieser Verantwortlichkeit; ausführlich dazu 47: 117f.). Zum Verantwortungsnetzwerk „Verantwortung im Straßenverkehr“ gehören alle am Straßenverkehr Beteiligten.

Verantwortungsnetzwerke haben häufig ungewöhnliche Ausmaße – als weitere Beispiele wären die Verantwortungsnetzwerke „Klimaverantwortung“, „Verantwortung in internationalen Beziehungen“ und auch „deutsche Verantwortung“ zu nennen – und bündeln in sich unterschiedliche Verantwortungsobjekte. Von Verantwortungsnetzwerken ist dann zu sprechen, wenn man sich – sehr schön zu veranschaulichen am Fall der Klimaverantwortung (47: Kapitel 13) – gar nicht recht sicher ist, ob hier in einem gehaltvollen Sinn Verantwortung definiert werden kann: wenn die Bestimmung eines Subjekts schwierig erscheint, wenn sich keine eindeutige Instanz ausmachen lässt oder die normativen Kriterien nicht benannt werden können. In einem Verantwortungsnetzwerk erfüllen die involvierten Parteien unterschiedliche Funktionen bzw. besetzen manchmal mehrere Relationspositionen zugleich. Sie können die Verantwortungssubjekte sein, in einem anderen Fall die Instanzen und wieder in einem anderen Fall zugleich Objekt und Adressat.

Es wäre nicht möglich, ein oder mehrere konkrete Verantwortungssubjekte für ‚die‘ Verantwortung im Straßenverkehr auszumachen, da diese viel zu umfassend ist, als dass eine Person oder eine geringe Anzahl Einzelner dafür Rede und Antwort stehen könnte. Im Verantwortungsnetzwerk „Verantwortung im Straßenverkehr“ werden mehrere Verantwortungsbereiche – moralische, juristische, politische, aber auch ästhetische und andere (definiert über moralische, juristische, politische sowie ästhetische Normen) – eingefasst. Der Straßenverkehr stellt das übergeordnete Verantwortungsobjekt dar, für das nicht eine oder mehrere Personen *die* Verantwortung tragen. Es differenziert sich vielmehr in mehrere weniger komplexe Verantwortungsgegenstände aus, für die dann die unterschiedlichen Parteien jeweils eine spezifische Verantwortung übernehmen. Verantwortung für den Straßenverkehr kann in einem Fall die Sicherheit der am Straßenverkehr beteiligten Menschen bedeuten, in einem anderen Verständnis die Verantwortung dafür, schnell und effizient von A nach B zu gelangen, aber auch die ästhetische Gestaltung von Straßenverlauf und Bürgersteigen und in noch einem anderen Fall die Verantwortung dafür, dass die moralischen und ethischen Herausforderungen, die mit einer Beteiligung am Straßenverkehr einhergehen, diskutiert bzw. denjenigen, die sich am Straßenverkehr beteiligen, zuvor hinreichend deutlich gemacht wurden. All

diese und zahlreiche weitere Verantwortlichkeiten sind Teil des Verantwortungsnetzwerkes „Verantwortung im Straßenverkehr“. Dabei werden unterschiedliche Subjekte in unterschiedlichem Ausmaß zur Verantwortungsübernahme angesprochen, verschiedene Instanzen, Adressaten und Normen.

Ein autonomes Fahrassistenzsystem kann gegenwärtig nur ein sehr schwacher Verantwortungsakteur sein. Es kann deshalb die Subjektposition einer Verantwortlichkeit innerhalb des Verantwortungsnetzwerkes „Verantwortung im Straßenverkehr“ nicht besetzen, gibt es doch immer potenziell qualifiziertere Verantwortungssubjekte – wie die Fahrerin bzw. den Fahrer. Allerdings ist denkbar, es als Verantwortungsobjekt und als Adressat in eine oder mehrere der Verantwortlichkeiten dieses Verantwortungsnetzwerkes einzubinden. In dieser Weise kann Verantwortung im Bereich der Roboterethik, die sich mit artifiziellen Systemen als Wertträgern befasst, letztlich alle denkbaren Maschinen in etwaige Verantwortungskonstellationen integrieren.

Fazit und Ausblick

Mit Einteilung der Roboterethik in zwei Arbeitsfelder – Roboter als *moral patients* und als *moral agents* – wird ein grundlegender Vergleich zwischen Roboter- und Tierethik möglich. Da künstliche Systeme bislang nur in einem operationalen bzw. in einem schwach funktionalen Sinne als moralische Akteure identifizierbar sind, betreffen die meisten Fragen, mit denen wir uns aktuell innerhalb der Robotik in Industrie, Service und Kriegsführung konfrontiert sehen, fast ausnahmslos den Bereich der Roboterethik zu artifiziellen Systemen als Wertträgern. Es lässt sich den Positionen eines Anthro-, Patho-, Bio- und Physiozentrismus eine weitere Sicht zur Lokalisierung von Phänomenen im moralischen Universum hinzufügen – vielleicht ein Mathenozentrismus (wie oben vorgeschlagen) –, die alle Wesen mit einem Eigenwert bemisst, die lernfähig sind. Lernfähigkeit bedeutet mindestens eine Programmierung durch nicht-determinierte Sets an Algorithmen. Solche Phänomene 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 (aber nicht-deterministischer) Sets an Algorithmen arbeiten und sich eher im Bereich operationaler Moralzuschreibung bewegen, immerhin einen hohen instrumentellen Wert zuzuschreiben.

Das Konzept der Verantwortungsnetzwerke erlaubt eine Lokalisierung auch der artifiziellen Systeme – aber letztlich aller Phänomene, da diese Idee nicht auf Maschinen beschränkt ist – im moralischen Universum, die die Kompetenzen moralischer Akteursfähigkeit und Verantwortungszuschreibung nicht oder nicht hinreichend mitbringen.

UNSERE AUTORIN:

Janina Loh (geb. Sombetzki) ist promovierte Philosophin und arbeitet im Bereich Technik- und Medienphilosophie an der Universität Wien.

Von ihr ist zum Thema erschienen:
Sombetzki, Janina (2016): „Verantwortung und Roboterethik – ein kleiner Überblick“. In: Humboldt Forum Recht. 03/2016, S. 10-30.

Demnächst erscheint:
Loh, Janina / Loh, Wulf (2017, im Erscheinen): „Autonomy and responsibility in hybrid systems – the example of autonomous cars“. In: Lin, Patrick/Jenkins, Ryan/Abney, Keith/Bekey, George (Hrsg.): Robot Ethics. Oxford University Press.

Die weiteren im Text genannten Literaturangaben finden sich auf der Frontseite unserer Internetseite:

www.information-philosophie.de