„Artificial Morality“

Möglichkeiten und Grenzen der Maschinenethik

CATRIN MISSELHORN
Geboren 1970 in Stuttgart, 2012 bis 2019 Lehrstuhlinhaberin für Wissenschaftstheorie und Technikphilosophie an der Universität Stuttgart, seit April 2019 Professorin für Philosophie an der Georg-August-Universität Göttingen, Forschungsprojekte unter anderem über philosophische Probleme der Künstlichen Intelligenz sowie Roboter- und Maschinenethik.

Durch die Fortschritte der Künstlichen Intelligenz (KI) und Robotik werden Maschinen in Zukunft mehr und mehr grundlegende moralische Entscheidungen fällen, die unser Leben betreffen. Maschinenethik ist eine neue Disziplin an der Schnittstelle von Informatik und Philosophie, der es um die Entwicklung einer Ethik für Maschinen im Gegensatz zur Entwicklung einer Ethik für Menschen im Umgang mit Maschinen geht. Man spricht in Analogie zu „Artificial Intelligence“ auch von „Artificial Morality“ (Misselhorn 2018). Während „Artificial Intelligence“ zum Ziel hat, die kognitiven Fähigkeiten von Menschen zu modellieren oder zu simulieren, geht es bei der „Artificial Morality“ darum, künstliche Systeme mit der Fähigkeit zu moralischem Entscheiden und Handeln auszustatten. Die Idee ist also, Computer so zu programmieren, dass sie moralische Entscheidungen treffen können.

Lange Zeit stand die Maschinenethik zu Unrecht im Verdacht, lediglich Science-Fiction zu sein. Doch schon eine so simple Maschine wie ein Staubsaugerroboter steht vor moralischen Entscheidungen: Soll er einen Marienkäfer einfach einsaugen oder soll er ihn verscheuchen beziehungsweise umfahren? Und wie sieht es mit einer Spinne aus? Soll er sie töten oder ebenfalls verschonen? Ein solcher Roboter ist in einem minimalen Sinn autonom, weil er im Unterschied zu einem konventionellen Staubsauger nicht von einem Menschen geführt oder überwacht wird.

„KILL-BUTTON“ FÜR SPINNEN

Die Entscheidung ist moralisch, weil sie sich darauf bezieht, ob man Tiere zu Reinigungszwecken töten darf. Gewöhnliche Staubsaugerroboter besitzen allerdings noch nicht die Fähigkeit, eine solche Entscheidung zu treffen. Es gibt jedoch erste Ansätze, einen um ein Ethikmodul erweiterten Prototyp des populären Modells Roomba zu schaffen (Bendel 2017), der das Leben von Insekten berücksichtigt. Der Prototyp ist mit einem optionalen „Kill-Button“ für Spinnen ausgestattet.

Je komplexer die Einsatzbereiche autonomer Systeme sind, desto anspruchsvoller werden die moralischen Entscheidungen, die sie treffen müssen. Dies zeigt sich beispielsweise an Pflegesystemen, Kriegsrobotern und autonomen Fahrzeugen, den zentralen Anwendungsfeldern der Maschinenethik (Misselhorn 2018). In allen drei Bereichen stehen autonome Systeme vor grundlegenden moralischen Entscheidungen, bei denen es manchmal sogar um Leben und Tod von Menschen geht. Kann man Maschinen solche Entscheidungen überlassen, darf man es oder sollte man es gar? Das sind die grundlegenden Fragen, mit denen sich die Maschinenethik auseinandersetzen muss.

Seit jeher verbindet sich mit dem Einsatz von Maschinen die Hoffnung, dass sie Menschen von Tätigkeiten entlasten. Doch je intelligenter und autonomer Maschinen werden, desto eher werden sie in Situationen geraten, die moralische Entscheidungen erfordern. Die zentrale Begründung der Maschinenethik ist deshalb, dass die Entwicklung von Maschinen mit moralischen Fähigkeiten unverzichtbar ist, wenn wir die Vorteile autonomer intelligenter Technologien voll ausnutzen möchten (Allen et al. 2011).

Ferner wird geltend gemacht, dass moralische Maschinen bessere Maschinen sind. Wie gut eine Maschine ist, bemisst sich daran, wie gut sie den menschlichen Bedürfnissen und Werten gerecht wird. Eine Maschine, der die Moral bereits einprogrammiert ist, wird den menschlichen Bedürfnissen und Werten besonders gut entsprechen.

Schließlich wird zugunsten künstlicher moralischer Akteure angeführt, dass sie moralisch sogar besser handeln als Menschen, weil sie weder irrationalen Impulsen, Psychopathologien noch emotionalem Stress unterliegen.

Nicht zuletzt können sie in Sekundenbruchteilen Entscheidungen treffen, in denen ein Mensch nicht mehr zu bewusstem Entscheiden in der Lage ist. Das spricht aus Sicht mancher Autoren dafür, ihnen moralische Entscheidungen in besonders prekären Situationen zu überlassen, beispielsweise im Krieg (Arkin 2009).

Nicht nur aus praktischen Erwägungen ist die Maschinenethik von Bedeutung: Sie stellt auch theoretisch gesehen ein interessantes Forschungsprogramm dar. Das gilt zum einen für die ethische Theoriebildung. Die menschliche Moral ist fragmentiert und teilweise widersprüchlich. Die Entwicklung künstlicher Systeme mit moralischen Fähigkeiten macht es erforderlich, die menschliche Moral (zumindest in den Anwendungsbereichen) zu vereinheitlichen und konsistent zu machen, weil künstliche Systeme nur auf dieser Grundlage operieren können. Insofern Einheitlichkeit und Widerspruchsfreiheit generell theoretische Tugenden darstellen, wäre das auch ein Fortschritt der Ethik als Theorie der Moral.

MODELL DES MENSCHLICHEN GEISTES

Aber auch für die Kognitionswissenschaften ist die Maschinenethik von großem Interesse. Einerseits ist der Mensch Vorbild bei der Entwicklung intelligenter Maschinen, die die Fähigkeit zum moralischen Handeln haben. Andererseits hat die wissenschaftliche und technische Entwicklung des Computers und der KI auch die Vorstellungen über die Beschaffenheit des menschlichen Geistes inspiriert, und vielfach wird der Computer als Modell für die Funktionsweise des menschlichen Geistes betrachtet.

So besteht die Hoffnung, dass der Versuch, künstliche Systeme mit moralischen Fähigkeiten zu konstruieren, auch Rückschlüsse darüber zulässt, wie moralische Fähigkeiten bei Menschen funktionieren könnten (Misselhorn 2019b). Im besten Fall gibt es grundlegende funktionale Strukturen moralischer Fähigkeiten, die sowohl in natürlichen als auch in künstlichen Systemen realisiert werden können.

Scheitern gewisse Erklärungsansätze moralischer Fähigkeiten an der Implementation, so ist auch das zumindest im negativen Sinn aufschlussreich im Hinblick darauf, wie die menschlichen Moralfähigkeiten nicht funktionieren. Maschinenethik besitzt also einen Wert als Instrument kognitionswissenschaftlicher Erkenntnis.

Den angeführten Möglichkeiten steht jedoch eine Reihe von Einwänden gegenüber, die die Grenzen der Maschinenethik thematisieren. Diese beziehen sich einerseits auf die technische Machbarkeit und andererseits auf die moralische Wünschbarkeit. So ist das Argument von der Unabdingbarkeit der Maschinenethik mit einer skeptischen Position konfrontiert, die ihre

Realisierbarkeit grundsätzlich infrage stellt. Die Skepsis gegenüber der Maschinenethik speist sich aus dem Zweifel daran, dass der menschliche Geist analog zu einem Computerprogramm funktioniert. Häufig wird argumentiert, dass ein Computer nicht über das gleiche Denken oder Bewusstsein verfügen kann wie der menschliche Geist (Searle 1980 [1986]). Aus diesem Grund sei jeder Versuch aussichtslos, eine starke Künstliche Intelligenz zu entwickeln, die der menschlichen Intelligenz entspricht. Infolgedessen gebe es auch keine Aussicht darauf, Maschinen zu entwickeln, die zu moralischem Entscheiden und Handeln in der Lage sind.

Gegen diesen grundsätzlichen Einwand ist zu sagen, dass die Maschinenethik nicht zwangsläufig mit dem Anspruch starker Künstlicher Intelligenz einhergehen muss. Für die Anwendungsbereiche ist es ausreichend, eine funktionale Moral zu entwickeln. Maschinen müssen lediglich über die entsprechenden moralischen Informationsverarbeitungsprozesse verfügen. Sie müssen in der Lage sein, moralisch relevante Merkmale einer Situation zu erkennen und nach entsprechenden moralischen Vorgaben zu verarbeiten. Das ist grundsätzlich möglich, ohne dass sie über Bewusstsein oder eine dem Menschen vergleichbare Denkfähigkeit verfügen.

FRAGWÜRDIGE MORALISCHE ENTSCHEIDUNGEN

Demzufolge handelt es sich bei Maschinen nicht um vollumfängliche moralische Akteure, wie Menschen es sind, da ihnen Fähigkeiten wie Bewusstsein, die Bezugnahme auf die Welt (Intentionalität), die Fähigkeit zur Selbstreflexion und Moralbegründung und daher auch Willensfreiheit abgehen (Misselhorn 2018). Das gehört zu den Gründen, warum Maschinen zwar moralisch handeln, aber keine Verantwortung für ihr Tun übernehmen können.

Gleichwohl können sie die Verantwortungszuschreibung an Menschen untergraben. Denn es besteht das Risiko, dass die Maschinen zu moralisch fragwürdigen Entscheidungen kommen, die niemand beabsichtigt oder vorhergesehen hat und über die niemand direkte Kontrolle besitzt. Das könnte systematisch dazu führen, dass niemand für die moralisch desaströse Entscheidung eines künstlichen Systems verantwortlich gemacht werden kann. Diese Konsequenz ist insbesondere dann problematisch, wenn es um Entscheidungen über Leben und Tod von Menschen geht (Sparrow 2007).

Nicht selten sind die kritischen Punkte Kehrseiten der positiven Aspekte der Maschinenethik. So kann man die Tatsache, dass sie uns zwingt, in bestimmten Fällen verbindliche moralische Entscheidungen zu treffen, die wir bislang offengehalten haben, auch negativ sehen. Möglicherweise werden dadurch Problemlagen eliminiert, ohne dass dies der Komplexität und existenziellen Bedeutung moralischer Situationen im Alltag gerecht wird.

Ein Beispiel hierfür sind die Dilemmasituationen beim autonomen Fahren. So ist beispielsweise nicht klar, wie sich ein autonomes Fahrzeug entscheiden sollte, wenn es ausschließlich die beiden Handlungsalternativen hat, das Leben seiner Insassen aufs Spiel zu setzen oder dasjenige von auf der Straße spielenden Kindern. Der Zwang zu einer Entscheidung ex ante erscheint in einem solchen Fall als problematisch. Ein Mensch hätte die Freiheit, sich in diesen Fällen situationsabhängig zu entscheiden. Doch das Verhalten eines autonomen Systems ist im Vorhinein festgelegt. Dadurch beschränken wir unseren Entscheidungsspielraum und die Möglichkeit, situativ von einer vorhergehenden moralischen Einschätzung abzuweichen, die uns in einer konkreten Situation nicht mehr angemessen erscheint.

SELBSTBESTIMMUNG VON MENSCHEN NICHT ERSETZEN

Dieser Gesichtspunkt führt uns zu einem anderen Einwand. So ist nach einer auf Immanuel Kant zurückgehenden Idee die Fähigkeit zu moralischem Handeln die Wurzel der menschlichen Würde. Daher kann man argumentieren, dass wir gerade dasjenige aus der Hand geben, was uns als Menschen ausmacht, wenn wir moralische Entscheidungen an Maschinen abgeben. Dieser Einwand spricht nicht prinzipiell gegen die Maschinenethik, gibt jedoch wichtige Hinweise darauf, wie moralische Maschinen gestaltet werden sollten.

Maschinen sollten die Selbstbestimmung von Menschen nicht ersetzen, sondern sie in ihrem selbstbestimmten Handeln unterstützen. So besteht eine Idealvorstellung im Bereich der häuslichen Pflege in der Entwicklung eines Systems, das sich durch Training und permanente Interaktion mit dem Nutzer auf dessen moralische Wertvorstellungen einstellen und Menschen nach ihren eigenen Moralvorstellungen behandeln kann (Misselhorn 2019a).

Ein solches System wäre vergleichbar mit einem verlängerten moralischen Arm des Nutzers, der es diesem ermöglicht, länger selbstbestimmt in seinen vier Wänden zu leben, wenn er dies möchte. Dabei ist allerdings zu beachten, dass Technologien allein den Pflegenotstand nicht lösen werden, sondern auch die sozialen und gesellschaftlichen Rahmenbedingungen einbezogen werden müssen. So sollte niemand gegen seinen Willen von Robotern gepflegt werden. Zudem darf der Einsatz von Pflegesystemen nicht zur Vereinsamung und sozialen Isolation der Gepflegten führen.

Vom moralischen Standpunkt sehr kritisch zu sehen ist es, Maschinen über Leben und Tod von Menschen entscheiden zu lassen. Diese Kritik wird gestützt durch die Tatsache, dass in Anwendungsbereichen, in denen über den Einsatz autonomer Systeme nachgedacht wird, keine moralische Pflicht besteht, zu töten (Misselhorn 2018). Eine solche Pflicht gibt es nicht einmal im Krieg. Folgt man etwa der üblichen Auslegung der Theorie des gerechten

Kriegs, so ist es bestenfalls moralisch zulässig, andere Menschen im Krieg zu töten, aber nicht moralisch verpflichtend (Childress 1979, Eser 2011). Aus diesem Grund sollte immer die Möglichkeit bestehen, etwa aus Mitleid von einer Tötungshandlung abzusehen. Der Einsatz autonomer Waffensysteme verschließt diesen Entscheidungsspielraum unweigerlich.

KRIEGSROBOTER UND AUTONOMES FAHREN

Eine wichtige Frage ist, ob die Einwände gegen Kriegsroboter sich auch auf andere Anwendungsbereiche übertragen lassen. So wurde eine Analogie zwischen der Programmierung autonomer Fahrzeuge zum Zweck der Unfalloptimierung und der Zielbestimmung autonomer Waffensysteme hergestellt (Lin 2016). Um Unfallergebnisse zu optimieren, ist es notwendig, Kostenfunktionen anzugeben, die bestimmen, wer im Zweifelsfall verletzt und getötet wird. Ähnlich wie bei autonomen Waffensystemen müssten also für den Fall einer unvermeidlichen Kollision legitime Ziele festgelegt werden, die dann vorsätzlich verletzt oder womöglich sogar getötet würden.

Lässt sich das Argument, dass es keine moralische Pflicht gibt, zu töten, auf das autonome Fahren übertragen? Dazu ist zu klären, ob eine moralische Pflicht besteht, unschuldige Menschen zu verletzen oder zu töten, sofern dies dazu dient, Schlimmeres zu verhindern. Eine solche Pflicht ist nicht nur moralisch problematisch (Misselhorn 2018). Sie stünde auch in einem Spannungsverhältnis zur deutschen Rechtsprechung. So hat das Bundesverfassungsgericht in seiner Entscheidung zum Luftsicherheitsgesetz im Jahr 2006 zum Abschuss entführter Passagierflugzeuge, die von Terroristen als Massenvernichtungswaffen eingesetzt werden sollen, festgestellt, dass ein Abschuss immer der Menschenwürde der Flugzeugpassagiere widerspricht (BVerfGE 115, 118, [160]).

Das Grundgesetz schließt aus, auf der Grundlage einer gesetzlichen Ermächtigung unschuldige Menschen vorsätzlich zu töten. Dieses Urteil steht zumindest auf den ersten Blick in einem Widerspruch zu einer Pflicht der Schadensminimierung, die die vorsätzliche Verletzung oder Tötung unschuldiger Menschen umfasst.

Aus der Diskussion der Möglichkeiten und Grenzen der Maschinenethik lassen sich drei Leitlinien für gute Maschinenethik gewinnen:
Erstens: Moralische Maschinen sollten die Selbstbestimmung von Menschen fördern, aber sie nicht beeinträchtigen.
Zweitens: Künstliche Systeme sollten nicht über Leben und Tod von Menschen entscheiden.
Drittens: Es muss sichergestellt werden, dass Menschen stets in einem substanziellen Sinn die Verantwortung übernehmen.

Ein diesen Richtlinien entsprechendes Einsatzbeispiel ist das beschriebene Pflegesystem, das als verlängerter moralischer Arm des Nutzers fungiert. Es kann diesem ermöglichen, länger selbstbestimmt in den eigenen vier Wänden zu leben, sofern dies sein Wunsch ist. Hierbei ist allerdings auf die entsprechende soziale und gesellschaftlich Einbettung solcher Technologien zu achten.

Kritisch sind im Licht dieser Leitlinien hingegen Kriegsroboter einzuschätzen. Auch das autonome Fahren sollte vor diesem Hintergrund nicht zu leichtfertig forciert werden, ohne die Alternativen zu erkunden, die das assistierte Fahren bietet. Denn es ist moralisch weniger problematisch, insofern es Maschinen keine Tötungsentscheidungen überträgt.

Literatur

Allen, Colin et al. „Why Machine Ethics?“, in: Michael Anderson / Susan Leigh Anderson (Hrsg.): Machine Ethics, Cambridge University Press, New York 2011, S. 51–61.

Bendel, Oliver: „2017: Ladybird – The Animal-Friendly Robot Vacuum Cleaner“, in: The AAAI 2017 Spring Symposium on Artificial Intelligence for the Social Good Technical Report SS-17-01, Palo Alto 2017, S. 2–6.

Childress, James F.: „Nonviolent Resistance – Trust and Risk-Taking. Twenty-Five Years Later“, in: Journal of Religious Ethics, 25. Jg., Nr. 2/1997, S. 213–220.

Lin, Patrick: „Why Ethics Matters for Autonomous Cars“, in: Markus Maurer et al. (Hrsg.): Autonomous Driving – Technical, Legal and Social Aspects, Springer Verlag, Berlin/Heidelberg 2016, S. 69–85.

Misselhorn, Catrin:

2018: Grundfragen der Maschinenethik, Reclam Verlag, Ditzingen 2018 (4. Auflage in Vorbereitung für 2019).

2019a: „Moralische Maschinen in der Pflege? Grundlagen und eine Roadmap für ein moralisch lernfähiges Altenpflegesystem“, in: Christiane Woopen / Marc Jannes (Hrsg.): Roboter in der Gesellschaft. Technische Möglichkeiten und menschliche Verantwortung, Springer Verlag, Wiesbaden. S. 53–68.

2019b: „Mensch und Maschine. Leonardo da Vinci als Vorbild für die gegenwärtige Roboterethik“, in: Ernst Seidl et al.: Ex machina. Leonardo da Vincis Maschinen zwischen Wissenschaft und Kunst, Schriften des Museums der Universität Tübingen (MUT), Band 18, Tübingen.

Searle, John R.: „Minds, Brains, and Programs“, in: The Behavioral and Brain Sciences, Nr. 3, Cambridge University Press 1980, S. 417–424; deutsche Fassung: „Geist, Gehirn, Programm“, in: Douglas R. Hofstadter / Daniel C. Dennett (Hrsg.): Einsicht ins Ich, Klett-Cotta Verlag, Stuttgart 1986, S. 337–356.

Sparrow, Robert: „Killer Robots“, in: Journal of Applied Philosophy, 24. Jg., Nr. 1/2007, S. 62–77.