Original paper UDC: 17:007.52(045)
doi: 10.21464/sp39104
Received: 13 July 2023

Arne Manzeschke
Evangelische Hochschule Nürnberg, Institut für Pflegeforschung, Gerontologie und Ethik, Bärenschanzstr. 4, DE–90429 Nürnberg
arne.manzeschke@evhn.de

Können Roboter Menschen moralisch verletzen?

Einige ethische Vorüberlegungen zu Mensch-Maschinen-Relationen

Zusammenfassung

In dem Maße, in dem Roboter als soziale Interaktionspartner des Menschen gestaltet werden, stellen sich gravierende moralische Fragen. Der Beitrag exploriert die Frage, ob Roboter Menschen moralisch verletzen können. Ich beginne die Ausarbeitung der Frage mit (1) einer kurzen Reflexion darüber, was unter „Roboter" verstanden wird. (2) skizziere ich, was es bedeutet, moralisch geschädigt zu werden. In einem weiteren Schritt (3) werden die Unterschiede zwischen physischer, psychischer und moralischer Schädigung untersucht. Dies führt (4) zu der Frage, welche Art von Akteur der Urheber eines solchen Schadens sein könnte. Daher beziehe ich mich (5) auf eine gut etablierte Taxonomie von Göttern, Menschen, Tieren, Pflanzen und Maschinen. Ich werde zu dem Schluss kommen (6), dass der Status der Handlungsfähigkeit eines Roboters von der Entscheidung der Menschen abhängt, wie sie ihn konstruieren wollen. Deshalb sollten wir als menschliche Gesellschaften gut informiert über diese wichtige Frage nachdenken. Es ist wirklich wichtig, wie wir über diese Frage entscheiden, die unser Selbstverständnis und das Gefüge des sozialen Lebens betrifft.

Schlüsselwörter

Roboter, Mensch-Roboter-Interaktion, moralische Schädigung, Technikgestaltung

Einleitung

In dem Maße, in dem Roboter als Interaktionspartner des Menschen gestaltet und in das soziale Leben der Menschen eingeführt werden, stellen sich einige schwerwiegende Fragen bezüglich des sozialen, rechtlichen und moralischen Status dieser Maschinen und der Weise, wie wir Menschen mit ihnen leben wollen.

In diesem Beitrag gehe ich der Frage nach, ob Roboter Menschen moralisch schaden können. Sollte dies der Fall sein, müsste die Konstruktion dieser Roboter genauer untersucht und entsprechend geändert werden, um solche Folgen zu vermeiden. Damit mache ich die starke Vorannahme, dass wir Menschen es nicht wollen können, dass von uns konstruierte technische Systeme Menschen verletzten – weder physisch noch psychisch noch moralisch. Wenn es bis heute unklar ist, wie die Frage zu beantworten ist, – und das scheint mir der Fall zu sein – wäre eine weitere Klärung in jedem Fall sinnvoll. Nur wenn die Frage ohne weiteres verneint werden könnte, wären weitere Überlegungen sinnlos. Ich bin davon überzeugt, dass die letztere Option keine ist, und beginne daher die Ausarbeitung der Frage mit (1.) einer kurzen Reflexion darüber, was im Folgenden unter „Roboter" verstanden wird. (2) skizziere ich, was es bedeutet, moralisch geschädigt zu werden. In einem weiteren Schritt (3.) werden die Unterschiede zwischen physischer, psychischer und moralischer Schädigung untersucht. Dies führt (4) zu der Frage, welche

Art von Akteur der Urheber eines solchen Schadens sein könnte. Daher beziehe ich mich (5) auf eine gut etablierte Taxonomie von Göttern, Menschen, Tieren, Pflanzen und Maschinen (mit einer weiteren Unterscheidung zwischen Automaten und Robotern). Ich werde zu dem Schluss kommen (6), dass der Status der Handlungsfähigkeit eines Roboters von der Entscheidung der Menschen abhängt, wie sie ihn konstruieren wollen. Deshalb sollten wir als menschliche Gesellschaften gut informiert über diese wichtige Frage nachdenken. Es ist wirklich wichtig, wie wir über diese Frage entscheiden, die unser Selbstverständnis und das Gefüge des sozialen Lebens betrifft.

1. Was meint „Roboter“?

Der Begriff Roboter findet erstmals Verwendung in Karel Čapeks Theaterstück „Rossums Universale Roboter“ von 1920 und hat von dort Eingang in viele Sprachen gefunden. Das Wort *robota* stammt aus dem Slawischen, reicht in der Etymologie aber mit einigen Lautverschiebungen bis zum Lateinischen *labor* zurück. *Rab* (slaw.) und *labor* (lat.) bezeichnen die »Arbeit«, den »Frondienst«. In Čapeks Stück sind die Roboter künstlich hergestellte Menschen (Androiden), die zu stupider Arbeit zugerichtet werden und ab einem bestimmten Zeitpunkt gegen ihre Ausbeutung rebellieren und die Menschheit vernichten. Damit sind bereits einige Motive angespielt, die die Debatte um Roboter immer noch mit bestimmen.

Betrachtet man Roboter als Weiterentwicklung von Maschinen und Automaten, so lassen sich gegenwärtig drei Stufen der Robotik verzeichnen.[1]

1) *Industrieroboter*, die nach ISO 8373 (2012) so definiert sind:

„Ein Roboter ist ein frei und wieder programmierbarer, multifunktionaler Manipulator mit mindestens drei unabhängigen Achsen, um Materialien, Teile, Werkzeuge oder spezielle Geräte auf programmierten, variablen Bahnen zu bewegen zur Erfüllung der verschiedensten Aufgaben.“

Dieser Typus von Robotern begegnet uns vor allem in den Industriehallen, wo er durch Sicherheitskäfige vom Menschen abgeschirmt seine repetitiven Aufgaben ausführt. Solche Roboter sind schnell, stark und für den Menschen gefährlich; sie verfügen über kein weiteres Orientierungsvermögen und würden einen Menschen, der ihren Operationsradius betritt, massiv schädigen.

2) Der zweite Typus von Robotern sind sogenannte *Serviceroboter*. Sie arbeiten aufgrund ihrer Kontextsensitivität in der unmittelbaren Umgebung von Menschen oder anderen Tieren (z. B. Staubsaugroboter, Melkroboter). Sie sind via Sensoren so weit über ihren Operationsraum orientiert, dass sie Menschen, Tiere, Gegenstände auch bei unvorhergesehenen Ereignissen in diesem Raum nicht verletzen können, sondern zuvor ihre Operation stoppen. Sie können mit Menschen auf einer materialen Ebene interagieren, wobei Sprache, Gestik oder Zeichen die Kommunikation unterstützen können. Die über den Industrieroboter hinausgehende Leistung des Serviceroboters besteht darin, dass er ein besseres Orientierungsvermögen über die Kontextfaktoren hat, die für seine Operationen relevant sind. Das gilt insbesondere für die Anwesenheit von Lebewesen oder sich immer wieder ändernde Umgebungsbedingungen (z. B. Lichtverhältnisse, örtliche Variationen der Objekte, unübliche Verhaltensweisen der Interaktionspartner).

3) Der dritte Typus von Robotern ist der *soziale Roboter*. Duffy *et al.* machen hierfür einen Definitionsvorschlag, der für unsere Überlegungen brauchbar

erscheint: Der soziale Roboter ist „eine physische Entität, die in einem komplexen, dynamischen und sozialen Umfeld ausreichend befähigt ist, um sich auf eine Weise zu verhalten, die förderlich ist für die eigenen Ziele und die der Gemeinschaft“.[2] Die Gemeinschaft, die hier angesprochen wird, sind zunächst einmal andere, kollaborative Roboter. Fähigkeiten solcher Roboter sind Bewegung, Manipulation und Wahrnehmung, so dass sie in eine vielschichtige Interaktion treten können: „Kollaboration, Kommunikation, Kooperation, Koordination, Gemeinschaft, Identität und Beziehungen mit reaktiven und pro-aktiven Verhaltensmodellen“ (ebd.). Die Autoren unterscheiden zwischen „social robots“, die mit anderen Robotern interagieren und „societal robots“ die mit Menschen interagieren. Die besondere Herausforderung der sozialen Interaktion zwischen Mensch und Roboter besteht vor allem darin, dass der für die Interaktion nötige Informationsaustausch und das gegenseitige „Verstehen“ weitaus komplexer sind als bei einer Roboter-Roboter-Interaktion. Um mit Menschen zu interagieren, sollen die Roboter ein „Wissen vom Menschen“ erlernen bzw. einprogrammiert bekommen, das sie instand setzt, Gestik, Sprache und unter Umständen auch Emotionen der Menschen zu erkennen, zu interpretieren, in einem weiteren Sinne zu „verstehen“. Die hierzu nötige Sprach-, Gesten- und Emotionserkennung, muss außerdem verbunden sein mit einer Orientierung im physischen Raum der Interaktion sowie einer Form der Erinnerung, die es erlaubt, aktuelle Interaktionen vor dem Hintergrund zurückliegender zu interpretieren und entsprechend zu gestalten.

Roboter mit diesen Fähigkeiten kämen dem schon sehr nahe, was wir als soziales Verhalten bezeichnen – und üblicherweise Menschen und zunehmend mehr Tierarten zusprechen. Die Leistung von sozialen oder sozio-emotionalen Roboter besteht derzeit weniger darin, bestimmte Tätigkeiten zu verrichten, als vielmehr mit den Menschen in eine soziale Interaktion zu treten. Es ist aber eine verbreitete Erwartung, dass mit entsprechendem technischen Fortschritt die Typen 2 und 3 zu einem neuen Robotertypus verschmelzen werden, der, umfassend orientiert (Raum, Personen, Interaktionsgeschichte, Sprache und Gestik, Emotionen) mit Menschen bei den verschiedensten Anforderungen in eine Interaktion (Kollaboration, Kooperation, usw.) treten könnte, die auch „handfeste Arbeit“ beinhaltet. Die gesamte und hier nur holzschnittartig dargestellte technische Entwicklung korrespondiert mit einer politischen Setzung, die das BMBF bereits 2013 in einer Förderausschreibung programmatisch so formuliert hat:

„Technische Systeme entwickeln sich immer mehr von reinen Werkzeugen zu kooperativen Interaktionspartnern. Das eröffnet vielfältige Chancen in unterschiedlichen Lebensbereichen. Sie werden Menschen zunehmend in Arbeitskontexten oder in Alltagssituationen unterstützen und einen wichtigen Beitrag leisten, ihre Produktivität, soziale Teilhabe, Gesundheit oder Alltagskompetenz zu stärken.“[3]

1
Zur Taxonomie vgl. Arne Manzeschke: „Roboter in der Pflege. Von Menschen, Maschinen und anderen hilfreichen Wesen“, *Ethik Journal* 5 (2019), S. 1–11.

2
Brian R. Duffy *et al.*., „What is a social robot?“, *10th Irish Conference on Artificial Intelligence & Cognitive Science, University College Cork, Ireland, 1-3 September, 1999* (1999). Hier erhältlich: http://hdl.handle.net/10197/4412 (abgerufen: der 26. August 2024).

3
Bundesministerium für Bildung und Forschung (BMBF) (2013) Förderbekanntmachung vom 10. Dezember 2013: »Vom technischen Werkzeug zum interaktiven Begleiter – sozial- und emotionssensitive Systeme für eine optimierte Mensch-Technik-Interaktion-Interemotio«. Hier erhältlich: https://www.bmbf.de/foerderungen/bekanntmachung-908.html (abgerufen: der 26. August 2024).

Wenn Roboter dazu befähigt werden, und die Integration der verschiedenen Robotertypen technisch tatsächlich gelingt, dann müssen wir Menschen auch die Fragen beantworten, die im sozialen Bereich durch diese neuen sozialen Agenten aufgeworfen werden. Eine dieser Fragen, die ich hier explorieren möchte, ist die, ob Menschen in der Interaktion mit Robotern moralisch geschädigt werden können.

2. Was bedeutet es, moralisch geschädigt zu werden?

Eine heutzutage weitverbreitete moralische Vorstellung ist, dass die Schädigung einer Person in allen Fällen vermieden oder so gering wie möglich gehalten werden sollte. In der Bioethik ist dieser Grundsatz als Nonmalefizienz bekannt[4] und auch über sie hinaus verbreitet worden. Die Schädigung eines Menschen ist sowohl auf physischer als auch auf psychischer Ebene vorstellbar. Beides ist moralisch abzulehnen. *Prima facie* ist es jedoch nicht möglich zu entscheiden, ob eine Schädigung auf körperlicher Ebene schwerwiegender ist als eine Schädigung auf psychischer Ebene oder umgekehrt – abgesehen davon, dass eine physische Schädigung häufig von einer psychischen begleitet wird. Was bedeutet es also, moralisch geschädigt zu sein? Geht dies über die physische oder psychische Ebene des Erlebens eines (schweren) Schadens hinaus? Um diese Frage beantworten zu können, soll im Folgenden zunächst nach dem Charakter physischer und psychischer Schädigungen gefragt werden, um so den Gehalt einer moralischen Schädigung genauer bestimmen zu können.

3. Physischer und moralischer Schaden

Zunächst möchte ich mich auf die Frage der physischen Schädigung konzentrieren. Wenn beispielsweise meine Hand von einer zuschlagenden Tür eingeklemmt wird, kann ich nicht – und sollte ich nicht – die Tür dafür verantwortlich machen. Es gibt zweifellos einen physikalischen Zusammenhang zwischen der Bewegung der Tür und dem Schmerz in meiner Hand. Aber es gibt kein moralisches Subjekt, das mir mit List oder Täuschung begegnet und die Bewegung der Tür verursacht, so dass ich zu Schaden komme. Auf Seiten der Tür lässt sich kein Handeln erkennen ebensowenig wie eine Intention. Wenn man einer Entität (bisher kommen dafür nur Menschen als moralische Personen infrage) Intention und Handlungsfähigkeit in einem substanzielleren Sinne zuschreibt, dann hat sie die Fähigkeit, Ziele zu setzen und die entsprechenden Mittel zu wählen, was einer Tür, die einfach nur den Gesetzen der Mechanik folgt, nicht zugeschrieben werden kann. Selbst eine Handlungsfähigkeit in einem schwächeren Sinne, die von Theoretikern der Akteur-Netzwerk-Theorie (ANT)[5]; oder dem „Double Dance of Agency“[6] nahegelegt wird, trifft im Fall von einfachen mechanischen Gegenständen, die nicht zu den Maschinen gezählt werden können, nicht zu.

Dennoch stellt sich bei den körperlichen Schmerzen die Frage, ob jemand für diese Misere verantwortlich gemacht werden kann. Sofern keine Fahrlässigkeit des Herstellers der Tür vorlag, war es letztlich meine eigene Unachtsamkeit, die den Schaden verursacht hat. Das Gleiche gilt für einfache Werkzeuge und Automaten: Liegt kein Produktionsfehler vor, wird der Benutzer von Werkzeugen oder Automaten als Urheber eines mit ihnen und nicht durch sie verursachten Schadens identifiziert. Es gibt also niemanden außer mir, der

Urheber und Adressat des körperlichen Schmerzes in einer Person ist. Daran schließt sich die Frage an, ob ich mir als Benutzer selbst diesen Schmerz aus Unachtsamkeit oder Unwissenheit über die Funktionsweise des schmerzverursachenden Gegenstandes beigefügt habe. Für die eigene Unachtsamkeit kann ich mich selbst zwar tadeln, aber eine moralische Schuld werde ich mir nicht zuweisen können. Anders verhält es sich mit der Unwissenheit; hier kommt es darauf an, ob ich *nicht wissen konnte* oder *nicht wissen wollte*. Im Falle, dass ich bei aller subjektiven Sorgfalt nicht wissen konnte, dass der Gebrauch des Gegenstands mich schädigen könnte, wäre eine moralische Schuld ausgeschlossen; *ultra scire nemo obligatur*. Beim Nicht-wissen-Wollen wäre noch – wie im Juristischen – zwischen einfacher und grober Fahrlässigkeit zu unterscheiden. In beiden Fällen aber scheint eine moralische Schuld gegenüber mir selbst nicht ausgeschlossen, je größer die Fahrlässigkeit desto stärker der Eindruck, dass ich mir selbst moralisch etwas schuldig geblieben bin.

Aber wie geht der körperliche Schaden mit einem moralischen Schaden zusammen? Neben dem physischen Schaden, den ein Mensch erfährt, entsteht ein moralischer Schaden, wenn in einer Interaktion mit mindestens einem anderen Wesen allgemein anerkannte moralische Grundsätze verletzt werden. Um einen moralischen Schaden zu erleiden, ist ein zweites Wesen erforderlich, ein anderer Urheber der Handlung, der der ersten Person körperliche Schmerzen zufügt und ihr gleichzeitig einen moralischen Schaden zufügt, weil er gegen vernünftige moralische Erwartungen oder Regeln verstößt. Dies kann am Ende den Schmerz noch verdoppeln.

Es liegt die Überlegung nahe, ob nicht eine Person, die sich physisch selbst verletzen kann, sich zugleich damit auch moralischen Schaden zufügt. Die maximale Form der Selbstverletzung ist zweifelsohne der Suizid. Für Kant wäre das eine Verletzung der Pflichten gegen sich selbst, indem der Mensch als sinnliches Wesen sich als moralische Person tötet.[7] Man kann den Eindruck gewinnen, dass hier zwei Menschen gegeneinander stehen (bei Kant sind es zwei Register menschlicher Existenz: die animalische und die noumenale), und der eine den anderen moralisch schädigt. Aber auch in einer zweiten Variante stellt der Suizid eine moralische Schädigung dar, und zwar in der Hinsicht, dass der Suizident gegen den Grundsatz der Nichtschädigung dessen verstößt, der ihn auffindet.[8] In beiden Fällen geht die Schädigung von einer Person aus und trifft eine andere. Bei Fairbairn ist es (mindestens) eine andere Person, die beim Auffinden der suizidierten Person erschrecken und seelischen Schaden nehmen kann. Bei Kant ist die andere Person, die

4
Tom L. Beauchamp, James F. Childress, *Principles of Biomedical Ethics*, Oxford University Press, Oxford [7]2013.

5
John Law, „Notes on the theory of the actor-network: Ordering, strategy, and heterogeneity“, *Systems Practice* 5 (1992), S. 379–393; Bruno Latour, *Enquêtes sur les modes d'existence. Une anthropologie des modernes*, Éditions La Decouverte, Paris 2012, bes. Kap. 8.

6
Jeremy Rose, Matthew Jones: „The double dance of agency: a socio-theoretic account of how machines and humans interact“, *Syst. Signs Actions* 1 (2005), S. 19–37.

7
Vgl. Immanuel Kant, *Metaphysik der Sitten*, in: Immanuel Kant. *Werke in sechs Bänden*, Wilhelm Weischedel (hg.). Bd. IV, *Schriften zur Ethik*, Wissenschaftliche Buchgesellschaft Darmstadt, Darmstadt 2005, bes. S. 549–556.

8
Vgl. Gavin J. Fairbairn, *Contemplating Suicide. The Language of Ethics and Self Harm*, Routledge, London – New York 1995.

„Menschheit in seiner Person“,[9] welche das Gegenüber für den mit innerer Freiheit (und damit zu Pflichten gegen sich selbst) begabten Menschen bildet.

Ein moralischer Schaden, so möchte ich vorläufig resümieren, bedarf a) einer zweiten Person, die dem Geschädigten gegenübersteht und neben dem hier zur Debatte stehenden physischen Schaden einen moralischen Schaden dadurch zufügt, indem sie b) gegen allgemein akzeptierte moralische Regeln und Erwartbarkeiten verstößt. Diese zweite Person kann einerseits ein konkretes moralisches Subjekt sein, das sich erkennbar nicht an die in dem zur Rede stehenden Kontext[10] moralischen Regeln hält. Sie kann aber auch in einem abstrakten Sinne die Gemeinschaft (im weitesten Sinne die Menschheit) repräsentieren, auf die hin Moral entworfen ist, weil Moral stets relational zu verstehen ist.

Dass es, um eine moralische Schädigung zu konstatieren, einer zweiten *Person* bedarf, will ich an einem weiteren Beispiel verdeutlichen. Wenn meine Katze mir beim gemeinsamen Spielen (wenn wir beide es denn als „Spiel“ verstehen!) die Hand zerkratzt, würde ich sie nicht für den physischen Schaden und mein Unbehagen verantwortlich machen, obwohl ich bereit bin, sie als Lebewesen mit eigenem Willen und eigenen Absichten wahrzunehmen. Aber es ist mir nie in den Sinn gekommen, dass sie mich absichtlich verletzt, dass sie böse oder bösartig oder, in abgeschwächter Form, fahrlässig handelt. Im philosophischen Diskurs werden Tiere höherer Stufen ähnlich wie Kleinkinder wahrgenommen, unverantwortlich für ihr Handeln und unfähig zu Verbrechen.[11] Dennoch werden Tieren bestimmte Formen moralischer Fähigkeiten als evolutionäre Frühform der menschlichen Moral zugeschrieben.[12] Wir Menschen können von anderen Tieren geschädigt werden, z. B. von einem fliehenden Pferd oder einem Wildschwein, das seine Frischlinge verteidigt. Aber auf keinen Fall würden wir uns als moralische Subjekte von diesen Tieren moralisch geschädigt fühlen, obwohl der physische Schaden von ihnen verursacht wurde und sie aus moralischen Motiven als moralische Akteure gehandelt haben mögen.[13] Selbst wenn wir uns auf die Tatsache von Vorläuferformen der Moral bestimmter Tiere einigen, würden wir sie nicht als moralisches Alter Ego (moralisches Subjekt) wahrnehmen, das uns moralischen Schaden zufügen kann. Sie stellen keine Instanz dar, an die wir uns in Form von moralischen Verpflichtungen oder Regeln wenden könnten. In einem weitaus stärkeren Sinne dürfte das für Pflanzen gelten, die wir als Lebewesen zu achten haben, die wir aber noch viel weniger als Urheber einer schädigenden Handlung und Adressaten einer moralischen Verantwortung ansehen.

Damit kehre ich zu den eingangs erwähnten mechanischen Objekten zurück und frage, ob Technik auf einer höheren Ebene der Komplexität und Selbstbewegung (Automation – die Eigenbewegung gepaart mit der Möglichkeit über verschiedene Bewegungsvarianten zu entscheiden) als Urheber einer moralischen Schädigung in Betracht kommen. Dass Industrieroboter Menschen physisch schädigen können, steht außer Frage. Genau aus diesem Grunde sind sie in den Produktionshallen durch Sicherheitskäfige von Menschen getrennt. Ihnen mangelt es an Orientierung und Kontextsensitivität, welche die Schädigung eines Menschen in den Operationsbahnen des Roboters ausschließen könnte. Sie können nichts wissen und nichts wollen, weshalb sie noch weniger als gewisse Tiere als handelnde Akteure in Betracht kommen. Ändert sich diese Einschätzung in dem Moment, in dem technische Artefakte über eine sensorbasierte Orientierung im Raum verfügen, Lebewesen detektieren und Kollisionen mit ihnen vermeiden können? Auch diese Serviceroboter

betrachten wir nicht als Akteure, denen wir Intentionen und Handlungen zuschreiben. Es sind programmierte Maschinen, denen ein gewisses Eigenmaß an Bewegung eingeräumt wird und die mit Stoppfunktionen ausgestattet sind, sofern ein belebtes oder unbelebtes Objekt ihre Bahnen kreuzt. Das lässt sich in einem sehr schwachen Sinne als ‚Entscheidung' für diese oder jene Aktion konzipieren,[14] aber sie unterscheidet sich doch in einem substanziellen Sinne von dem, was wir uns Menschen in Bezug auf Entscheidungen zuschreiben: Motive, Wünsche, Willen, Präferenzbildung, Abwägung von Ziel-Mittel-Verhältnissen usw. Sollte ein physischer Schaden durch solche Roboter verursacht werden, so wird die Verantwortung dafür bei den Herstellern gesucht und in Deutschland über eine Gewährleistungshaftung eingefordert. Entsprechend wird man bei diesem Typus von Roboter nicht davon sprechen können, dass moralische Schädigungen von ihnen ausgehen.

4. Psychischer und moralischer Schaden

Gibt es psychische Schäden, die durch Maschinen oder Roboter verursacht werden können? Das süchtige Verhalten beim Glücksspiel an Automaten mit all seinen Nebenwirkungen kann als psychischer Schaden eingestuft werden – Spielsucht wird im ICD-11 (Internationale statistische Klassifikation der Krankheiten und verwandter Gesundheitsprobleme, 11. Revision) als eigene Diagnose unter Impulskontrollstörungen eingestuft. Der Umgang mit Spielautomaten (offline oder online) trägt, wenn vielleicht auch nicht ursächlich, zu einer psychischen Schädigung bei. Allerdings kann man in diesem Fall nicht den Automaten als Verursacher identifizieren. Vielmehr sind es die persönliche Veranlagung und das individuelle Verhalten in Kombination mit (halb-)öffentlichen Glücksspielangeboten, die den Schaden ausmachen. Dasselbe gilt für den exzessiven Internetkonsum. In diesen Fällen lassen sich psychische Schäden im Zusammenhang mit dem Einsatz technischer Systeme konstatieren. Allerdings haben diese keinen Subjektcharakter; sie können nicht als Urheber, oder auch nur als Mitverursacher (im Sinne eines willentlichen und gezielten Betrags) des Schadens identifiziert werden. Von moralischer Schädigung durch sie kann in diesem Fall keine Rede sein.

Man kann fragen, ob sich an dieser Einschätzung etwas ändert, wenn das Spiel ‚intelligent' ist und sich an die Spielgewohnheiten einer Person adaptiert bzw. diese gezielt stimuliert. In diesen Fällen wird man sagen können, dass die Konstrukteure solcher Spiele die Spiellust (und wohl auch Spielsucht) der

9
I. Kant, *Metaphysik der Sitten*, S. 550.

10
Es handelt sich hierbei um Konventionen und Regeln in der Kultur; vgl. zu den „Üblichkeiten" Gernot Böhme: *Ethik im Kontext. Über den Umgang mit ernsten Fragen*, Suhrkamp, Frankfurt am Main 1997, bes. S. 28ff.

11
Vgl. Mark Rowlands, *Can Animals be Moral?*, Oxford University Press, Oxford 2012.

12
Vgl. etwa Michael Tomasello, *Eine Naturgeschichte der menschlichen Moral*, Suhrkamp, Frankfurt am Main 2016; Frans de Waal, *Der Mensch, der Bonobo und die zehn Gebote. Moral ist älter als Religion*, Übers. von Cathrine Hornung, Klett-Cotta, Stuttgart 2015.

13
Vgl. M. Rowland, *Can Animals be Moral?*, bes. Kap. 1.

14
Streng genommen handelt es sich um das Ergebnis eines logischen Wenn-Dann-Satzes.

Nutzenden gezielt stimulieren und zumindest billigend eine Schädigung in Kauf nehmen. Damit gilt aber der konstruierende Mensch als moralisches Gegenüber, der adressiert und auf seine Verantwortung angesprochen werden muss – und nicht das technische System.

Kann man sich durch einen Chatbot rassistisch beleidigt fühlen? Würde das als eine psychische Schädigung klassifiziert werden können? Das könnte in der Tat bei Microsofts Chatbot Tay passiert sein.[15] Er wurde vom Betreiber nach nur 16 Stunden wieder abgeschaltet, weil er in kürzester Zeit maschinell gelernt hatte, dass rassistische Äußerungen viel Aufmerksamkeit erregen. Menschen könnten sich von den Äußerungen des Chatbot Tay rassistisch beleidigt fühlen. Dass es sich um rassistische und beleidigende Äußerungen gehandelt hat, steht außer Frage. Ob Menschen sich tatsächlich von diesen Äußerungen beleidigt fühlten, kann ich aufgrund der Datenlage nicht sagen. Noch weniger, ob sie sich durch den Chatbot beleidigt fühlten.

Unter bestimmten Umständen kann hier eine Schädigung von Menschen beobachtet werden. Dem Chatbot kann jedoch weder Vorsatz noch Fahrlässigkeit angelastet werden. Adressat solcher Eigenschaften wäre eindeutig das Unternehmen, das den Chatbot konstruiert und ins Internet gestellt hat. Das Unternehmen könnte seinerseits auf eine Lücke in der Handlungsverantwortung hinweisen, aber das Selbstlernen rassistischer Äußerungen durch die Maschine wäre nicht mit einem Konzept des bewussten, vorsätzlichen Handelns verbunden. Die Verantwortungslücke, die beim Einsatz selbstlernender Systeme entsteht, weil weder der Konstrukteur noch der Nutzer das System vollständig determinieren kann, wird derzeit dadurch – nur unvollkommen – aufgefangen, dass der Mensch als letztverantwortende Instanz angerufen wird.[16]

Gegenwärtig werden bei neueren Chatbots große Anstrengungen unternommen, damit ein solches ‚Fehlverhalten' nicht wieder auftritt. Bei ChatGBT gibt es eine starke menschliche Steuerung im Hintergrund, die solche ‚Ausfälle' verhindern soll. Das zeigt meines Erachtens, dass ein moralisches Problem erkannt und adressiert wird. Die Adresse der ‚moralischen Korrektur' ist jedoch der Mensch und nicht der Chatbot. Dieser erscheint (gegenwärtig noch) nicht so lernfähig, dass er die Beleidigung als solche erkennen und aus moralischen Gründen unterlassen würde – wenn programmierte Regeln für einen Chatbot überhaupt als Gründe angesehen werden können.

5. Ein Roboter beleidigt einen Menschen – ein Gedankenspiel

Macht es einen Unterschied, ob ein Chatbot sich in rassistischer Weise äußert, oder ein Roboter das tut? Ist die moralische Schädigung durch einen Roboter wahrscheinlicher, intensiver – oder trifft eher das Gegenteil zu? Hier gibt es noch keine soliden empirischen Belege, aber es steht zu vermuten, dass die Beleidigung durch einen Roboter, der die ‚sprachliche Nachricht' durch physische Präsenz, Mimik und Gestik unterstützt, wohl einen stärkeren Eindruck beim Menschen hinterlässt. Das ließe sich so interpretieren, dass der Mensch die beleidigenden Äußerungen eines physischen Gegenübers ‚persönlich' nimmt und sie deshalb schwerer wiegen. Andererseits könnte ein Roboter als physisches Gegenüber je nach äußerem Design weniger ernst genommen und so die Beleidigung geringer oder gar nicht empfunden werden. Im Gegensatz dazu ist die Kommunikation mit einem Chatbot, also nur mit einer Stimme, die aus einem Lautsprecher dringt, ungleich virtueller und abstrakter. Anders als

beim Telefonat mit einem Menschen, bei dem wir zur Stimme ein mehr oder weniger konkretes Gegenüber imaginieren können, gibt es hier keinen Körper in einem realen Raum, welche die Kommunikationssituation konkretisieren. Die Stimme kommt aus dem ‚Off', gleichsam aus einer anderen Welt, und das macht es für Menschen – mindestens der aktuellen Generationen – ungleich schwerer, diese Stimme als ein Gegenüber für soziale Aushandlungsprozesse anzusehen und mit ihr entsprechend zu interagieren.

Anders Roboter. Sie treten in ihrer physischen Präsenz als ‚Gegenüber' auf, die nicht nur beleidigen könnten, sondern mit denen die Beleidigung augenscheinlich auch sozial verhandelt werden könnte. Das könnte für den Roboter Sophia gelten, dem von Saudi Arabien die Staatsbürgerrechte verliehen worden sind. Ich möchte in einem Gedankenspiel das Problem weiter explorieren. Würde dieser Roboter sich rassistisch gegenüber einer einzelnen Person oder auch einer Gruppe äußern, und diese Person oder Gruppe fühlte sich in der Folge auch tatsächlich rassistisch beleidigt, so wäre hier die offene Frage, ob diese Bürgerin des Staates Saudi Arabien für diese Normverletzung gegenüber den Beleidigten vor einer zuständigen Instanz zur Verantwortung gezogen werden könnte. Ein mit Staatsbürgerschaft versehener Roboter ist (von den Spezifika des saudi-arabischen Staatsbürgerrechts einmal abgesehen) nicht nur ein technisches Gerät, sondern ein Wesen, dem Rechtsansprüche, aber auch Pflichten zukommen. Die Herstellerfirma Hanson Robotics aus Hongkong als moralischen Ansprechpartner zu wählen, wäre in einem Analogieschluss zu dem Fall des Microsoft-Chatbots Tay naheliegend, verfehlte aber das Problem. In diesem Fall würde man Hanson Robotics als Hersteller und den Roboter als sein Produkt betrachten, für den der Hersteller Gewährleistung zu tragen habe. In unserem Gedankenspiel hat der Roboter als Staatsbürger jedoch Personenrechte und deshalb wird die Beleidigung auf diese Person Sophia zurückgeführt. Nun ließe sich argumentieren, dass es auch andere Staatsbürger oder Staatsbürgerinnen gibt, denen aufgrund fehlender kognitiver Orientierung bestimmte Bürgerrechte eingeschränkt bzw. diese Rechte durch einen bestellten Stellvertreter (Bevollmächtigte, Verfügungsberechtigte) im Sinne des Betreuten wahrgenommen werden. Hanson Robotics hätte hier den Status ‚nächster Verwandter'. Im Falle von Sophia müsste nach dem hier fiktiv gesetzten Fehlverhalten eine ‚Entmündigung' folgen, um der Verantwortung zu entsprechen. Wie diese Entmündigung physisch-technisch aussähe, kann hier zunächst offen bleiben. Vermutlich würde eine Entschuldigung der Firma folgen. Damit träte das Problem umso schärfer hervor: Die Beleidigung wäre in der Welt, die vom Roboter beleidigten Menschen haben einen psychischen und moralischen Schaden erfahren. Sophia zu entmündigen, konzedierte, dass ›sie‹ etwas getan hat, wofür ‚sie' nicht (mehr) verantwortlich sein kann, weil ‚sie' nicht (noch nicht oder nicht mehr?) zureichend orientiert in dieser Welt ist. Ja, es würde signalisieren, dass dieser Roboter (oder solche Roboter)

15
Vgl. Patrick Beuth: „Twitter-Nutzer machen Chatbot zur Rassistin", *Die Zeit* (25. 3. 2016). Hier erhältlich: https://www.zeit.de/digital/internet/2016-03/microsoft-tay-chatbot-twitter-rassistisch (abgerufen: der 26. August 2024); Dave Lee: „Tay: Microsoft issues apology over racist chatbot fiasco", *BBC News* (25. 3. 2016). Hier erhältlich: https://www.bbc.com/news/technology-35902104 (abgerufen: der 26. August 2024).

16
Es geht dabei um die Forderung eines *Human in the Loop*; sehr früh dazu Dieter Sturma: „Ersetzbarkeit des Menschen? Robotik und menschliche Lebensform", *Jahrbuch für Wissenschaft und Ethik* 9 (2004), S. 141–162.

prinzipiell in der Lage ist, sich in dieser Welt so zu verhalten, dass moralische Schädigungen passieren können. Da nun aber Roboter, anders als Menschen, so konstruiert werden sollen, dass Schädigungen vermieden werden, träte hier ein offenkundiger Konstruktionsfehler zutage.

6. Fazit

Die Argumentation könnte allerdings auch anders lauten, dass wir Roboter lernen lassen sollten, so wie auch wir Menschen lernen. In einer fundamentalen Weise gehört es zu unserem menschlichen Zusammenleben, dass wir Fehler und Schädigungen durch andere weder à priori ausschließen können noch wollen. Und so könnte man dann auch mit Robotern verfahren. Damit sie lernen könnten, müssten wir Menschen es in näher zu bestimmenden Maßen aushalten, dass sie Fehler machen, dass sie uns Menschen, vielleicht aber auch andere Tiere oder andere Teile der Umwelt schädigen. Dass wir Menschen neben physischen und psychischen Schädigungen auch moralische Schädigungen davontragen könnten, wäre dann gleichsam der Preis, den wir zahlen müssten für die durch die Roboter hinzugewonnenen Leistungen – worin diese im Bereich der sozialen Robotik auch immer bestehen mögen.

Ich erinnere an dieser Stelle noch einmal an die drei Asimovschen Roboter-Gesetze:

„1. A robot may not injure a human being or, through inaction, allow a human being to come to harm. 2. A robot must obey the orders given to it by human beings, except where such orders would conflict with the First Law. 3. A robot must protect its own existence, as long as such protection does not conflict with the First or Second Law.“[17]

Und Asimov fügt hinzu, dass diese drei Regeln keinesfalls ein ethisches Rahmenwerk ersetzen.

Mit der künstlichen Intelligenz, die ein wesentlicher Baustein von Robotern darstellt, die mit Menschen in eine soziale Interaktion treten, ist immer eine Determinationslücke gegeben. Die künstliche Intelligenz als maschinelles Lernen kommt zu Kalkulationsergebnissen, die sich nicht vorausberechnen oder auch nur *ex post* erklären lassen. Es ist die Frage eines ethischen Rahmenwerks, wie groß und an welchen Stellen unseres gesellschaftlichen Lebens wir diese Lücke zulassen wollen.

Arne Manzeschke

Mogu li roboti moralno povrijediti ljude?

Neka preliminarna etička razmatranja o odnosu čovjek – stroj

Sažetak

U mjeri u kojoj su roboti oblikovani kao partneri u društvenoj interakciji ljudi, postavljaju se ozbiljna moralna pitanja. Članak istražuje pitanje mogu li roboti moralno povrijediti ljude. Razradu pitanja počinjem s (1) kratkom refleksijom o tome što se podrazumijeva pod »roboti«. (2) Skiciram što znači biti moralno povrijeđen. U daljem koraku (3) ispituju se razlike među fizičkom, psihičkom i moralnom nanošenju štete. To nas dovodi (4) do pitanja koja vrsta aktera bi

17
Isaac Asimov, *I, Robot*, Dobson, London 1950.

mogla biti uzročnik takvog nanošenja štete. Stoga, upućujem (5) na jasno ustanovljenu taksonomiju bogova, ljudi, životinja, biljaka i strojeva. Doći ću do zaključka (6) da status sposobnosti za djelovanje nekoga robota zavisi od odluke ljudi o tome kako ga žele konstituirati. Stoga bismo kao ljudska društva trebali biti dobro informirani kada razmišljamo o ovom važnom pitanju. Zaista je važno na koji način donosimo odluke o ovom pitanju koje utječe na naše samorazumijevanje i na strukturu društvenoga života.

Ključne riječi
roboti, interakcija čovjeka i robota, moralno nanošenje štete, oblikovanje tehnike

Arne Manzeschke

Can Robots Morally Harm Human Beings?

Some Preliminary Ethical Considerations on Human-Machine Relations

Abstract
To the extent that robots are designed as social interaction partners of humans, serious moral questions arise. This article explores the question of whether robots can morally harm humans. I begin the elaboration of the question with (1) a brief reflection on what is meant by "robot". (2) I outline what it means to be morally harmed. In a further step (3), the differences among physical, psychological, and moral harm are analysed. This leads (4) to the question of what kind of actor might be the author of such harm. Therefore, I refer (5) to a well-established taxonomy of gods, humans, animals, plants, and machines. I will conclude (6) that the status of a robot's agency depends on how humans choose to construct it. Therefore, as human societies, we should think about this important question in a well-informed way. It really matters how we decide on this question that affects our self-understanding and the fabric of social life.

Keywords
robots, human-robot interaction, moral harm, technology design

Arne Manzeschke

Les robots peuvent-ils nuire moralement aux humains ?

Quelques réflexions éthiques sur la relation homme-machine

Résumé
Dans la mesure où les robots sont conçus pour être des partenaires d'interaction sociale des êtres humains, des questions morales importantes se posent. Cet article analyse la question liée à la possibilité qu'ont les robots de nuire moralement aux humains. Je commence par analyser la question par (1) une brève réflexion sur ce que l'on entend par « robot ». Ensuite, (2) je discute de ce que signifie être moralement lésé. Plus tard, (3) les différences entre préjudice physique, psychologique et moral sont examinés. Cela mène (4) à la question d'identification du type d'acteur qui pourrait être l'auteur d'un tel préjudice. Par conséquent, je me réfère (5) à une taxonomie bien établie des dieux, des humains, des animaux, des plantes et des machines. J'en arrive à la conclusion (6) que le statut d'agentivité d'un robot est dépendant des décisions humaines sur la base desquelles il a été conçu. Par conséquent, en tant que sociétés humaines, il est dans notre devoir de nous informer afin d'être en mesure de réfléchir à cette question importante. La manière dont nous apporterons des décisions sur la question est véritablement cruciale, car elle touche à notre compréhension de nous-mêmes et au tissu de la vie sociale.

Mots-clés
robots, interaction homme-robot, atteinte morale, formation technique