Wie es ist einem autonomen artifiziellen Agenten zu begegnen

Karsten Weber

Lehrstuhl für Allgemeine Technikwissenschaften
Brandenburgische Technische Universität Cottbus
Erich Weinert Str. 1, LG 10, Raum 114
03044 Cottbus
Karsten.Weber@tu-cottbus.de

Abstract: Ausgehend von Thomas Nagels Aufsatz „What is it like to be a bat?“ und Alan Turings Aufsatz „Computing machinery and intelligence“ wird argumentiert, das eine erfolgreiche Interaktion von Menschen und autonomen artifiziellen Agenten vor allem darauf beruht, welche Eigenschaften Menschen jenen Agenten zuschreiben und nicht so sehr, ob diese Agenten jene Eigenschaften wirklich besitzen. Diese Annahme bestätigen sowohl Masahiro Moris Idee des “uncanny valley” ebenso wie zahlreiche empirische Studien. Zum Schluss werden einige der moralischen Konsequenzen des hier Gesagten skizziert.

1 Ausgangspunkt: Erste- und Dritte-Person-Perspektive

In einem seit seiner Publikation im Jahr 1974 oft zitierten Aufsatz mit dem Titel „What is it like to be a bat?“ [Na74] oder in der deutschen Übersetzung „Wie es ist eine Fledermaus zu sein“ ergriff der US-amerikanische Philosoph Thomas Nagel in der Debatte der Philosophie des Geistes Partei und argumentierte für die Irreduzibilität der sogenannten Erste-Person-Perspektive: Das innere mentale Erleben, so Nagel, einer Spinne, einer Fledermaus oder eines Menschen sei grundsätzlich nicht reduzierbar auf rein physikalische Beschreibungen von Prozessen im Gehirn oder im Nervensystem. Dabei ging es Nagel nicht darum, für einen Dualismus von Körper und Geist zu plädieren. Er war jedoch der Ansicht, dass die Existenz des inneren mentalen Erlebens, das subjektive Erleben, das Haben von Qualia – all dies sind Versuche der Umschreibung dessen, worum es geht –, dass all dies ein Faktum in der Welt darstelle, das den gleichen ontologischen Status besäße wie das Faktum, dass, während Menschen solche Qualia haben, bestimmte neurophysiologische Prozesse im Hirn ablaufen. Es mag sein, so würde Nagel vermutlich sagen, dass wir irgendwann wissen werden, dass diese und jene Qualia immer mit diesen und jenen neurophysiologischen Prozessen einhergehen; es mag aber auch sein, dass wir irgendwann wissen werden, dass diese und jene Typen von mentalen Zuständen immer mit diesen und jenen Typen neurophysiologischer Prozesse einhergehen. Es mag also sein, dass wir irgendwann mithilfe einer Theorie der Token- oder aber der Typen-Identität (bspw. [JPP82]) mentale

Zustände erklären und beschreiben können. Das aber, so würde Nagel sagen, bedeute nicht, dass wir Qualia auf neurophysiologische Prozesse reduziert hätten; sie behielten ihren ontologischen Status.

Nun sprach Nagel nicht nur von Menschen, sondern auch von Spinnen und vor allem von Fledermäusen. Er hatte insbesondere diese gewählt, um zu verdeutlichen, dass wir anerkennen müssten, dass Lebewesen, deren Sinneserfahrungen völlig andere als unsere sind, ein mentales Innenleben besitzen, das nicht dadurch vollständig erschließbar wird, dass wir bspw. Fledermäuse sezieren, uns klar machen, wie deren Echolot funktioniert und das Ganze dann in physikalischen Termen beschreiben. Nagel behauptete, dass es auf eine ganz bestimmte Art ist eine Fledermaus zu sein, und dass diese spezifische Erfahrung eine eigenständige Existenz hat und eben nicht restlos reduziert werden kann auf neurophysiologische Prozesse. Um es etwas plastischer auszudrücken: Nur weil wir vielleicht einmal wissen werden, wie wir die Signal- und Informationsverarbeitung in einem Fledermaushirn auf neurophysiologischer Ebene beschreiben und erklären können, wissen wir nicht – und werden wir auch nicht wissen können –, wie es ist, bei Dunkelheit durch die Nacht zu flattern, Insekten zu jagen und am Tag dann mit Hunderten von anderen Fledermäusen mit dem Kopf nach unten in einer Höhle zu hängen. Die spezifische Erlebnisqualität einer Fledermaus als Fledermaus und nicht als Mensch in einer Fledermaussimulation wird uns stets verschlossen bleiben; und doch ist es ein unbestreitbares Faktum, dass es irgendwie ist eine Fledermaus zu sein – nämlich genauso unbestreitbar, wie es ist ein bestimmter Menschen zu sein. Wichtig dabei ist, dass Nagel ein Kontinuum der Befähigung zu einer solchen inneren mentalen Erfahrung annahm – auch die im Baum der Evolution unter uns eingruppierten Tiere, zumindest alle mit einem Nervensystem, seien dazu im Prinzip fähig. Wenn also gilt: „there is something that it is like to be that organism"– es ist irgendwie dieser Organismus zu sein – dann muss man sich möglicherweise darüber Gedanken machen, ob es auch irgendwie sein kann eine Maschine, ein Roboter oder ein autonomer artifizieller Agent zu sein, wenn jene Maschine, jener Roboter oder jener autonome artifizielle Agent über etwas verfügt, das einem Nervensystem analog ist.

Obwohl dieser Gedanke nun verlockend und interessant ist, soll er hier nicht weiter verfolgt werden. Nur so viel sei gesagt: Selbst wenn es gelänge, einen Roboter zu bauen, der sich wie ein Mensch verhielte oder, vielleicht weniger anspruchsvoll, wie eine Katze, so wüssten wir dennoch nichts über das innere Erleben dieser Maschine – denn dem stünde das gerade mit Thomas Nagel Festgestellte entgegen. Entscheidend ist nun jedoch, dass wir dies in den allermeisten lebensweltlichen Kontexten auch gar nicht wissen müssen, um erfolgreich mit anderen Lebewesen oder auch Maschinen zu interagieren. Nagel macht in seinem Text nämlich eine Bemerkung, die er zwar nicht selbst weiter verfolgt, aber die gerade in Bezug auf autonome artifizielle Agenten fruchtbar gemacht werden kann. Denn nach einigen einleitenden Absätzen schreibt Nagel ([Na74], 438, seine Hervorhebung):

„Even without the benefit of philosophical reflection, anyone who has spent some time in an enclosed space with an excited bat knows what it is to encounter a fundamentally *alien* form of life."

Man kann Nagel so verstehen, dass die Dritte-Person-Perspektive auf ein Lebewesen mindestens ebenso wichtig ist wie die Erste-Person-Perspektive – vor allem, wenn es um die Interaktion mit diesem Lebewesen geht. Die nun zu belegende These ist, dass dies nicht nur für den Umgang mit Lebewesen gilt, sondern für alle Entitäten, die bestimmte Eigenschaften zeigen, über die noch zu reden sein wird. Die künstlichen Vertreter dieser Entitäten sollen im Folgenden als autonome artifizielle Agenten (engl.: „autonomous artificial agent“) bezeichnet werden.

2 Sein oder Schein

Um diese These plausibilisieren zu können, ist ein Blick in einen anderen berühmten Aufsatz hilfreich. In seinem Maßstäbe setzenden Aufsatz „Computing Machinery and Intelligence“ aus dem Jahr 1950 [Tu50] schlug der britische Mathematiker Alan M. Turing einen Test vor, mit dessen Hilfe entschieden werden könne, ob eine Maschine intelligent sei bzw. denken würde – allerdings muss man das Wort „denken“ in diesem Kontext in Anführungszeichen setzen. Gelingt es einer Maschine in diesem Test nicht als Maschine identifiziert, sondern für einen Mensch gehalten zu werden, so gilt der Turing-Test als bestanden: Die Maschine ist „intelligent“ oder „denkt“. „Denken“ heißt hier bestimmte kommunikative Fähigkeiten zu besitzen, ohne dass damit gesagt werden würde, wie diese realisiert sein müssen. Mit dem Bestehen des Turing-Tests geht somit nicht einher, dass behauptet wird, dass eine Maschine genauso denke wie ein Mensch, sondern eben nur, dass es dieser Maschine gelingt, in einem Menschen die Überzeugung zu wecken, es mit einem denkenden Wesen zu tun zu haben. Um den Turing-Test bestehen zu können, ist es also nicht wichtig, dass die Maschine tatsächlich denkt, sondern es ist wichtig, dass Menschen die Überzeugung haben, dass sie denkt. „Denken“ ist hier also eine Zuschreibung; der Unterschied zwischen „X denkt“ und „ich glaube, dass X denkt“ ist, so kann man Turing verstehen, im Grunde nicht sinnvoll zu ziehen. Mit anderen Worten: Bezüglich des Denkens ist nicht so sehr die Erste-Person-Perspektive wichtig, sondern die Dritte-Person-Perspektive.

Nun kann man diese Aussage über das Denken auch auf andere psychische Phänomene wie Emotionen, Wünsche, Ziele, Intentionen, Motive und so fort ausweiten und sagen, dass auch der Unterschied zwischen „X hat Gefühle“ und „ich glaube, dass X Gefühle hat“ oder „X hegt Überzeugungen“ und „ich glaube, dass X Überzeugungen hegt“ und vielen anderen ähnlichen psychischen Phänomenen nicht sinnvoll zu ziehen ist, da wir stets nur das beobachtbare äußere Verhalten zur Beurteilung darüber, ob unser Gegenüber denkt, Gefühle hat oder Überzeugungen hegt, heranziehen können. John McCarthy [Mc79] hat diesen Gedanken schließlich in seinen Arbeiten soweit auf die Spitze getrieben, dass er sogar so einfachen Mechanismen wie Heizungsthermostaten Überzeugungen zuschrieb. Man kann dies auch so ausdrücken, dass wir zur Operationalisierung der Messung von Intelligenz oder auch von Emotionen immer nur das äußere Verhalten des jeweiligen Untersuchungsobjekts heranziehen können; hierauf bezogen macht die Unterscheidung von „X denkt“ und „ich glaube, dass X denkt“ aber tatsächlich keinen Sinn mehr, weil wir die Aussage „X denkt“ nur auf Grundlage des äußeren Anscheins von X treffen können. Selbst wenn es also tatsächlich eine ontologische Differenz zwischen „X denkt“ und „ich glaube, dass X denkt“ geben sollte,

kann diese Differenz auf der epistemologischen Ebene nicht nachvollzogen werden (vgl. [Fl08]).

Dem Gesagten folgend könnte daher Thomas Nagel entgegengehalten werden, dass er in dem obigen Zitat einen Fehler in Bezug auf seine eigene Position gemacht habe, denn er spricht dort von der aufgeregten („excited“) Fledermaus. Doch eigentlich müsste er von der „aufgeregt erscheinenden Fledermaus“ sprechen, denn da wir nach seinen eigenen Worten nicht wissen können, wie das innere Erleben der Fledermaus ist, sollten wir auch nicht unterstellen zu wissen, wie es um die Fledermaus tatsächlich bestellt ist. Doch diese Form der Textkritik soll hier nicht weiter verfolgt werden. Wichtig ist stattdessen zu fragen, wie weit solche Zuschreibungen gehen können. Hierzu schreibt nun Claude Draude ([Dr11], 324) in Anknüpfung an E.T.A. Hoffmanns „Der Sandmann“:

„When it comes to the encounter between Nathanael and Olimpia, it is his agency that animates the object. The fact that his lips spread warmth to hers, that the spark of his eyes activates hers, is noteworthy for the field of human-computer interaction: the ability of the user to construct a meaningful scenario should not be underestimated.“

Das ist so zu verstehen, dass es gar nicht so wichtig ist, wie sich ein Sachverhalt wirklich verhält, sondern wie wir eine Situation deuten oder vielleicht auch missverstehen. John McCarthy mag der Überzeugung gewesen sein, dass Heizungsthermostaten wirklich Überzeugungen besitzen – oder auch nicht, das ist gar nicht entscheidend. Wichtig ist, dass Menschen solche Zuschreibungen vornehmen, um Ereignisse und Prozesse, die sie kognitiv ansonsten nicht durchdringen (können), trotzdem verständlich und erklärbar zu machen. Sie nehmen dann einen – wie es der US-amerikanische Philosoph und Kognitionswissenschaftler Daniel Dennett [De94] genannt hat – intentionalen Standpunkt ein in der Interaktion mit einer Entität, um das Verhalten dieser Entität verstehbar und erklärbar zu machen. Etwas plastischer ausgedrückt: Wenige von uns wissen und verstehen, was im Innern all der Geräte abläuft, die wir alltäglich und ständig nutzen. Deshalb neigen viele von uns dazu, diesen Geräten einen eigenen Willen zu unterstellen, um zu erklären, warum diese Geräte zuweilen sich nicht so verhalten, wie wir das wollen. Wir tun das auch bei anderen Interaktionspartnern, bspw. mit unseren Haustieren wie Hunden und Katzen. Wir schreiben ihnen einen, bei Katzen meist durch viele Widersprüche geprägten, Charakter zu. Bei Hunden wiederum, gerade bei jenen, die plötzlich und unvorhergesehen einen Menschen anfallen, neigen wir sogar dazu, sie als bösartig zu bezeichnen und sie damit moralisch zu beurteilen bzw. ihnen ein moralisch verwerfliches Handeln vorzuwerfen; dabei ist schon die Nutzung des Ausdrucks „Handeln“ in diesem Zusammenhang höchst zweifelhaft.

Die Haltung diesen verschiedenen Entitäten gegenüber beruht in erster Linie auf Zuschreibung; ob es tatsächlich so ist, wie wir denken, spielt keine Rolle – zumindest in den allermeisten außerwissenschaftlichen Situationen. Damit soll nun nicht behauptet werden, dass es mit dem entsprechenden wissenschaftlichen Instrumentarium grundsätzlich unmöglich wäre, Genaueres zu erfahren, doch im Rahmen dessen, was im Englischen als „folk psychology“ bezeichnet wird und im Deutschen mit „Alltagspsychologie“ übersetzt werden könnte, ist es unerheblich, ob unser Gegenüber wirklich mentale Zustände hat, wirklich denkt, fühlt, glaubt, wünscht, oder ob wir nur

der Überzeugung sind, dass es so sei; für unser Handeln spielt das keine Rolle. In der Gestaltung von Geräten und insbesondere im Bereich der Mensch-Maschine-Interaktion kann man sich dies nun erfolgreich zunutze machen. Wenn die Designer solcher Geräte es erreichen können, in den Nutzern die Überzeugung zu wecken, dass das Gerät denkt, fühlt, glaubt, wünscht – kurz: Augenscheinlich etwas Ähnliches wie man selbst ist, so sind Interaktionen meist erfolgreicher.

Denn die Nutzer begegnen diesen Geräten dann auch emotional anders. Ja noch mehr; teilweise funktioniert die Interaktion mit entsprechenden Geräten nur noch auf der emotionalen Ebene. Beim GPS-Navigationsgerät im Auto, das uns mit einer angenehmen Stimme den Weg weist, mag dies noch nicht völlig der Fall sein, aber diese Stimme wurde sicher auch nach dem Kriterium ausgesucht, dass sie Vertrauen erwecken kann. Bei der Stimme, die uns dazu anhält uns anzuschnallen, geht es um Überredung (bspw. [IKM06]). Bei bestimmten Therapien mithilfe maschineller bzw. Roboterhilfe sind Emotionen schließlich fast alles, was zählt; ein gutes Beispiel hierfür ist Paro, die künstliche Robbe. Sie wird insbesondere bei der Therapie von Menschen mit demenziellen Veränderungen eingesetzt und soll dazu beitragen, die soziale Isolation dieser Menschen aufzubrechen und die Interaktion zwischen Patienten und Pflegepersonal zu erleichtern (vgl. [TTK05], [KTT06]). Es gibt viele weitere Beispiele für autonome artifizielle Agenten, die mit Menschen – insbesondere Menschen, die an Krankheiten wie Alzheimer, Demenz oder Autismus leiden – interagieren sollen:

„Delirium, dementia, and depression caused by social isolation pose a serious threat to the quality of life of older adults. Robots have the potential to head off or postpone the onset of these conditions by providing cognitive therapy, alleviating loneliness, and encouraging exercise to reduce obesity and improve cardiovascular health.“ ([MVH09], 506)

Im weiteren Zusammenhang der sozialen Interaktion kann zudem auf den Roboterkopf KISMET (vgl. [Br01], [SB05]) oder auf die weitaus komplexeren sogenannten Geminoiden, die von einem Team um Hiroshi Ishiguro (bspw. [IN07]) in Japan gebaut und untersucht werden, verweisen. In diesen und anderen Projekten versucht man Roboter zu bauen, die zur sozialen Interaktion fähig sind und so auf möglichst natürliche Weise bspw. mit Menschen zusammen arbeiten können. Auch in einem anderen Bereich, der Unterhaltungsindustrie und hier speziell das Kino, lässt sich beobachten, dass den in den Filmen auftretenden Robotern durch deren äußere Gestaltung bestimmte Eigenschaften zugewiesen werden, die diese Roboter sympathisch oder unsympathisch machen, in jedem Fall aber eine emotionale Regung provozieren sollen (siehe [Mi09]). Zu nennen sind bspw. Gort, der Roboter aus dem Film „Der Tag an dem die Erde stillstand“ aus dem Jahr 1951, Robby aus „Gefahr im Weltall“ von 1956, die kindlich anmutenden Roboter aus „Lautlos im Weltraum“ aus dem Jahr 1972 oder aber „Wall-E“ aus dem gleichnamigen Film von 2008. In allen diesen Fällen soll das jeweilige Aussehen der Maschine die Beziehung zu den menschlichen Interaktionspartnern betonen und unterstützen: Gort sieht wie ein furchtloser Ritter aus, der Recht und Ordnung verteidigt; Robbys „Gesicht“ soll betonen, dass diese Maschine nichts anderes ist als ein Sklave – sein tumbes Auftreten entspricht den Stereotypen und Klischees des schwarzen Sklaven aus Texten und Filmen dieser Zeit. Das spezifische Aussehen dieser

und anderer Roboter ermöglicht jeweils eine emotionale Beziehung zwischen der Maschine und dem Publikum. Am besten lässt sich die Bedeutung des Aussehens für die Zuschreibung von bestimmten mentalen Eigenschaften an einer Ikone des Films verdeutlichen: das Cyberdyne Systems Model 101 bzw. der T-101 oder einfach der Terminator aus dem gleichnamigen Film von 1984 und den Nachfolgern „Terminator 2: Judgment Day", der 1991 in die Kinos kam, und „Terminator 3: Rise of the Machines" aus dem Jahr 2003. Vor allem am Ende des ersten Films kann man die Bedeutung von Ähnlichkeit insbesondere des Gesichts demonstrieren: Der Terminator, der eine Killermaschine ist, verliert seine gesamte „Haut" und sein Menschengesicht – der Roboter sieht jetzt wie eine Maschine bzw. ein Skelett aus. Diese Entmenschlichung erlaubt es der Protagonistin nun uneingeschränkte Gewaltmittel gegen den Roboter einzusetzen (dieses Schema wird in „Terminator 3: Rise of the Machines" reproduziert, als der „weibliche" Roboter zerstört wird). Anders in „Terminator 2: Judgment Day": Hier ist der T-101 ein Beschützer der Menschen und daher, moralisch gesprochen, eine gute Maschine. Auch dieser Roboter wird zum Schluss des Films hin schwer beschädigt, doch die Maschine behält ihr Gesicht und ihre Menschenähnlichkeit – die Schäden an der Maschine sehen beinahe aus wie Wunden an einem menschlichen Körper. Die positive emotionale Beziehung zwischen dem Roboter und dem Publikum bleibt daher intakt. Somit könnte man argumentieren, dass die Filmemacher vom sogenannten „uncanny valley" profitiert haben.

3 Das unheimliche Tal

Man könnte erwarten, dass es einen positiven Zusammenhang zwischen dem Aussehen und dem Verhalten einer Maschine bzw. eines Roboters und der Zuschreibung eines mentalen Innenlebens gibt, so dass es sowohl notwendig wie auch hinreichend für den Aufbau einer positiven emotionalen Beziehung zwischen einer Maschine und einem Menschen sei die Maschine so menschenähnlich wie möglich zu gestalten; man kann sie bspw. Emotionen zeigen lassen:

„[…] many examples could be given to illustrate the variety of rules that emotion-inspired mechanisms and abilities could serve a robot that must make decisions in complex and uncertain circumstance, either working alone or with other robots. Our interest, however, concerns how emotion-inspired mechanisms can improve the way robots function in the human environment, and how such mechanisms can improve robots ability to work effectively in partnership with people." ([BB04], 273)

Doch trifft jene Annahme nicht zu. Bereits 1970 stellte Masahiro Mori [Mo70] ein Schema vor, mit dem der Zusammenhang zwischen Aussehen und Verhalten einer Maschine und der Zuschreibung von Menschenähnlichkeit beschrieben werden soll. Dieses Schema ist unter der Bezeichnung „uncanny valley" oder „unheimliches Tal" in die Literatur eingegangen. Es zeigt einen verblüffenden Verlauf: Bis zu einem gewissen Punkt gibt es den angesprochenen Zusammenhang, doch dann fällt die Kurve plötzlich ab. Claude Draude ([Dr11], 319) schreibt dazu:

„[…] human likeness evokes trust only up to a certain point. If the robot comes very close to appearing human, but of course is not quite the real thing, minor lapses will produce irritations. On its way to reach the peak of humaneness, the robot falls into the depths of the Uncanny Valley.“

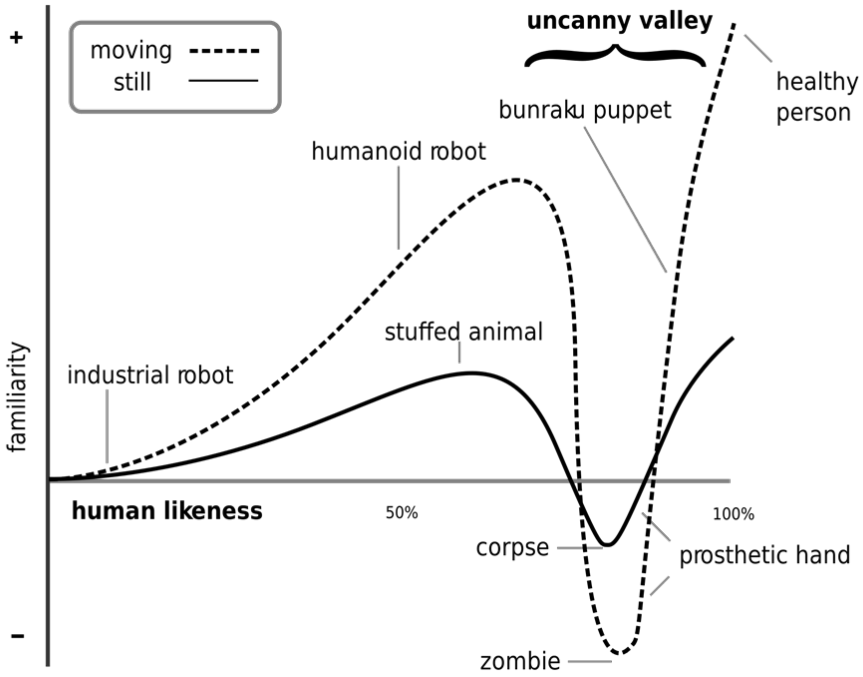

Abbildung 1: The uncanny valley [Mo70]

Folgt man Mori, so ist die Vermutung, dass eine möglichst große Menschenähnlichkeit notwendig und hinreichend für eine erfolgreiche Mensch-Maschine-Interaktion sei, schlicht falsch und möglicherweise sogar kontraproduktiv. Zu einer Zeit, als es angesichts des Stands der Technik wenig Sinn machte, über autonome artifizielle Agenten nachzudenken, verneint Mori die Annahme, dass es einen linearen oder zumindest monoton steigenden Zusammenhang zwischen Menschenähnlichkeit und Vertrautheit von Maschinen bzw. Robotern gäbe. Er behauptet stattdessen, dass viele Objekte trotz deren Menschenähnlichkeit beinahe jede Art der Vertrautheit vermissen ließen. Wie in Abbildung 1 gezeigt, sind dies bspw. Leichen oder Zombies; diese Objekte fallen in das unheimliche Tal und besitzen für Menschen nicht nur keine, sondern sogar so etwas wie eine negative Vertrautheit.

Man kann Mori zudem so verstehen, dass er annimmt, dass alle Menschen auf die gleiche Art reagieren, wenn sie sich mit Maschinen, Robotern oder ähnlichen autonomen artifiziellen Agenten konfrontiert sehen. Ob diese Annahme zutrifft oder ob es Unterschiede gibt, die entweder auf Eigenschaften der Maschinen oder der Menschen,

die auf Maschinen treffen, basieren, ist eine nur empirisch zu beantwortende Frage. Grundsätzlich gilt zu beachten, was Draude ([Dr11], 320) über das unheimliche Tal schreibt: „Mori's concept is discussed controversially; it has been considered non-scientific […] and questionable […] or served as inspiration […].“ Somit macht es Sinn einen genaueren Blick auf empirische Untersuchungen zur Mensch-Maschine-Interaktion und zum unheimlichen Tal zu werfen.

4 Empirie

Tatsächlich zeigen Studien, dass Menschen im Prinzip reagieren, wie der zunächst ansteigende Teil der Kurve in Moris Schema voraussagt. So berichten Slater et al. [SAD06], dass Testpersonen, denen Gewalt einer virtuellen Person bzw. einem sogenannten Avatar gegenüber präsentiert wurde, psychische Reaktionen zeigten, wie sie auch auftreten, wenn Menschen mit Gewalt gegen reale Personen konfrontiert werden – und das, obwohl die Testpersonen wussten, dass es sich um eine virtuelle Figur handelte. Das könnte als Bestätigung für die Vermutung gedeutet werden, dass es für das Design von autonomen artifiziellen Agenten keine Rolle spielt, ob der jeweilige Agent tatsächlich bestimmte mentale Zustände – hier eben Emotionen und Schmerzen – einnimmt oder nur ein Verhalten zeigt, dass so interpretiert werden kann, als ob der Agent diese mentalen Zustände hat – man kann zudem darauf verweisen, dass Untersuchungen mit sogenannten „embodied conversational agents“ (ECA) ebenfalls in diese Richtung deuten (bspw. [ID05]). Sie zeigen, dass der autonome artifizielle Agent nicht verkörpert sein muss; computeranimierte Charaktere – selbst sehr einfache – können ebenfalls starke Emotionen aufseiten der menschlichen Interaktionspartner auslösen (bspw. [RR00]). Allerdings betont Catrin Misselhorn ([Mi09], 346) in einer Diskussion der Ergebnisse von Slater et al., dass die Stärke der empathischen Reaktion bei den Testpersonen bei der Gewalt gegenüber einem künstlichen Agenten niedriger war als im Fall realer Personen.

Misselhorn ([Mi09], 347) schlägt zudem drei Fragen vor, die zukünftige Untersuchungen anleiten sollten: Als erstes sollte geklärt werden, warum wir überhaupt eine gewisse Empathie gegenüber Maschinen mit menschenähnlichen Eigenschaften verspüren. Auf den ersten Blick mag diese Frage trivial klingen, doch bedenkt man die Ergebnisse von Slater et al., so findet sie ihre Berechtigung. Denn offensichtlich wird menschliche Empathie nur wenig durch das Wissen über die Situation bestimmt. Gerade bezüglich der Gestaltung von Maschinen, die mit Menschen interagieren sollen, kann diese Einsicht sehr hilfreich sein, da der Aufwand, den Designer und Hersteller von Robotern, die bspw. in der Pflege alter, hochbetagter oder auch psychisch gehandicapter Personen eingesetzt werden sollen, betreiben müssen, sehr gering sein könnte.

Die zweite Frage ist, warum Menschen aufhören Empathie gegenüber Objekten zu empfinden, wenn diese sehr menschenähnlich sind – dies ist letztlich die Frage, ob und warum das unheimliche Tal überhaupt existiert. Wiederum kann eine Antwort hierauf hilfreich für die Arbeit von Gestaltern und Produzenten autonomer artifizieller Agenten sein. Denn wenn sich ergeben sollte, dass das unheimliche Tal in Wirklichkeit sehr weit ist oder gar eine Ebene ohne erneuten Anstieg, dann würde der Aufwand zur Erzeugung

sehr menschenähnlicher Maschinen kaum Sinn ergeben. Aus ökonomischer Sicht würde es stattdessen ausreichen Maschinen zu entwickeln, die in Moris Diagramm das lokale Maximum auf der linken Seite des unheimlichen Tals erreichen, aber es wäre unnötig, unvernünftig und vielleicht sogar unmöglich zu versuchen, den erneuten Anstieg auf der rechten Seite des Tals zu erreichen.

Als dritte und letzte stellt Misselhorn eine Frage, die mit der zweiten verbunden ist: Sie möchte wissen, warum Menschen nicht einfach nur aufhören Empathie gegenüber sehr menschenähnlichen Objekten zu empfinden, sondern eine seltsame und beängstigende Atmosphäre verspüren (engl.: „eeriness"). Denn von einer praktischen Warte gesehen wird es – abgesehen vielleicht von Geisterbahnen auf dem Jahrmarkt – nur wenige Anwendungsfälle geben, die es sinnvoll erscheinen lassen Maschinen zu gestalten und herzustellen, die solche befremdenden Reaktionen auslösen.

Ebenfalls nicht endgültig geklärt ist, ob der von Mori unterstellte Zusammenhang für alle Menschen auf gleiche Weise zutrifft oder kulturabhängig ist. Ergebnisse einer Studie von Li et al. ([LRL10]) lassen vermuten, dass bspw. Chinesen, Deutsche und Amerikaner ganz unterschiedlich auf die Konfrontation mit Robotern reagieren – nämlich genauso, wie sie auch auf die Interaktion mit anderen Menschen reagieren. Sie schreiben ([LRL10], 177): „[…] German participants greatly preferred the explicit communication style of a robot (e.g. says "I think this choice is not correct" to express disagreement), while their Chinese counterparts preferred an implicit one (e.g. says "Are you sure?" to show disagreement)." Das kann sowohl im Sinne einer Kulturabhängigkeit wie -unabhängigkeit gedeutet werden: Ersteres ergibt sich daraus, dass soziale Interaktionen stets kulturabhängig stattfinden, die zweite Deutung daraus, dass – gegeben, die Maschine folgt den jeweiligen Gepflogenheiten – die Interaktion so ablaufen kann, wie dies auch für die Interaktion zwischen Menschen gilt: Das heißt, dass die Zuschreibung dem Schema Moris entspricht, sofern die Maschine kulturgerecht gestaltet wurde. Doch zu betonen ist, dass hier alles andere als Sicherheit in den wissenschaftlichen Erkenntnissen vorliegt. So verweisen bspw. MacDorman et al. ([MVH09]) darauf, dass Japaner im Vergleich zu Amerikanern anders auf Roboter reagieren. Allerdings betonen sie auch, dass „the cross-cultural similarities in attitudes toward robots were more striking than the differences" ([MVH09], 506); dies würde Moris Annahme einer universalen Gültigkeit bzw. einer kulturellen Unabhängigkeit des unheimlichen Tals bestätigen. Zieht man jedoch Ergebnisse weiterer Studien in Betracht (bspw. [KIR09], [WSD08]), muss man festhalten, dass diese Frage nicht endgültig geklärt ist: Einige Projekte finden kulturelle Unterschiede in Bezug auf die Interaktion von Menschen mit Robotern oder ECAs, andere nicht.

Man kann die Ergebnisse bisheriger Studien vielleicht so zusammenfassen, dass es Belege für die Existenz des unheimlichen Tals gibt, aber dass sich Gestalter und Hersteller von Robotern und ECAs deshalb keine Sorgen machen müssen, da „[…] it is more likely that in the near future relations between humans and robots will mainly take the form of strong attachments to robots that do not appear human" ([Co11], 197). Wie Mark Coeckelbergh betont – und viele der oben genannten Studien bestätigen dies –, müssen Roboter und ECAs für viele Anwendungsfälle gar nicht sonderlich

menschenähnlich sein, um eine erfolgreiche Mensch-Maschine-Interaktion zu ermöglichen.

5 Schlussfolgerungen

Im Gegenteil spricht viel dafür, dass es nur notwendig ist, dass Menschen, die mit Robotern und ECAs interagieren, ein Verhalten zeigen, das Anlass zur Annahme gibt, dass sie glauben, ihre mechanischen Gegenüber hätten tatsächlich Emotionen, Überzeugungen, Intentionen, Wünsche und Ähnliches. Die bisher genannten Studien legen zudem nahe, dass ein eher geringes Maß von Menschenähnlichkeit und Vertrautheit erreicht werden muss, damit Maschinen als autonome artifizielle Agenten akzeptiert werden.

Unglücklicherweise fragen die meisten wenn nicht sogar alle Studien in diesem Zusammenhang jedoch nicht, ob diese Akzeptanz auch die Anerkennung von Maschinen als künstliche moralische Agenten umfasst. Der Ausdruck „künstlicher moralischer Agent“ soll dabei heißen, dass Menschen eher die Maschine für deren Handlungen und den Folgen dieser Handlungen verantwortlich machen als deren Designer oder Hersteller. Wie im schon diskutierten Fall der Emotionen, Überzeugungen, Intentionen, Wünsche u.Ä. wird moralische Verantwortlichkeit hier als Zuschreibung verstanden. Denn zumindest vom Standpunkt jener Menschen, die mit solchen Maschinen interagieren, kann die Frage, ob diese tatsächlich ein moralischer Agent ist oder nur so erscheint, letztlich nur auf der Grundlage von Menschenähnlichkeit, Vertrautheit, sichtbarem Verhalten und ähnlichen beobachtbaren Hinweisen entschieden werden.

Designer und Hersteller von Robotern und ECAs könnten daher versucht sein Maschinen zu gestalten, die von Menschen als künstliche moralische Agenten angesehen werden. Mit Sicherheit werden diese Agenten versagen, Fehler begehen, Besitz zerstören, Menschen schaden und andere moralisch schlechte Dinge tun – ganz genauso wie ordinäre Maschinen das zuweilen ebenfalls tun. Wenn dies geschieht und diese Maschinen als künstliche moralische Agenten akzeptiert würden, machten jene Menschen, die mit ihnen interagieren, vermutlich die Maschinen für ihre Aktionen und deren Folgen verantwortlich – zumindest bis zu einem gewissen Maß. Designer und Hersteller würden sich dann eher in der Position von Eltern als von Ingenieuren wiederfinden. Doch auch ohne tiefe philosophische Reflektion ist es offensichtlich, dass es unangemessen wäre, Hersteller mit Eltern zu vergleichen. Tatsächlich bleibt nur festzustellen, dass derzeit keine auf autonome artifizielle Agenten anwendbaren moralischen Normen existieren. Aber eher früher als später werden solche Normen dringend benötigt: Nicht weil wir bereits in der Lage wären, künstliche moralische Agenten zu bauen, sondern weil sie durch schlichte Zuschreibung erzeugt werden – und das ist sehr einfach. Es ist naheliegend, dass entsprechende moralische und soziale Normen die Designer und Gestalter solcher Maschinen ansprechen müssen, nicht die Maschinen selbst (im Gegensatz dazu [FS01]).

Ähnlich wie Friedman und Kahn ([FK92]) könnte man nun argumentieren, dass hier ein Strohmann verbrannt und ein Scheinproblem aufgeworfen wird, da derzeit künstliche

moralische Agenten nicht existierten, weil heutige Roboter und ECAs keinerlei Intentionen hätten und damit "a necessary condition of moral agency" ([FK92], 9) nicht erfüllten. Dies aber würde an der Argumentation des vorliegenden Texts vorbeigehen. Menschen glauben nämlich viele Dinge, obwohl diese bereits wissenschaftlich als falsch erwiesen wurden. In Bezug auf die Mensch-Maschine-Interaktion kann man sehr leicht Szenarien bspw. für den Bereich des E-Commerce entwerfen, in denen es nützlich wäre, die Überzeugung bei Menschen zu erzeugen oder zu stärken, dass das jeweilige mechanische Gegenüber ein wohlinformierter, moralisch verantwortlicher und vertrauenswürdiger Agent ist, der nur im Interesse der Menschen agiert. Um es direkt zu sagen: Autonome artifizielle Agenten könnten dazu benutzt werden, Menschen zu verführen oder gar zu betrügen (vgl. [SS10]).

Daher sollte folgende moralische Norm stets beachtet werden: Gestalte keine Maschinen, die die menschlichen Interaktionspartner vergessen lassen, dass sie mit Maschinen interagieren (vgl. [CP10], 205). Diese Norm wäre jedoch problematisch, da deren Befolgung es unmöglich machen könnte, Maschinen zu bauen, mit denen Menschen auf möglichst natürliche Art und Weise interagieren können. Da eine solche natürliche Interaktion aber für viele Anwendungen wünschenswert oder gar notwendig ist und sowohl in praktischer Hinsicht als auch von einer moralischen Warte aus als wertvoll angesehen werden kann, sind wir mit sich widersprechenden moralischen Ansprüchen konfrontiert. Bis jetzt bleibt diese Herausforderung ohne Antwort.

Literaturverzeichnis

[BB04] Breazeal, C.; Brooks R.: Robot emotions: A functional perspective. In (Fellous, J.; Arbib, M., Hrsg.): Who Needs Emotions? Oxford University Press, New York, 2004; S. 271-310.

[Br01] Breazeal, C.: Affective interaction between humans and robots. Advances in Artificial Life, LNCS 2159, 2011; S. 582-591.

[Co11] Coeckelbergh, M.: Humans, Animals, and Robots: A Phenomenological Approach to Human-Robot Relations. International Journal of Social Robotics, 3, 2011; S. 197-204.

[CP10] Castellano, G.; Peters, Chr.: Socially perceptive robots: Challenges and concerns. Interaction Studies, 11, 2010; S. 201-207.

[De94] Dennett, D.: Philosophie des menschlichen Bewusstseins, Hoffmann und Campe, Hamburg, 1994.

[Dr11] Draude, Cl.: Intermediaries: reflections on virtual humans, gender, and the Uncanny Valley. AI & Society, 26, 2011; S. 319-327.

[FK92] Friedman, B.; Kahn, P. H. Jr.: Human agency and responsible computing: Implications for Computer System Design. Journal of Systems Software, 17, 1992; S. 7-14.

[Fl08] Floridi, L.: The Method of Levels of Abstraction. Minds and Machines, 18, 2008; S. 303-329.

[FS01] Floridi, L.; Sanders, J. W.: Artificial evil and the foundation of computer ethics. Ethics and Information Technology, 3, 2001; S. 55-66.

[ID05] Isbister, K.; Doyle, P.: The blind men and the elephant revisited. In (Ruttkay, Zs.; Pelachaud, C., Hrsg.): From Brows to Trust: Evaluating Embodied Conversational Agents, Kluwer Academic Publishers, Dordrecht, 2005; S. 3-26.

[IKM06] IJsselsteijn, W. A.; Kort, Y. A. W.; Midden, C.; Eggen, B.; Hoven, E.: Persuasive Technology for Human Well-Being: Setting the Scene. In (IJsselsteijn, W. A.; Kort, Y.

A. W.; Midden, C.; Eggen, B.; Hoven, E., Hrsg.): Persuasive Technology, LNCS 3962, 2006; S. 1-5.
[IN07] Ishiguro, H.; Nishio, Sh.: Building artificial humans to understand humans. Journal of Artificial Organs, 10, 2007; S. 133-142.
[JPP82] Jackson, Fr.; Pargetter, R.; Prior, E. W.: Functionalism and type-type identity theories. Philosophical Studies 42 1982; S. 209-225.
[KIR09] Koda, T.; Ishida, T.; Rehm, M.; André, E.: Avatar culture: cross-cultural evaluations of avatar facial expressions. AI & Society, 24, 2009; S. 237-250.
[KTT06] Kidd, C. D.; Taggart, W.; Turkle, S.: A sociable robot to encourage social interaction among the elderly. Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006; S. 3972-3976.
[LRL10] Li, D.; Rau, P.; Li, Y.: A Cross-cultural Study: Effect of Robot Appearance and Task. International Journal of Social Robotics, 2, 2010; S. 175-186.
[Mc79] McCarthy, J.: Ascribing Mental Qualities to Machines. In (Ringle, M., Hrsg.): Philosophical Perspectives in Artificial Intelligence, Harvester Press, Brighton, 1979; S. 161-195.
[Mi09] Misselhorn, C.: Empathy with Inanimate Objects and the Uncanny Valley. Minds and Machines, 19, 2009; S. 345-359.
[Mo70] Mori, M.: The uncanny valley. Energy, 7, 1970; S. 33-35.
[MVH09] MacDorman, K.; Vasudevan, S.; Ho, Ch.-Ch.: Does Japan really have robot mania? Comparing attitudes by implicit and explicit measures. AI & Society, 23, 2009; S. 485-510.
[Na74] Nagel, Th.: What is it like to be a bat? The Philosophical Review, 83, 1974; S. 435-450.
[RR00] Rickenberg, R.; Reeves, B.: The effects of animated characters on anxiety, task performance, and evaluations of user interfaces. Proceedings of the SIGCHI'00 conference on Human factors in computing systems, ACM, New York, 2000; S. 49-56.
[SAD06] Slater, M.; Antley, A.; Davison, A.; Swapp, D.; Guger, Chr.; Barker, Chr.; Pistrang, N.; Sanchez-Vives, M. V.: A Virtual Reprise of the Stanley Milgram Obedience Experiments. PLoS ONE, 1, 2006; S. e39.
[SB05] Stiehl, W. D.; Breazeal, C.: Affective touch for robotic companions. Affective Computing and Intelligent Interaction, LNCS, 3784, 2005; S. 747-754.
[SS10] Sharkey, N.; Sharkey, A.: The crying shame of robot nannies: An ethical appraisal. Interaction Studies, 11, 2010; S. 161-190.
[TTK05] Taggart, W.; Turkle, S.; Kidd, C. D.: An interactive robot in a nursing home: Preliminary remarks. Toward Social Mechanisms of Android Science. Cognitive Science Society, Stresa/Italy, 2005; S. 56-61.
[Tu50] Turing, A. M.: Computing Machinery and Intelligence. Mind, 54, 1950; S. 433-457.
[WSD08] Walters, M.; Syrdal, D.; Dautenhahn, K.; te Boekhorst, R.; Koay, Kh.: Avoiding the uncanny valley: robot appearance, personality and consistency of behavior in an attention-seeking home scenario for a robot companion. Autonomous Robots, 24, 2008; S. 159-178.