Zur Ethik künstlicher Intelligenz – philosophische Untersuchung der Möglichkeit moralisch handelnder und ethisch denkender technischer Systeme

Dissertation
zur Erlangung des Grades eines Doktors
der Philosophie (Dr. phil.)

der Pädagogischen Hochschule Ludwigsburg

vorgelegt von Leonie Seng aus Stuttgart

Ludwigsburg
2023

Erstgutachter:
Prof. Dr. phil. habil. Matthias Rath
Pädagogische Hochschule Ludwigsburg
Institut für Theologie und Philosophie

Zweitgutachterin:
PD Dr. Jessica Heesen
Eberhard Karls Universität Tübingen
Internationales Zentrum für Ethik in den Wissenschaften

Datum des Abschlusses der mündlichen Prüfung: 11. Januar 2023

Inhaltsverzeichnis

Danksagung

Allen voran danke ich meinem Erstbetreuer, Prof. Dr. phil. habil. Matthias Rath, und meiner Zweitbetreuerin, PD Dr. Jessica Heesen, für Ihre Geduld, ihre ebenso fachwissenschaftliche wie persönliche Unterstützung sowie ihr unermüdliches Engagement bei der Begleitung des Entstehungsprozesses der vorliegenden Arbeit.

Ich danke der Pädagogischen Hochschule Ludwigsburg für die großzügige finanzielle Unterstützung mittels eines Stipendiums zum Abschluss der Promotion von April 2021 bis Juni 2021. Vom Karlsruher Institut für Technologie, insbesondere durch Prof. Dr. Armin Grunwald, Dr. Linda Nierling, Constanze Scherz und Dr. Ulrich Ufer vom Institut für Technikfolgenabschätzung und Systemanalyse, habe ich große Unterstützung bei der Fertigstellung der Dissertation erfahren.

Mein besonderer Dank gilt darüber hinaus denjenigen, die den Werdegang meiner Dissertationsschrift eng begleitet und immer wieder Textteile kritisch gelesen und kommentiert haben. Ich danke allen Kolleg∗innen[1], egal, ob vom Fach oder fachfremd, die mit mir diskutiert und meine Gedanken um kluge Anregungen bereichert haben. Der fachwissenschaftliche Austausch mit Kolleg∗innen bei Vorträgen ebenso wie Vorträge vor Laienpublikum und Diskussionen mit Studierenden haben meine Arbeit um viele gute Ideen und kritische Nachfragen erweitert. Frau Dipl.-Päd. Adeline Schach bin ich sehr dankbar für ihre professionelle Unterstützung und ihre hilfreichen Ratschläge.

Von Herzen danke ich meiner Familie und guten Freund∗innen, die mich während des gesamten Schreibprozesses begleitet und in sämtlichen Phasen mit Verständnis, Interesse und Zuspruch unterstützt haben.

[1] Mit dem Genderstern wird im Folgenden versucht, der sprachlichen Abbildung möglichst vielfältiger Identitätsvarianten und einer möglichst inklusiven Ausdrucksweise Rechnung zu tragen. Mir ist bewusst, dass diese Schreibweise von der aktuellen Regelung gemäß Duden abweicht und aus verschiedenen Gründen auch umstritten ist (vgl. z. B. Witte 2021: 119–153). Der Genderstern ist daher nur eine Variante, womit in der vorliegenden Arbeit versucht wird, eine möglichst gendersensible Sprache zu realisieren. Für die bessere Lesbarkeit verwende ich die Schreibweise auf x-Höhe (vgl. ebd.: 148–152).

1 Vorwort

Die folgenden Zitate aus unterschiedlichen Entstehungszeiten und Disziplinen dienen dem Zweck, das Spannungsfeld, in dem sich die vorliegende Arbeit bewegt, aufzumachen. Welche der folgenden Thesen ich dann mit welcher Begründung auch in meiner Arbeit vertrete oder zu widerlegen versuche, wird in den folgenden Kapiteln deutlich werden.

> „Eine moderne Grafikkarte schafft 10.000.000.000.000 Rechenoperationen pro Sekunde, doch reicht das aus, um *ausrechnen* zu lassen, ob mein Gegenüber ein feindlicher Kämpfer, ranghoher Terrorist oder einfach nur ein spielendes Kind ist?“
> (Ullrich 2019b: 253. Herv. i. Orig.)

> „Don't seek to build ethical robots; seek to build robots ethically.“
> (Chrisley 2020: 473)

> „So, the central question of machine ethics asks whether the machine could exhibit a simulacrum of ethical deliberation. It will be no slight to the machine if all it achieves is a simulacrum. It could be that a great many humans do no better.“
> (Powers 2011: 464)

> „Bei unartigen oder schwer erziehbaren Robotern ist von körperlicher Züchtigung abzuraten [...]. In Notfällen erweisen sich Kneifzangen als wirksam.“
> (Loriot 1972b: 67)

2 Einleitung

> And can machines ever really create new things?
> Initial forays into artificial creativity have been made, but the machine is only rolling dice to arrive at possible solutions to known problems and then asking people if the solution is any good.
> (Ramge 2019: 105 f.)

Moralische Maschinen und so genannte *künstliche Intelligenz* (KI)[2] werden seit einigen Jahrzehnten vielfach diskutiert; in wissenschaftlichen Kontexten verschiedener Disziplinen ebenso wie im Alltag und in öffentlichen Medien. Eine zentrale Frage der analytischen Metaethik, mit der ich mich in der vorliegenden Arbeit unter anderem beschäftigen werde, ist, wie moralische Maschinen im Detail definiert werden und welche (primär ethischen) Konsequenzen daraus folgen.

Im Englischen ist auch häufig von *ethical machines* die Rede, also von ethischen Maschinen[3]; der Ausdruck wird dort meist äquivalent zu *moral machines* verwendet.[4] Im deutschsprachigen Raum werden die Begriffe *Moral* und *Ethik* jedoch im wissenschaftlichen Gebrauch eindeutig voneinander getrennt: Unter Ethik wird in der Regel die wissenschaftliche Disziplin verstanden, welche die vielen Aspekte von Moral – und im Wesentlichen moralisches Handeln[5] – zum Gegenstand hat.[6]

Ausgangspunkt für die vorliegende Arbeit war die folgende Überlegung: Was passiert, wenn man die englische Redeweise von *ethical machines* mit der Definition von *Ethik* als wissenschaftliche Disziplin kombiniert? Zunächst scheint es ebenso sinnvoll zu sein, von genuin *ethischen* Maschinen zu sprechen wie von *biologischen*

[2] Die Ausdrücke *Maschine*, *Algorithmus*, *Computerprogramm*, *Roboter* werden häufig äquivalent verwendet. In Abschnitt 2.3.1 erläutere ich die in der vorliegenden Arbeit verwendete Terminologie sowie die zugrundeliegenden technologischen Unterschiede in dem Rahmen, der für das Verständnis der Argumentation notwendig ist. Die Begriffe *Technik* und *technische Systeme* werden dabei als Überbegriffe verstanden, *Maschine* und *Computerprogramme* als technische Spezifizierungen, die in Kombination mit lernenden Systemen (Stichwort *maschinelles Lernen*) ethisch relevant sein können. Unter *künstlicher Intelligenz* wird häufig vor allen Dingen Letzteres verstanden; mehr zum Stichwort *Intelligenz* in Abschnitt 5.2.6.

[3] Im Folgenden wird entweder die Rede von *ethischen technischen Systemen*, oder *ethischen Techniksystemen* sein; alle Ausdrücke bezeichnen denselben Gegenstand, nämlich genuin ethische Techniksysteme.

[4] Vgl. hierzu u. a. Wallach und Allen 2009: 6.

[5] Vgl. u. a. Lutz-Bachmann 2013b. Es wird aber auch zwischen *moralisch sein* und *moralisch handeln* unterschieden, s. hierzu Abschnitt 4.1.

[6] Wie im Folgenden deutlich werden wird, kann Ethik dabei auch unabhängig von der wissenschaftlichen Disziplin über die moralische Fragen *Was soll ich tun?* definiert werden. Mein engeres Verständnis von Ethik für die vorliegende Arbeit schließt jenes weitere nicht prinzipiell aus.

Wissenschaftler∗innen oder *mathematischen* Professor∗innen. Die Aussage der aus den jeweiligen Disziplinen generierten Adjektive kann jedoch – angesichts der Tatsache, dass die Ausdrücke *ethische Maschine* und *ethical machines* nun einmal existieren und verwendet werden[7] – dann sinnvoll sein, wenn darunter die Zuschreibung der Fähigkeiten verstanden wird, derer es bedarf, um in der jeweiligen Disziplin tätig zu sein, beziehungsweise sich mit den in der Disziplin relevanten Fragen zu beschäftigen. Ein *ethischer Mensch*[8] ist demnach einer, der in der Lage ist, ethische Argumente und Theorien aufzustellen, diese zu reflektieren und plausibel zu begründen, Positionen gegeneinander abzuwägen und – je nach Spezialisierung – deskriptiv oder normativ im Feld der wissenschaftlichen Ethik zu arbeiten. Ein moralischer Mensch ist im Gegensatz dazu einer, der sich in einem bestimmten Kontext an geltende, ethisch begründete Normen oder Richtlinien hält.[9]

In Anlehnung an die obige Begriffskombination stellt sich folglich die technikethisch entscheidende Frage: Sind *ethische Maschinen* (in jenem Sinn von Ethik) möglich? Die im Titel formulierte „Ethik von Maschinen“ ist demnach nicht als Ethik von Menschen *für* (potenziell moralische) Maschinen zu verstehen, sondern als *Ethik einer Maschine selbst*, also einer Maschine, die alle Voraussetzungen erfüllt, die auch bei Menschen vorausgesetzt werden, um im Feld der Ethik tätig sein bzw. ethische Fragen reflektieren und beantworten zu können.[10] Matthias Rath beschreibt die Anforderungen an ethische Maschinen so:

[7] Gemäß der ersten Sätzen in Ludwig Wittgensteins Tractatus logico-philosophicus (S. 11): „I Die Welt ist alles, was der Fall ist. I.I Die Welt ist die Gesamtheit der Tatsachen, nicht der Dinge.“ Ich verstehe die Sätze in Anwendung auf meine Thematik so, dass die Ausdrücke *ethische Maschine* und *ethical machines* einer Auseinandersetzung bedürfen (beziehungsweise diese legitim ist), aufgrund der Tatsache, dass es sie gibt und dass sie verwendet werden – unabhängig davon, ob sie auf dem weißen Papier, also unabhängig von der tatsächlichen Verwendung, als sinnvoll erachtet werden können und sollten oder nicht. Andere Autor∗innen kommen zu dem Schluss, dass die Rede von ethischen Maschinen *nonsensical* ist, „because the question attempts to endow a computer with some metaphysical qualities which, because of its nature, a computer does not have and considering what computing and ethics are, in high probability never will have“ (vgl. Krzanowski et al. 2016: 1). Welche Voraussetzungen für die Ausübung der Disziplin der Ethik genau benötigt werden, ist eine Frage, die ich immer wieder im Verlauf meiner Arbeit aufgreifen werde; eine Analyse und Diskussion einzelner, u. a., kognitiver Qualitäten findet insbesondere in Kapitel 5 statt.

[8] Die Rede von *ethischen Menschen* ist unüblich und wird hier im Sinn eines Gedankenexperiments verwendet, in dessen Rahmen überlegt werden soll, was es bedeuten könnte, ein Ethik auch adjektivisch zu verstehen.

[9] Auch zum Begriff der Moral und insbesondere der Definition von *moral machines* gehen die Ansichten und Konzepte auseinander. Ich werde in Kapitel 4 genauer darauf eingehen.

[10] Auf die verschiedenen Verständnismöglichkeiten von Maschinenethik gehe ich später noch genauer ein, s. S. 30.

> Als *ethische* Agenten hingegen müssten sie [Maschinen; L. S.] im Stande sein, Regeln nicht nur zu befolgen, sondern sie auch zu *verstehen*. Verstehen meint hier, diese in einem größeren Sinnzusammenhang einzuordnen.
> (Rath 2019: 237. Herv. i. Orig.)

An diesem Punkt stellen sich viele Fragen. Die erste Herausforderung wird sein, zu spezifizieren, was Ethik bei Menschen bedeutet beziehungsweise welche Anforderungen wir an ethische Maschinen, die nicht nur moralisch handeln, sondern auch moralische Werte reflektieren und ethische Prinzipien generieren können sollen, stellen wollen. Dies setzt die Definition von Moral bei Menschen und moralischen Maschinen voraus, die ich jenen Überlegungen voran stellen werde.

Daneben besteht eine zweite Herausforderung darin, zu klären, welche Konsequenzen aus dem hypothetischen Konzept der ethischen Maschine folgen. Unabhängig davon, ob man zu dem Schluss kommt, dass es, gemäß derzeitigem Stand der Technik, ethische Maschinen geben kann oder nicht, sehen wir uns in der heutigen Welt mit zahlreichen, im weitesten Sinn, technischen Entwicklungen konfrontiert, die immer wieder ethische Expertise auf den Plan rufen. Gleichzeitig werden technische Systeme im Hinblick auf verschiedene Fähigkeiten derart weiterentwickelt, dass eine technisch-ethische Unabhängigkeit vom Menschen nicht mehr unmöglich erscheint. Dies rechtfertigt nicht nur die Überlegungen, sondern macht sie geradezu notwendig, da in der heutigen Zeit Entscheidungen für zukünftige Entwicklungen getroffen werden müssen.[11]

Grundlegend gibt es mindestens zwei Herangehensweisen an die Frage, wie Ethik in technische Systeme hineinkommt. Die eine geht von einer prinzipiellen Ethik von Menschen aus und versucht, diese auf Maschinen anzuwenden – oft ist auch die Rede davon, dass bestimmte Werte oder die Fähigkeit, diese abzuwägen in Maschinen *implementiert* werden soll. Dieses Vorgehen läuft in der Regel darauf hinaus, dass die Ansprüche, gemessen am heutigen Stand der Technik, sehr hoch sind und einige Autor∗innen zu dem Schluss kommen, dass es ethische Maschinen gar nicht geben kann, weil diese demnach eigentlich Menschen sein müssten.

> An important question in this context is whether computers can question rules – and thus themselves – like a critically thinking human being can. A strong artificial intelligence would have to master this trick in order to continuously reinvent itself in

[11] Welche Bedeutung die systematische Reflexion eigenständig *denkender* Programme und Algorithmen insbesondere in Situationen technischer Unsicherheit spielen können, ist von dem im August 2022 erscheinenden Buch *Algorithms for Decision Making* von Merkel Kochenderfer, Tim Wheeler und Kyle Wray zu erwarten. Cambridge, MA: MIT Press, https://mitpress.mit.edu/books/algorithms-decision-making, aufgerufen am 01.03.2022.

> the same way that humans have been for thousands of years. And can machines ever really create new things? Initial forays into artificial creativity have been made, but the machine is only rolling dice to arrive at possible solutions to known problems and then asking people if the solution is any good. In this realm, it is not apparent that current AI research could endow machines themselves with the ability to be truly innovative without a human being first defining the problem. This would suggest that there might be a limit to strong AI's capacity to evolve.[12]
>
> (Ramge 2019: 105 f.)

Auch Rath (2019: 237) meint, dass die oben zitierte Fähigkeit, „Regeln nicht nur zu befolgen, sondern sie auch zu *verstehen*“ (Herv. i. Orig.) selbst „mit einer auch flexiblen Regelprogrammierung – bislang zumindest noch – nicht zu leisten[. . .]“ sei. Ein Problem, das dieser Annahme innewohnt, ist, dass viele unterschiedliche Vorstellungen von *Ethik* schon in Bezug auf Menschen in verschiedenen Ansätzen kursieren; dies macht es schwer, eine adäquate Umsetzung von genuiner Ethik in technischen Systemen zu entwickeln. Es ist also sinnvoll, zunächst zu definieren, aufgrund welcher Theorien die Voraussetzungen für Ethik bestimmt werden, wie plausibel diese sind und welche Konsequenzen sich daraus ergeben – zunächst für Menschen, dann für Maschinen. Häufig werden – anscheinend, um Maschinen, aus verschiedenen Gründen, dezidiert von Menschen abzugrenzen[13] – besonders hohe Ansprüche für ethische Kompetenzen vorausgesetzt, die jedoch noch nicht einmal sinnvollerweise für Menschen angenommen werden können. Im Zusammenhang mit der Definition von Moral und moralischen Maschinen werde ich die (mittlerweile häufig zitierten) Überlegungen James Moors genauer darlegen und diskutieren, der beispielsweise hiervon ausgeht: „A full ethical agent can make explicit ethical judgments and generally is competent to reasonably justify them. An average adult human is a full ethical agent“ (Moor 2006: 6).

Eine zweite Herangehensweise an genuine Maschinenethik besteht darin, in Analogien zu denken, also mögliche limitierende Faktoren technischer Systeme (zumindest im direkten Vergleich mit Menschen) von vorneherein mitzudenken und mögliche Umsetzungen ethischer Fähigkeiten und Eigenschaften daran anzupassen. Thomas Powers schlägt zum Beispiel vor, nicht von technischen Systemen auszugehen, die *selbst* ethisch denken, sondern die vielmehr ethische Überlegungen von Menschen *simulieren* können. Dies kommt einer Form von *schwacher KI*, wie in Fußnote 12 beschrieben, nahe:

[12] Vertreter *starker* künstlicher Intelligenz „streben die vollständige maschinelle Nachbildung psychischer Prozesse wie Denken, Lernen oder Problemlösen an“ (Seng 2019a: 2) oder halten diese zumindest für möglich. Hinter dem Ausdruck *schwache künstliche Intelligenz* werden „bereits an menschliche Intelligenz angelehnte Teillösungen“ (ebd.) als technische Imitation menschlicher Fähigkeiten akzeptiert.

[13] Vgl. hierzu auch Selinger 2021 Mögliche Gründe hierfür erläutere ich in Abschnitt 3.1.

> We get closer to machine ethics when the tool is a computer that's programmed to effect good as a result of the programmer's intentions. But to be ethical in a deeper sense – to be ethical in themselves – machines must have something like practical reasoning that results in action that causes or avoids morally relevant harm or benefit. So the central question of machine ethics asks whether the machine could exhibit a simulacrum of ethical deliberation.
>
> (Powers 2011: 464)

Dem ersten Ansatz liegt eine Identitätsbehauptung zugrunde, dem zweiten eine Analogiebehauptung. Gemäß Konzeptionen, denen die zweite Annahme zugrundeliegt, können Maschinen also allenfalls *menschenähnliche* Formen von Denken, Bewusstsein usw. hervorbringen. Der zweite Ansatz scheint aus mehreren Gründen sowohl argumentatorisch-begrifflich plausibler als auch technisch praktikabler zu sein. Die Gründe dafür werden im Lauf der Arbeit, und insbesondere bei der Erörterung der Möglichkeit ethischer Maschinen, ausführlich dargelegt. Doch auch bei der Verfolgung des zweiten Ansatzes bleibt die Anforderung einer Definition dessen, was menschliche ethische Fähigkeiten ausmacht, und welche ethischen Fähigkeiten für Maschinen gesellschaftlich erwünscht sind (und welche nicht), nicht aus.

> Technische Innovationen schaffen häufig neue Handlungsräume, für die bekannte ethische Verhaltensformen keine Antwort geben. Sie müssen also neu diskutiert, reflektiert und ausgehandelt werden. Um diesen Prozess gesellschaftlich für KI voran zu bringen [sic!], müssen ethische Leitlinien erstellt und vor allem auf ihre praktische Umsetzbarkeit geprüft werden.
>
> (Heesen 2021b: 48)

Ethische Leitlinien werden derzeit einerseits an vielen Stellen gefordert; andererseits stehen sie unter anderem aufgrund ihrer (zu) stark limitierenden Auswirkungen auf technische und wirtschaftliche Prozesse bei vielen Personen, die mit direkten Anwendungen und möglichen Umsetzungen genuin ethischer Maschinen arbeiten, nicht besonders hoch im Kurs. Thilo Hagendorff kommt nach einer systematischen Analyse von 22 Ethikrichtlinien zu dem Schluss:

> Currently, AI ethics is failing in many cases. Ethics lacks a reinforcement mechanism. Deviations from the various codes of ethics have no consequences. And in cases where ethics is integrated into institutions, it mainly serves as a marketing strategy. Furthermore, empirical experiments show that reading ethics guidelines has no significant influence on the decision-making of software developers. In practice, AI ethics is often considered as extraneous, as surplus or some kind of ‚add-on' to technical concerns, as unbinding framework that is imposed from institutions ‚outside' of the technical community. Distributed responsibility in conjunction with a lack of knowledge about long-term or broader societal technological consequences

> causes software developers to lack a feeling of accountability or a view of the moral significance of their work. Especially economic incentives are easily overriding commitment to ethical principles and values. This implies that the purposes for which AI systems are developed and applied are not in accordance with societal values or fundamental rights such as beneficence, non-maleficence, justice, and explicability [...].“
>
> (Hagendorff 2020: 113 f.)

in Bezug auf ethische Maschinen stellt sich die Frage, inwieweit aktuelle Ethikrichtlinien mit jenen vereinbar sein können, was ich in Abschnitt 6.3 untersuchen werde. Denn selbst, wenn man, wie Moor (2006: 7), zu dem Schluss kommt, dass „[... w]e won't resolve the question of whether machines can become full ethical agents by philosophical argument or empirical research in the near future“, so gibt es bei heutigen technischen Entwicklungen bereits viele ethische Fragen, die zu klären sind – die Möglichkeit ethischer Maschinen ist angesichts des aktuellen technischen Fortschritts mindestens möglich, wenn nicht sogar naheliegend.

Was in Diskussionen um KI-Ethik häufig nicht beachtet wird: Dabei geht es nicht nur um die Fragen nach dem – an bestimmten gesellschaftlichen Konventionen gemessenen – moralisch richtigen Verhalten, sondern auch um den ursprünglichsten Zweck von Ethik: gelingende, menschliche Lebensweisen (s. hierzu auch Abschnitt 4.3). Ob Ethik auch bezüglich ethischer Maschinen ein von Menschen für Menschen gemachtes Konstrukt bleibt, das uneingeschränkt und vorrangig ihrem eigenen Wohl dient, wird zu diskutieren sein.

2.1 Überblick über die Kapitel

Im Folgenden werde ich zunächst in Kapitel 3 einen kurzen Überblick über die Geschichte maschineller Moral und Ethik (sofern in letzterem Fall möglich) geben sowie einige technikanthropologische Aspekte anführen. Die Überlegungen zu Technik und dem damit verbundene Menschenbild allgemein (Abschnitt 3.1) münden dabei in Überlegungen, was Technikentwicklungen für Menschen im Lauf der Zeit bedeutet haben und wie jene bewertet werden können (Abschnitt 3.2). Anhand des Beispiels *Sprache* (Abschnitt 3.3) zeichne ich eine spezifische Technikentwicklung nach, um die Relevanz der Technikgeschichte für die Beurteilung möglicher zukünftiger Entwicklungen zu verdeutlichen. Die Zukunftsvorstellungen (Abschnitt 3.4) können dabei in engem Zusammenhang mit Science-Fiction-Narrativen (Abschnitt 3.5) gelesen werden.

In Kapitel 4 folgt die philosophische Analyse des Begriffs *moralische Maschine*. Nach den Fragen *Was ist Moral?* (Abschnitt 4.1) und einer Analyse der Möglichkeit moralisch handelnder Maschinen (Abschnitt 4.2) folgt ein Abschnitt, in dem ich auf das Verständnis von Moral als Bedingung für ein *gutes Leben* einge-

hen werde (Abschnitt 4.3). Anschließend folgen ein paar Beispiele für moralische Akteure (Abschnitt 4.4) sowie einige Beispiele für mögliche Risiken moralischer Akteure (Abschnitt 4.5). In Abschnitt 4.6 werde ich ein scheinbares Paradox bezüglich moralischer Maschinen diskutieren, dann die Bedeutung der Inter- und Transdisziplinarität für das Themenfeld ansprechen (Abschnitt 4.7) und schließlich die Ergebnisse des Kapitels in einer Zusammenfassung festhalten (Abschnitt 4.8).

In Kapitel 5 stehen theoretische Ansätze zur Umsetzung technischer Systeme als moralische und potenziell ethische Akteure im Vordergrund. In Abschnitt 5.1 werden hierzu zunächst grundlegend *bottom-up-* und *top-down*-Ansätze vorgestellt. Anschließend werden die Grundlagen für die Untersuchung der Notwendigkeit einiger Begriffe für sowohl menschliche als auch technische ethische Akteure gelegt (Abschnitt 5.2). Daran schließt im selben Abschnitt die Diskussion und Analyse zentraler Begriffe und ihrer Notwendigkeit für ethische Systeme an. Nach einem Zwischenfazit in Abschnitt 5.3 stelle ich dann einige ethische Theorien und ihre mögliche Anwendung auf technische Systeme vor (Abschnitt 5.4) und fasse das Wesentliche des Kapitels anschließend kurz zusammen (Abschnitt 5.5).

In Abschnitt 6 werde ich neben einigen Beispielen für ethische Techniksysteme (Abschnitt 6.1) einige Ergänzungen zur konkreten technischen Realisierung anführen (Abschnitt 6.2) und anschließend das Konzept jener anhand einiger Aspekte des *Weißbuchs zur Künstlichen Intelligenz* der Europäischen Kommission diskutieren (Abschnitt 6.3). In Kapitel 7 folgen schließlich ein Fazit und Ausblick.

2.2 Prélude: Kann eine KI eine Dissertation schreiben?

Bemerkung: Die vorliegende Dissertationsarbeit wurde am 15. März 2022 an der Pädagogischen Hochschule in Ludwigsburg eingereicht. Die Anfang des Jahres 2023 beginnenden Diskussionen um das Programm ChatGPT konnten daher nicht berücksichtigt werden; Vorgängerversionen des Programms und der technische Stand bis dato werden natürlich einbezogen.

Angenommen, es bedürfe der im ersten Einleitungsteil angesprochenen Fähigkeiten, um eine wissenschaftliche Abschlussarbeit im Fachgebiet der Ethik zu schreiben – logische Argumentation, Reflexion moralischer Werte und ethischer Theorien, Abstraktionsvermögen usw. –, dann lässt sich die Frage, ob es ethische Maschinen geben kann auch etwas plakativer umformulieren: Kann ein Algorithmus oder eine Maschine eine Dissertation schreiben? Diese Frage werde ich im folgenden Abschnitt hypothetisch und als Vorbereitung für den Hauptteil diskutieren. Die Richtigkeit der Analogie von *ethisch sein* und *eine Dissertation im Fach Ethik schreiben*, kann natürlich bestritten werden. Allenfalls kann *eine Dissertation im Fach Ethik schreiben* als eine mögliche Form des *Ethischseins* verstanden werden,

die weder notwendig noch hinreichend für dieses ist.

Ist es möglich beziehungsweise wahrscheinlich, dass dieser Text von einem Computerprogramm geschrieben worden ist? Spontan ist das gar nicht so abwegig in Zeiten scheinbar omnipräsenter und offenbar multipotenter Algorithmen und angesichts der Rede von *künstlicher Intelligenz*. Schauen wir uns daher zunächst genauer an, von wem Dissertationen normalerweise geschrieben werden (der Einfachheit halber beschränke ich mich hier auf Deutschland, wobei die formalen Anforderungen selbst dort stark variieren): Menschen schreiben Dissertationen, die in der Regel eine bestimmte akademische Laufbahn durchschritten haben – im Fall der Disziplinen Philosophie und Ethik handelt es sich um Personen, die meistens akademisch in den Bereichen Philosophie und Ethik ausgebildet wurden. Rein formal müssen diese Personen ein Diplom- oder Masterstudium abgeschlossen haben. Dieses erreicht man durch den Erwerb von ECTS-Punkten durch das Besuchen und Bestehen von Seminaren, Vorlesungen und Tutorien (ggf. auch durch andere Aktivitäten; die Anforderungen sind von Universität zu Universität zuweilen recht unterschiedlich) sowie durch das Schreiben – und manchmal auch mündliche Verteidigen – einer Abschlussarbeit. Diese muss (auch da gehen die Konditionen zuweilen auseinander) eine bestimmte Seitenzahl umfassen,[14] und sich wissenschaftlich mit einem bestimmten Thema befassen. *Wissenschaftlich* heißt im Fall der Philosophie, dass Methoden der Philosophie und Ethik angewandt, Thesen aufgestellt und diese begründet (also Argumente verfasst) werden sollten und gegebenenfalls eine philosophische oder ethische Theorie aufgestellt wird (also ein System plausibel begründeter Aussagen, vgl. z. B. Beckermann 2003; Tetens 2006; Tugendhat und Wolf 1986). Weitere Voraussetzungen mögen hinzukommen. Für all diese Anforderungen bedarf es bestimmter Fähigkeiten, die das Vorherige wiederum voraussetzen: die Fähigkeit zur kritischen Reflexion von Theorien, zur Prüfung von Argumenten und Theorien im Hinblick auf Plausibilität ebenso wie logische Schlüssigkeit und Gültigkeit und die Aufstellung eigener Argumente und Theorie-Systeme.

[14] Dieses Kriterium ist schwer zu greifen, in der Philosophie ebenso wie in anderen Disziplinen. In den Naturwissenschaften sind Dissertationen von drei Seiten möglich, wie im Zuge der Plagiatsaffäre um die nachträglich abgelehnte Doktorarbeit der CDU-Politikerin Annette Schavan bekannt wurde (vgl. Mühlbauer 2013). Rudolf Carnap wurde 1921 mit einer neo-kantianischen Arbeit von 80 Seiten promoviert. Edmund Gettier schrieb 1961 eine Dissertation, die um die 200 Seiten umfasste. (Unter Wissenschaftstheoretikern bekannt wurde er jedoch mit seiner ersten Publikation, dem dreiseitigen Aufsatz *Is Justified True Belief Knowledge?*, der ihm auch zu einer unbefristeten Anstellung verhalf; vgl. Gettier 1963; vgl. hierzu auch Weber und Yolcu 2019: 30 ff.).

Wenngleich der Katalog an Voraussetzungen zur Verfassung einer Dissertation im Fach Ethik sicherlich noch ergänzt werden könnte, so bedarf es notwendigerweise bestimmter kognitiver Fähigkeiten.[15] Auf diese möchte ich mich im Folgenden fokussieren, da sie wiederum auch für die Frage, ob Maschinen ethisch sein können (und auch für andere Fähigkeiten im Rahmen starker künstlicher Intelligenz, s. u.) im Zentrum stehen. Die zentrale Frage für diesen Abschnitt könnte also auch lauten: Können Algorithmen denken? Diese Frage stellte der Mathematiker und Informatiker Alan Turing bereits 1950 in seinem Aufsatz *Computing Machinery and Intelligence* (Turing 1950). Allerdings ersetzte er die Frage durch ein Gedankenexperiment, aus Sorge, „the meaning of the words ‚machine' and ‚think' are to be found by examining how they are commonly used" (ebd.: 433). Um die Begriffe *machine* und *think* wissenschaftlich untersuchen zu können, erfand Turing ein *Imitation Game* (ebd.):

> The new form of the problem can be described in terms of a game which we call the ‚imitation game.' It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart front the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either ‚X is A and Y is B' or ‚X is B and Y is A.' The interrogator is allowed to put questions to A and B thus:
>
> C: Will X please tell me the length of his or her hair?
>
> Now suppose X is actually A, then A must answer. It is A's object in the game to try and cause C to make the wrong identification. His answer might therefore be:
>
> ‚My hair is shingled, and the longest strands are about nine inches long.'
>
> In order that tones of voice may not help the interrogator the answers should be written, or better still, typewritten. The ideal arrangement is to have a teleprinter communicating between the two rooms. Alternatively the question and answers can be repeated by an intermediary. The object of the game for the third player (B) is to help the interrogator. The best strategy for her is probably to give truthful answers. She can add such things as ‚I am the woman, don't listen to him¡ to her answers, but it will avail nothing as the man can make similar remarks.
>
> We now ask the question, ‚What will happen when a machine takes the part of A in this game?' Will the interrogator decide wrongly as often when the game is played

[15] Im Folgenden werde ich die Ausdrücke *kognitive Qualitäten*, *kognitive Eigenschaften* und *kognitive Fähigkeiten* verwenden. Das genaue Verständnis wird vor allem in Kapitel 5 deutlich werden. Die Verwendung der Ausdrücke beinhaltet ein weites Verständnis der Begriffe, nicht zu verwechseln mit bspw. *rational abilities* im Verständnis Gilbert Ryles (vgl. Ryle 1945). Für diesen Hinweis danke ich Michael Poznic.

> like this as he does when the game is played between a man and a woman? These questions replace our original, ‚Can machines think?'

An dieser Stelle ist Alan Turing also trotz des Titels seines Aufsatzes nicht daran interessiert, die Frage zu beantworten, ob Maschinen wirklich *denken* können in einem (zumindest ähnlichen) Sinn, wie man bei Menschen von *denken* spricht. Daher kann Turings Gedankenexperiment bei der Frage, ob Algorithmen und Computerprogramme denken und eine Dissertation im Fach Ethik verfassen können, nicht direkt zur Erhellung beitragen; ich werde es aber im Rahmen der Untersuchung menschlicher und künstlicher Intelligenz (S. 118 ff.) noch einmal aufgreifen.

Die intuitive Antwort auf die Frage, ob Maschinen, Computer und Algorithmen denken können, lautet vermutlich bei vielen Menschen *Nein*. Algorithmen folgen bestimmten mathematischen Gesetzen, die sie aber selbst nicht infrage stellen oder kritisch reflektieren können. Der Journalist und Autor Thomas Ramge (2019: 105 f.) kommt daher in Bezug auf Maschinen zu dem Schluss: „At the moment, there is no discernible development path that would make an intelligence explosion probable."[16] Viele Wissenschaftler∗innen stimmen ihm zu; nur ein Beispiel von vielen, das dies belegt, ist der Artikel von Braga und Logan (2017) mit dem süffisanten Titel „The Emperor of Strong AI Has No Cloths. Limits to Artificial Intelligence".

Auf der anderen Seite gibt es viele Stimmen, die große Bedenken vor der zunehmenden Selbstständigkeit technischer Systeme bis hin zu deren völliger Unabhängigkeit von Menschen und der Macht von Maschinen über Menschen äußern – und somit doch zumindest die Möglichkeit von Formen starker KI annehmen. Neben dem Stichwort *starke KI* ist dabei auch von *technologischer Singularität* die Rede.

> Das, was operationalisierbar ist, lässt sich grundsätzlich auch mit formalen Systemen darstellen und auf einem Computer berechnen. Vieles aber, was das menschliche Denken kennzeichnet und was wir mit intentionalen Termini wie Kreativität oder Bewusstsein benennen, entzieht sich weitgehend einer Operationalisierung. Dies wird jedoch angezweifelt von Vertretern der sog. ‚starken KI-These', die besagt, dass Bewusstseinsprozesse *nichts anderes* sind, die also Intelligenz und Kognition auf bloße Informationsverarbeitung reduziert. Ein solcher Nachweis konnte aber bisher nicht erbracht werden [...]. Hingegen wird kaum bestritten, dass Intelligenz *auch* Informationsverarbeitung ist – dies entspricht der ‚schwachen KI-These'.
>
> (Görz, Schmid und Braun 2021: 6 f. Herv. i. Orig.)

Für *starke KI* wurde allerdings nicht immer ein radikaler Materialismus vorausgesetzt; vielmehr bestand in einem früheren Stadium der Diskussion um maschinelle

[16] Die Rede ist hier zwar von *intelligence* und nicht von Denken; im Kontext wird aber klar, dass es um weitergefasste kognitive Fähigkeiten geht.

Intelligenz (definiert vor allem über bestimmte kognitive Fähigkeiten) eine Definition in der *Gleichsetzung* von Menschen und Maschinen (*starke KI*) im Gegensatz zu einer bloßen *Simulation* kognitiver Fähigkeiten und Eigenschaften (*schwache KI*) – und zwar zunächst unabhängig von der zugrundeliegenden philosophischen Theorie des Geistes (vgl. z. B. Russell und Norvig 2012: 26). Für die angrenzende Idee technologischer Singularität[17] ist neben einigen anderen der Autor Ray Kurzweil bekannt.

> What, then, is Singularity? It's a future period during which the pace of technological change will be so rapid, its impact so deep, that human life will be irreversibly transformed. Although neither utopian nor dystopian, this epoch will transform the concepts that we rely on to give meaning to our lives, from our business models to the cycle of human life, including death itself. [...] For example, computers are diagnosing electrocardiograms and medical images, flying and landing airplanes, controlling the tactical decisions of automated weapons, making credit and financial decisions, and being given responsibility for many other tasks that used to require human intelligence. The performance of these systems is increasingly based on integrating multiple types of artificial intelligence (AI). But as long as there is an AI shortcoming in any such area of endeavor, skeptics will point to that area as an inherent bastion of permanent human superiority over the capabilities of our own creations.
>
> This book will argue, however, that within several decades information-based technologies will encompass all human knowledge and proficiency, ultimately including the pattern-recognition powers, problem-solving skills, and emotional and moral intelligence of the human brain itself.
>
> (Kurzweil 2005: 8)

Da es sich hierbei um hypothetische Annahmen über die Zukunft handelt, bleiben wir zunächst bei der eingangs genannten, spontanen Reaktion, dass Algorithmen nicht denken können. Ein Argument dafür könnte wie folgt aussehen:

These A: Eine ethische Maschine muss, um sinnvollerweise als solche bezeichnet werden zu können, eine Dissertation im Bereich der Ethik verfassen können.[18]

These B: Um eine Dissertation im Bereich Ethik verfassen zu können, bedarf es kognitiver Fähigkeiten, wie unter anderem Reflexionsfähigkeit, Abstraktionsfähigkeit und die Fähigkeit zu logischem Denken.

[17] Krüger (2020: 270 ff.) grenzt technologische Singularität von kosmologischer ab; weitere Differenzierungen sind denkbar.

[18] Etwas anders formuliert, handelt es sich hierbei um eine materiale Implikation: *Wenn* eine Maschine eine Dissertation im Bereich Ethik verfassen kann, *dann* handelt es sich um eine ethische Maschine.

These C: Intelligente Maschinen basieren zwar auf logischen Systemen, können aber nicht selbstständig in jenem Sinn denken. Damit erfüllen Sie die Bedingungen für das Verfassen von Dissertationen im Bereich Ethik nicht.

Konklusion: Es kann keine ethischen Maschinen geben.

Bei dem formal schlüssigen Argument (das heißt, die Konklusion folgt logisch aus den Prämissen), handelt es sich schlicht um eine verneinte Implikation der Form:

Wenn eine Maschine eine Dissertation im Bereich Ethik verfassen kann (A), dann handelt es sich um eine ethische Maschine (B); also: $A \rightarrow B$.

Maschinen können *keine* Dissertation im Bereich Ethik verfassen (A), also gibt es *keine* ethischen Maschinen (B); also: $\neg A \rightarrow \neg B$.

Die Antwort auf die im Titel dieses Abschnitts gestellte Frage muss demnach lauten: Dieser Text kann nicht von einem Algorithmus geschrieben worden sein, da Algorithmen nicht denken können. Philosophisch gesehen gibt es an dieser Stelle mindestens drei Einwandmöglichkeiten, die die Gültigkeit des Arguments[19] infrage stellen:

1. Kritiker∗innen könnten verlangen, dass die Bedingungen für die Ausübung von Tätigkeiten innerhalb der ethischen Disziplin bei Menschen und damit des Verfassens einer Dissertation in den Bereichen Philosophie und Ethik überprüft, hinterfragt und gegebenenfalls angepasst werden müssen. Sie könnten also die Wahrheit von These B infrage stellen. Es wäre dann zum Beispiel denkbar, dass die Bedingungen anders formuliert werden und damit ethische Maschinen möglich gemacht werden könnten.[20]

2. Der zweite Einwand besteht in der Rückweisung von These A: Nur, weil eine Maschine keine Dissertation verfassen kann, muss das noch lange nicht heißen, dass es keine ethischen Maschinen geben kann, die andere Fähigkeiten haben, deren es bei der Ausübung der ethischen Disziplin bedarf und die bislang noch nicht genannt oder nicht genügend berücksichtigt wurden. Die Frage ist, ob es sich bei der Fähigkeit, eine Dissertation im Bereich Ethik verfassen zu können, um eine *notwendige* Bedingung für ethische Menschen oder Maschinen in unserem Sinn

[19] Hierfür müssen zusätzlich zur Schlüssigkeit auch die Prämissen als wahr angenommen werden.

[20] Wie ich weiter unten zeigen werde, wird die Unmöglichkeit bestimmter Fähigkeiten oder Eigenschaften bei Maschinen häufig in Abgrenzung zu einem sehr anspruchsvollen und voraussetzungsreichen Menschenbild zu begründen versucht, was auch (technik-)anthropologische Überlegungen mit sich führt und grundlegend infrage gestellt werden kann.

handelt – oder nicht. Kritiker∗innen könnten einwenden, dass ethisches Denken auch ohne die Fähigkeit möglich ist, eine Dissertation im Bereich Ethik zu verfassen.

3. Drittens kann man einwenden, dass es möglicherweise analoge Kompetenzen und Fähigkeiten bei bereits entwickelten (oder in naher Zukunft absehbaren) technischen Systemen gibt, die zwar gewöhnlich nicht als Reflexionsfähigkeit *bezeichnet* werden, die aber bei genauerer Betrachtung mindestens ähnliche Strukturen beinhalten. Dies könnten zu menschlichem Denken analoge, logische Prozessabläufe sein, was die Zuschreibung menschlicher Fähigkeiten oder Eigenschaften – wissend, dass es sich nicht um eine Form der Identität, sondern um eine Analogie handelt – rechtfertigen könnte. Dieser Einwand widerspricht These C.

Zum zweiten Einwand hatte ich bereits eingangs klargestellt: Bei der Definition ethischer Fähigkeiten über die Fähigkeit des Schreibens einer Dissertation handelt es sich um ein stilistisches Mittel, das zugunsten der Klarheit der Argumentation jederzeit aufgegeben werden kann. Der erste und der dritte Einwand müssen ernstgenommen werden und stehen bei der folgenden Analyse im Vordergrund. Beim dritten Einwand handelt es sich, genau genommen, um eine Spezifizierung des ersten, weshalb ich die beiden im Folgenden zusammenfassen werde.

Der Aspekt der Analogie menschlicher kognitiver Fähigkeiten und ähnlicher technischer Kapazitäten, der bereits bei der Erläuterung der Ausdrücke *starke* und *schwache KI* zur Sprache kam, wird entscheidend sein für die Beantwortung der Frage, ob es selbstständig denkende, ethische Maschinen geben kann. Eine der Hauptaufgaben der vorliegenden Arbeit wird darin bestehen, zu prüfen, ob Maschinen in demselben Sinn über die für Ethik kognitiven (und ggf. anderen) Fähigkeiten verfügen können wie Menschen – oder in einem anderen, analogen Sinn. Ein Blick auf verschiedene Science-Fiction-Narrative zeigt zudem, dass bei möglichen Zukunftsszenarien – egal, ob utopischer oder dystopischer Art – von einer Übersteigerung menschlicher Fähigkeiten durch Maschinen ausgegangen werden kann.[21] Es ist also durchaus auch eine dritte Form genuin maschineller Ethik denkbar, die weder identisch mit der menschlichen ist, noch analog zu dieser betrachtet werden kann.

Bei der Entwicklung technischer Systeme geht es, wie im folgenden Kapitel noch deutlich werden wird, primär meist darum, menschliche Tätigkeiten durch *bessere* und *effizientere* maschinelle zu ersetzen, was (aus heutiger Perspektive) anhand verhältnismäßig einfacher technischer Systeme, wie Waschmaschinen oder Einparkhilfen, verdeutlicht werden kann: Es handelt sich um Unterstützungen alltäglicher

[21] Einige Beispiele folgen im Verlauf der Arbeit, u. a. in Abschnitt 3.5.

menschlicher Tätigkeiten, die das Leben von Menschen vereinfachen sollen und dies in vielen Fällen auch leisten. Ohne Ergänzungen und Einschränkungen wäre diese Sicht natürlich naiv. Zum einen gibt es wirtschaftliche Motive, die nicht nur funktionelle Zwecke verfolgen, sondern auch finanzielle – mehr oder weniger unabhängig vom Nutzen der Technik für Menschen. Zum anderen kann auch die Hilfsfunktion technischer Systeme nicht pauschal als notwendige Bedingung technischer Systeme angenommen werden, angesichts des immensen Missbrauchs technischer Systeme zu Ungunsten vieler Menschen (man denke an aktuelle Beispiele wie *deep fakes*, vgl. van Huijstee et al. 2021). Bei manchen technischen Entwicklungen wird Wert auf eine äußerliche Ähnlichkeit zu Menschen gelegt (zum Beispiel bei der Entwicklung sogenannter humanoider Roboter, vgl. z. B. Decker 2010), bei anderen nicht. Insofern stellt sich grundlegend die Frage, in welche Kategorie technischer Systeme ethische Maschinen eingeordnet werden können: humanoid oder nicht? Menschlichen Zwecken dienend oder nicht? Möglicherweise eigene *Interessen* verfolgend – und wenn ja, wie und mit welchen Konsequenzen?

All diese Fragen sowie die Details der Analogie zwischen Menschen und Technik, ihre Bedingungen und die Antwort auf die Frage, ob Maschinen im Sinn einer Analogie zur Redeweise über Menschen *selbstständig denken* können, werden im Folgenden genauer ausgeführt werden. Bevor ich jedoch die Bedingungen genuin *ethischer* Maschinen diskutiere, ist es begrifflich sinnvoll, mit den theoretischen Bedingungen und möglichen praktischen Umsetzungen *moralischer* Maschinen zu beginnen. Zuvor folgt im kommenden Abschnitt die Definition zentraler Begriffe in der Arbeit.

2.3 Grundlegende Begriffsklärung

2.3.1 Maschinen, Roboter, Algorithmen und künstliche Intelligenz

Der Begriff der Technik wird in der vorliegenden Arbeit als Überbegriff über verschiedene Ausformungen technischer Systeme verstanden. Darunter fallen Roboter ebenso wie Maschinen und Computerprogramme (u. a.). Da die spezifischen Ausformungen von Technik bereits an anderen Stellen hinreichend differenziert wurden (vgl. z. B. Grunwald und Hillerbrand 2021: Abschnitt „B Spezifische Techniken“, S. 57 ff.), werde ich hier nicht genauer auf die technologischen Unterschiede eingehen, jedoch kurz den Hintergrund zu den hier verwendeten Terminologien erläutern.

In den Bereichen der Technikethik und -philosophie ist im Zusammenhang mit moralischen Handlungen nicht nur von *Maschinen* die Rede, sondern häufig auch von anderen technischen Systemen, wie Robotern, Algorithmen oder Computerprogrammen. Ich verwende in meiner Arbeit vorrangig die Ausdrücke *Maschine* oder allgemeiner *technische Systeme*. Ethisch relevant und für die Argumentation in der vorliegenden Arbeit interessant sind Maschinen und Computerprogramme

(seltener wird von *Software* die Rede sein) insofern, da diese in Kombination mit lernenden Systemen Outputs erzeugen können, die ethisch diskutiert werden können. Ob dabei im selben Sinn wie bei Menschen von *Handlungen* und / oder *Entscheidungen* die Rede sein kann, ist eine von vielen Detailfragen, die zu klären sein wird. Bei Umsetzungsversuchen von menschlicher Kognition ähnlichen Fähigkeiten wie *Bewusstsein* und *Denken* wird in der Regel auch von softwarebasierten Anwendungen ausgegangen. Die äußeren Darstellungsformen lernender Systeme ist dabei zunächst nebensächlich. Der Ausdruck *Roboter* wird aber häufig mit einer bestimmten, primär an menschlichen Eigenschaften orientierten, äußeren Form verbunden, weshalb ich ihn für meine Argumentation für potenziell irreführend halte. Wenngleich im Folgenden immer wieder ein direkter Vergleich zwischen Menschen und Maschinen aufs Tableau gebracht wird, kommt es dabei nicht primär auf eine ähnliche äußere Form an, sondern auf bestimmte kognitive Qualitäten, die als Voraussetzungen für bestimmte Denk- und Handlungsweisen angenommen werden können.

Ein häufig, insbesondere im Kontext von KI verwendeter, Ausdruck ist *Algorithmus*. Hierbei handelt es sich um ein metaphysisches Prinzip, das zunächst unabhängig von einer jeweiligen (technischen) Umsetzung mit gesehen werden sollte und auch als *Handlungsvorschrift* beschrieben werden kann. Die im Kontext selbstständig denkender und lernender technischer Systeme sehr bedeutende Rolle von Algorithmen wird an den jeweils relevanten Stellen in der vorliegenden Arbeit aufgegriffen. Eine weitere Differenzierung ist für meine Argumentation an vielen Stellen nicht notwendig und falls doch, nehme ich diese vor. Davon abweichende Ausdrücke aus Texten von anderen Autor∗innen setze ich, falls nötig, in Bezug zu meiner Terminologie.

Schließlich kurz zu klären sind die Ausdrücke *maschinelles Lernen* (oder *machine learning*), *big data* und *deep learning*. Alle Methoden und Verfahren wurden in anderen Werken bereits hinreichend beschrieben, weshalb ich die Begriffe hier nur grob skizzieren will. Maschinelles Lernen ist eine Methode der Informatik, bei der ein System

> keine Regeln vorgegeben [bekommt; L. S.], wie zur Lösung gelangt wird, sondern Lernregeln, die, angewendet auf Trainingsdaten, es dem System ermöglichen [sic!] Regeln für die Problemlösung zu entwickeln. Im Fall der Sprachsoftware dient eine große Anzahl an Beispielsätzen als Input. Solche Systeme sind oft sehr leistungsfähig, versagen aber häufig, sobald es darum geht [sic!] zu generalisieren, d. h. wenn sie in der Anwendung mit Fällen konfrontiert werden, die sich deutlich von den Fällen [sic!] anhand derer gelernt wurde, unterscheiden.
>
> (Heil 2021: 424)

Die Varianten des *überwachten* und *unüberwachten Lernens* werde ich weiter unten im Abschnitt 5.4.3 im Kontext der konkreten technischen Umsetzung ethischer

Theorien vorstellen. Je mehr Daten einem Systeme für Lernprozesse zur Verfügung stehen, desto genauer kann es seine Prognosen machen, weshalb das Narrativ vieler Daten (*big data*) als Vorteil kursiert – ebenso können die Daten aber auch missbraucht werden, wie aktuelle Beispiele von *deep fakes* zeigen (s. o.). *Deep learning* kann, die beiden vorherigen Ausdrücke zusammenfassend, als Methode des maschinellen Lernens beschrieben werden, die auf großen Datenmengen basiert. Die Ausdrücke *Intelligenz* und *künstliche Intellilgenz* werden in Abschnitt 5.2.6 erörtert.

2.3.2 Technische Systeme und Moral – erste Annäherung

Die Grundfrage der Ethik ist gemäß Immanuel Kant (*Kritik der reinen Vernunft*) „Was soll ich tun?“ (KrV, AA: 805). In der zeitgenössischen Philosophie wird die Frage als moralische bezeichnet und die Ethik als das Feld, die Metaebene oder Disziplin, von denen aus moralische Fragen reflektiert und zu beantworten versucht werden. Gerhard Ernst unterscheidet moralische Fragen *im engeren Sinn*, die nicht nur von Bedeutung für das Wohl der eigenen Person, sondern mindestens einer anderen Person oder eines anderen Lebewesens sind, von moralischen Fragen *im weiteren Sinn*, worunter sämtliche Fragen fallen, also beispielsweise auch „Soll ich lieber den Vanillepudding oder doch besser das Erdbeertörtchen als Nachtisch nehmen?“ (Ernst 2009: 11). Moralische Fragen im engeren Sinn gelten auch gemeinhin als moralisch relevant.

> Aber eine Frage wie ‚Soll ich lieber den Vanillepudding oder doch besser das Erdbeertörtchen als Nachtisch nehmen?‘ würden wir normalerweise nicht als eine moralische Frage bezeichnen, obwohl es sich nach meiner Charakterisierung der Moral um eine solche handelt: Schließlich geht es auch hier darum, was zu tun richtig ist – auch wenn vermutlich nicht viel davon abhängt, ob ich die richtige oder falsche Wahl treffe, und es gut sein kann, dass sogar beide Handlungsoptionen gleich gut sind (und damit jede Wahl richtig ist). Entsprechend würden wir normalerweise bei diesem Beispiel die falsche Entscheidung (wenn es eine gibt) nicht ‚moralisch falsch‘ nennen. Wenn Hans lieber Vanillepudding als Erdbeertörtchen isst und andere Aspekte, wie etwa der Preis, gesundheitliche Rücksichten etc. [sic!] keine Rolle spielen, dann wäre es sicher falsch von Hans, das Erdbeertörtchen zu wählen. Aber wir würden nicht sagen, dass er sich einer moralischen Verfehlung schuldig macht, wenn er nicht den Pudding nimmt, sondern, dass er unklug handelt. Von einer moralischen Verfehlung sprechen wir nur dann, wenn eine Handlung in ganz bestimmten Hinsichten schlecht ist. Wenn sich beispielsweise Hans das Erdbeertörtchen aneignet, ohne es zu bezahlen, dann verletzt er durch seine Handlung die Besitzrechte einer anderen Person. Und das würden viele als moralische Verfehlung ansehen. Es ginge dann um das, was ich die *Moral im engeren Sinn* nennen möchte.
>
> (Ebd.)

Robin Celikates und Stefan Gosepath definieren:

> Als Kerndisziplin der Praktischen Philosophie beschäftigt sich die Moralphilosophie (oder wie sie manchmal auch genannt wird: die Ethik) mit der Frage, was wir tun sollen. Im Unterschied zur Politischen Philosophie, in der es um die richtige politische und soziale Ordnung geht, wird hier mit Bezug auf das Individuum gefragt: Wie soll *ich* mich verhalten? Welches ist die richtige Handlung?
> (Celikates und Gosepath 2017a: 7)

Im Unterschied zu Gerhard Ernst sprechen Robin Celikates und Stefan Gosepath nicht von Moral im *engeren* und im *weiteren* Sinn, sondern sie unterscheiden mit Bezug auf Jürgen Habermas (1991: 100–118) *drei Dimensionen* der Frage *Was soll ich tun?*:

> (1) Die Frage danach, wie ich möglichst effizient ein bestimmtes Ziel erreiche, ist eine technische beziehungsweise pragmatische Frage; (2) die Frage, was gut für mich ist beziehungsweise was ich eigentlich will, ist eine ethisch-existenzielle Frage; (3) erst die Frage, was ich tun oder lassen soll, unabhängig davon, was meine jetzigen Ziele sind und was ich will, ist die genuin moralische Form der praktischen Frage ‚Was soll ich tun?' Nur auf sie ist eine genuin moralische Antwort erforderlich.
> (Celikates und Gosepath 2017a: 7)

Auf die Frage, ob auch *ethisch-existenzielle* Fragen, wie Celikates und Gosepath sie nennen, zu den genuin moralischen zählen oder nicht, – Gerhard Ernst bringt als Beispiel die Frage, was man zum Nachtisch essen soll, für Celikates und Gosepath fällt auch die Frage, wie man sich einem Bahnübergang verhalten soll, in diese Kategorie (vgl. Ernst 2009: 11; Celikates und Gosepath 2017a: 7) – kann für die vorliegende Arbeit geantwortet werden: Es scheint sinnvoll zu sein, Ernsts Unterscheidung von Moral im *engeren* und im *weiteren* Sinn zu übernehmen, um nicht bestimmte Situationen von vornherein aus dem Kreis der moralischen Relevanz auszuschließen. Gleichwohl fallen die meisten maschinenethischen Fragen in den Bereich engerer – beziehungsweise geuiner – Moral, insofern die Verhaltensweisen technischer Systeme meist Auswirkungen auf mehrere Personen in größeren Kontexten haben und moralische Fragen diesbezüglich grundsätzlicher und allumfassender Lösungen bedürfen.

Bei der Frage, ob Maschinen moralisch handeln können, wenn ja, unter welchen Umständen und was dies auf Mikro-, Meso- und Makroebene bedeutet, handelt es sich in aller Regel um moralische Fragen im engeren Sinn, für die die Rolle einzelner Entscheidungsträger bedeutsam ist, wie die Beschreibung Jessica Heesens zeigt:

> Die KI-Entwicklung soll der Gesellschaft dienen und nicht dazu führen, dass neue technische oder wirtschaftliche Zwänge entstehen, die ethische Normen des Zusammenlebens verletzen oder positive Entwicklungen beschränken. Das ist ja eigentlich damit gemeint, wenn in vielen politischen Dokumenten und Reden von ‚KI für den Menschen' die Rede ist. Generell ist es wichtig, die Handlungsroutinen und Handlungsempfehlungen, die durch KI-Systeme vorgegeben werden, nicht als alternativlos bzw. als ‚Sachzwang' anzuerkennen. Wie in andere technische Produkte auch, sind in KI bestimmte, veränderbare Zwecke und Präferenzen eingeschrieben, die bestimmten Gruppen und Individuen nutzen, anderen aber schaden können.
> (Heesen 2021b: 49)

Als Beispiele werden Fragen rund um die Moralfähigkeit von Maschinen und technischen Systemen häufig in Bezug auf den Einsatz von Pflegerobotern oder anderen Technologien, die im Pflegesektor als Unterstützung eingesetzt werden, diskutiert, in Bezug auf automatisierte Fahrzeuge, militärisch eingesetzte Drohnen oder Systeme, die im Medizinsektor die Erstellung von Diagnosen unterstützen sollen (vgl. u. a. Misselhorn 2019: 136). Aber auch der Einsatz von Sexpuppen wird diskutiert (vgl. Bendel 2018a), die Verankerung von Grundwerten, wie *Selbstbestimmung, Gerechtigkeit* und *Privatheit* (vgl. Heesen 2021b: 49 f.), beziehungsweise Grundrechten in technischen Systemen (vgl. Orwat und Bless 2016) ebenso wie mögliche Folgen, die sich aus einem moralischen Status für die Maschinen selbst ergeben, zum Beispiel, ob diese Rechte haben sollten (vgl. u. a. Gunkel 2018; Mamak 2021). Ferner stellt sich die Frage, ob Roboter und Algorithmen selbst moralische Ansprüche stellen können, zum Beispiel auf Unversehrtheit, eine Form von *Würde* (s. z. B. das Zitat von McEwan auf S. 33) oder Rechte.

> Doch selbst wenn sich nach vorheriger Prüfung herausstellen sollte, dass Roboter keine moralischen Handlungssubjekte sein können und damit zu moralischem Handeln selbst nicht befähigt wären, sollte man ihnen einen Platz im moralischen Universum zuweisen. Schließlich sind alle möglichen Wesen und Entitäten – wie etwa Tiere, Pflanzen, Häuser, Autos, Smartphones, Landschaften oder ganze Ökosysteme – Objekte moralischen Handelns, und wir sprechen einer ganzen Reihe von nichtmenschlichen und zum Teil auch unbelebten Entitäten einen Wert, ja, in manchen Fällen sogar Rechte zu.[22]
> (Loh 2019a: 13)

Moralische Maschinen werden zuweilen als Problem oder Bedrohung wahrgenommen – beziehungsweise wird diskutiert, welche (positiven und negativen) Folgen

[22] Ein aktuelles Beispiel ist eine Gruppe von Jurist∗innen, die derzeit versuchen, mittels eines Volksbegehrens Rechte für die Natur (in der Rechtssprache die „natürliche Mitwelt") in der Verfassung zu verankern. https://www.riffreporter.de/de/umwelt/volksbegehren-klimaklage-recht-der-natur-bayern-verfassung, aufgerufen am 01.03.2022. Vorreiter für solche Vorhaben gibt es in Ecuador, Bolivien und den USA. https://www.klimareporter.de/protest/welche-rechte-hat-ein-fluss-wald-oder-berg, aufgerufen am 01.03.2022.

die (potenzielle) Moralität von Maschinen hat. Diskutiert wird dabei auch, ob und inwiefern technische Systeme moralische Akteure[23] sein können (vgl. u. a. Rath 2019) beziehungsweise inwiefern sie Handlungen ausüben und Verantwortung übernehmen können (vgl. u. a. Loh 2019a). Die Frage, ob technische Systeme selbst als moralische Akteure gewertet werden können oder nicht, ist bedeutend für die moralische Bewertung ihrer Folgen. Inwiefern sind Algorithmen aus sich heraus moralisch relevant? Oder ist dies allein ihre Herstellung und Benutzung durch Menschen? Diese Fragen betreffen den moralischen Status insbesondere komplexerer technischer Systeme, auf die ich kurz eingehen möchte.

2.3.3 Zum moralischen Status technischer Systeme.

Welchen moralischen Status haben technische Systeme: einen prinzipiell passiven oder zumindest zu Teilen aktiven? Sind Algorithmen & Co. also lediglich Werkzeuge, die von Menschen (oder Tieren) benutzt werden, womit ihnen ein rein passiver Objektstatus zukommt? Oder können sie auch aktiv sein und selbst *handeln*, *Entscheidungen treffen*, *Ziele verfolgen*[24] und somit moralische Subjekte sein? Ersteres wird auch als Neutralitätsthese bezeichnet (vgl. u. a. Radder 2009; Ullrich 2014). Stefan Ullrich (2019b: 250) schreibt dazu: „Bislang war in Bezug auf Ethik und Moral immer die Rede vom Technik gebrauchenden Menschen und nicht von der Technik selbst“. Inzwischen wird jedoch, mit unterschiedlichen Begründungen und in unterschiedlichen Ausführungen (auch abhängig von der jeweiligen Komplexität der jeweiligen Anwendung), häufig die Ansicht vertreten, dass technische Systeme keineswegs nur neutrales Werkzeug sind, sondern bereits bestimmte Werte genuin beinhaltet.[25] Ein Beispiel:

[23] Den Begriff *Akteur* verstehe ich im Folgenden in einem weiten Sinn, der sowohl menschliche Handlungen als auch maschinellen Output impliziert. Voraussetzung für Akteurschaft sind gemäß dem in Abschnitt 4.2 dargelegten technischen Pragmatismus in erster Linie die Folgen von Handlungen oder Aktionen und dafür notwendige Bedingungen.

[24] Diese Zuschreibungen sind gemäß der eingangs definierten Analogie von maschinellen zu menschlichen Fähigkeiten zu verstehen (s. S. 19 f.), auf die ich im Folgenden noch näher eingehen werde.

[25] Dasselbe gilt im Übrigen auch für die Beobachtung der Folgen von Technikentwicklungen. Nierling und Torgersen (vgl. 2019: 12 ff.) beschreiben *drei Ebenen der Normativität* für die Technikfolgenabschätzung (TA): „TA als Instrument der Politik- bzw. Gesellschaftsberatung“, der „Umgang mit offenen oder verdeckten normativen Setzungen im TA-Forschungsprozess“, die Diskussion um einen „normativen Kern“ (ebd.: 12 ff.) von TA im Sinne eines „shared ethos of TA“ (Hennen und Nierling 2019: 20).

> At least when used for specific purposes, ICT systems [Information and Communication Technology; L. S.] intentionally or unintentionally *embed certain values* and have indirect effects on other values [...]. In most cases, the values directly addressed in engineering are technical and economic values like system security or economic efficiency.
>
> (Orwat und Bless 2016: 26. Herv. i. Orig.)[26]

Die Werte nicht-neutraler Maschinen und Computerprogramme können dabei entweder in den Systemen selbst verankert sein, wie im eben zitierten Beispiel. Die Rede ist dann auch von *values in design* (vgl. hierzu Knobel und Bowker 2011 in Orwat und Bless 2016). Oder moralische Werte spiegeln sich in der *menschlichen* Umgangsweise *mit* technischen Systemen wider – die zum Beispiel auch der ursprünglichen Konzeptions-Motivation widersprechen kann: „Ein Buttermesser und ein Assassinen-Dolch sind mit höchst unterschiedlichen Vorüberlegungen gestaltet worden“ (Ullrich 2019b: 256). Das Beispiel legt den Gedanken nahe, dass auch Geräte für unmoralische Handlungen verwendet werden können, die ursprünglich nicht dafür vorgesehen waren. Das zeigt, dass die Bewertung von Technik nicht zwangsläufig an unveränderlichen, Technik inhärenten Werten, geschehen muss, sondern diese auch durch die Benutzung von Menschen veränderbar sind. Auch Janina Loh sieht menschliche Handlungen mit Technik als zentral für deren Bewertung:

> Gegen eine häufig anzutreffende Intuition, dass Technik neutral sei, lässt sich einwenden, dass Technik allgemein Produkt menschlichen Handelns und damit immer (ob bewusst oder unbewusst) durch Normen und Werte bestimmt wird. Denn gerade durch ihre Intention unterscheidet sich eine Handlung vom Instinkt oder bloßen Verhalten [...]. Durch die Intentionen werden Werte in eine Handlung eingeschrieben. Menschen wählen über Gründe zwischen unterschiedlichen Handlungsalternativen, die zuvor implizit oder explizit gegen andere Gründe abgewogen wurden. Mit Robotern als spezifischen Technologien, den in ihnen implementierten Werten und den sich aus ihnen ergebenden (gesellschaftlichen) Konsequenzen gehen also immer moralische Fragen einher [...].
>
> (Loh 2019a: 9)

Ob die Werte letztendlich in technischen Systemen genuin enthalten sind, wie Loh beschreibt oder ob die Werte in den menschlichen Entscheidungen, Handlungen usw. zu verorten sind (und Technik in diesem Fall als reines Mittel zu bewerten ist), wird im Folgenden insbesondere bei komplexen technischen Systemen wie ethischen Maschinen eine wichtige Rolle spielen. Gemäß der Neutralitätsthese sind technische Dinge in erster Linie Instrumente, die *von Menschen* zum Vor-

[26] Vgl. hierzu auch van de Poel 2021a.

oder Nachteil anderer Menschen, von Tieren oder der weiteren Umwelt eingesetzt werden können (vgl. auch Pitt 1999). Franssen, Lokhorst und van de Poel (2018) unterstützen, wie Loh, die Argumentation gegen die Neutralitätsthese:

> This view [die Neutralität von Technik; L. S.] might have some plausibility in as far as technology is considered to be just a bare physical structure. Most philosophers of technology, however, agree that technological development is a goal-oriented process and that technological artifacts by definition have certain functions, so that they can be used for certain goals but not, or far more difficulty or less effectively, for other goals. This conceptual connection between technological artifacts, functions and goals makes it hard to maintain that technology is value-neutral.

Inwiefern Maschinen auch als Akteure gelten können, welche Konsequenzen das für Zuschreibungen von Handlungen und Verantwortung hat, werde ich in Kapitel 4 ausführlich darlegen und diskutieren. Es scheint jedoch so zu sein, dass die potenzielle Moralität von Maschinen keineswegs so problematisch ist, wie es oft in wissenschaftlichen Diskursen dargestellt wird. Geht man davon aus, dass Algorithmen nicht nur „Kunstwerke der Faulheit“ sind (Stiller 2015: 9), die eine „mystische Aura“ umgibt (Ullrich 2019a: 2), sondern vielmehr als „kodifizierte Handlungsvorschriften zur Lösung eines kodifizierbaren Problems“, so kann eine Übereinstimmung mit der oben festgelegten Definition von Moral gefunden werden. Wird moralisches Handeln (im weiten oder engen Sinn), wie oben beschrieben, über eine Antwort auf die Frage *Was soll ich tun?* definiert, so ist ein Algorithmus eine „Problemlösungsmaschine“ (ebd.: 4), indem sie einen spezifischen Lösungsweg für ein Problem oder eine Aufgabe darstellt. Gemäß diesen Bestimmungen ist ein Algorithmus die technische Verkörperung moralischer Handlung par excellence und somit per definitionem moralisches Handlungssubjekt. Diese grundlegende Definition von Moral beziehungsweise *moralischen Maschinen* ersetzt dabei keineswegs eine ausführliche Analyse und Diskussion der Bedingungen maschineller Moral, potenzieller Akteurs- und Handlungsfähigkeit, wie sie später folgen wird.

2.3.4 Maschinenethik im Feld der Bereichsethiken

Maschinenethik, Computerethik, Roboterethik, digitale Ethik und Ethik künstlicher Intelligenz sind sogenannte Bereichsethiken, die der angewandten Ethik zugeordnet werden.[27] Diesen Bereichsethiken übergeordnet ist, je nach wissenschaftlicher Perspektive, die Technikethik (vgl. Grunwald 2020) beziehungsweise

[27] Vgl. u. a. Bendel 2018a; Grimm, Keber und Zöllner 2019; Johnson und Snapper 1985; Loh 2019a; Misselhorn 2019; Nida-Rümelin 2005; Rath, Krotz und Karmasin 2019. Für weitere angrenzende philosophische Bereiche vgl. auch Abschnitt IV C in Grunwald und Hillerbrand 2021.

die Informationsethik (vgl. Heesen 2016). Ihr Gegenstand sind moralische Fragen im Zusammenhang mit technischen Systemen und nach Janina Loh (2019a: 13) „[. . .] Sonderbereiche des menschlichen Lebens, in denen Werte vertreten, Normen geltend gemacht und Regeln formuliert werden [. . .]“. Inwiefern die Bereiche der Ethik als *Sonder*bereiche zu verstehen sind oder schlicht als völlig *normale* Bereiche menschlichen Lebens, kann freilich diskutiert werden. Rath, Karmasin und Krotz (2019: 8) verstehen Maschinenethik wesentlich umfassender:

> Maschinen*ethik* im eigentlichen Sinne, also nicht nur eine maschinisierte Moralanwendung, ist dann als eine Spezifizierung der Medienethik zu verstehen, sofern digitale Maschinen, denen Künstliche Intelligenz zuzuschreiben und ggf. sogar moralische Intelligenz zu unterstellen wäre, nicht nur Objekte, sondern Akteure medial vermittelter Normansprüche sind.
>
> (Herv. i. Orig.)

In diesem Verständnis fallen jedoch alle möglichen Akteure medial vermittelter Normansprüche in den Bereich der Medienethik und es stellt sich – auch vor dem Hintergrund des sehr grundlegenden Medienbegriffs von Matthias Rath (2014) – die Frage, welche Gegenstände dann nicht (zumindest potenziell) der Medienethik zugeordnet werden können. Wie ich im Lauf der Arbeit zeigen werde, müssen sich die Zuschreibungen „Akteure medial vermittelter Normansprüche“ und „maschinisierte Moralanwendungen“ außerdem nicht ausschließen.

Ebenso wie sich die Roboterethik von der Roboterphilosophie unterscheidet (Loh 2019a: 11), gibt es auch die Technikphilosophie (vgl. z. B. Irrgang 2011), wobei Maschinenphilosophie als Begriff eher unüblich ist; hier ist im Deutschen auch die Rede von der *Philosophie der Maschine* (vgl. Burckhardt 2018); im Englischen gibt es entsprechend den Ausdruck *philosophy of artificial intelligence* (vgl. z. B. Carter 2007; Nath 2009).[28] Die Untersuchungsgegenstände und Fragen der Bereichsethik und Teilbereiche der Philosophie können sich dabei natürlich auch überlappen:

> Roboterethik ist eine Disziplin der Roboterphilosophie, die zusätzlich beispielsweise epistemologische, ästhetische, politikphilosophische und rechtsphilosophische Themen behandelt, wobei es innerhalb der Roboterphilosophie natürlich zu einer Überschneidung zahlreicher Fragen – wie etwa von ethischen und politikphilosophischen Fragen – kommt [. . .].
>
> (Loh 2019a: 11 f.)

[28] Im englischsprachigen Raum scheinen die wissenschaftlichen Diskussionen in diesem Bereich aber häufiger dem Feld der *philosophy of science* zugeordnet zu werden, vgl. z. B. Korb 2004; Williamson 2009.

Janina Loh (2019a) merkt zudem an, dass eine „roboterspezifische Entsprechung der philosophischen Anthropologie, also der philosophischen Disziplin, die nach dem Wesen ‚des' Menschen fragt" nicht existiere. Zwar existiert der Ausdruck *Machinologie* nicht als Name für eine Disziplin, die die Untersuchung des Wesens technischer Systeme zum Ziel hat; allerdings wird in den vorgestellten Teilbereichen nichts anderes gemacht, als ebendas aus verschiedenen Perspektiven zu untersuchen.

2.3.5 Technische Systeme und Ethik – erste Annäherung

Im Englischen werden die Ausdrücke *moral* und *ethical*[29] – ebenso wie die Ausdrücke *morals* und *ethics* – häufig gleichbedeutend verwendet, wie zum Beispiel in dem folgenden Zitat von James Gips deutlich wird:

> When our mobile robots are free-ranging how ought they to behave? [...] We want our robots to behave more like equals, more like ethical people. [...] How do we program a robot to behave ethically? Well, what does it mean for a person to behave ethically? People have discussed how we ought to behave for centuries. Indeed, it has been said that we really have only one question that we answer over and over: What do I do now? Given the current situation what action should I take?
> (Gips 1995: 1)

To behave ethically – das ist die Formulierung, von der ich eingangs geschrieben habe, dass sie im Deutschen unüblich ist; im Englischen ist sie jedoch möglich, da Moral und Ethik nicht nur umgangssprachlich, sondern auch wissenschaftlich häufig äquivalent verstanden werden. Im Deutschen würde man eher von *moralischem* Verhalten sprechen – jedoch gibt es auch hier Ausnahmen, wie die Definition Dietmar Hübners zeigt, der Moral als ein „*Normensystem*" beschreibt, „dessen Gegenstand *menschliches Verhalten* ist" (Hübner 2021: 13. Herv. i. Orig.).

Neben der Gleichstellung von Ethik und Moral beziehungsweise *ethical machines* und *moral machines* gibt es noch ein weiteres Verständnis, in dem der Ausdruck *Ethik*, insbesondere im Englischen, verwendet wird. Gemeint sind dann ethische Reflexionen über moralische Maschinen, wenn zum Beispiel von *ethical artificial intelligence* die Rede ist oder allgemeiner „the ethical issues of AI" (Pavaloiu und Kose 2017: 15) adressiert werden. Mit Annemarie Pieper kann man Ethik verstehen

[29] Wie bereits in Abschnitt 2.3.1 angemerkt, gelten Aussagen über *ethische Maschinen* oder *ethical machines* im Folgenden auch für angrenzende Ausdrücke, wie *ethical robot* oder *ethical neural network* – auch wenn ich die letzten beiden Termini teilweise als irreführend einschätze.

> als eine Disziplin der Philosophie [...,] als *Wissenschaft vom moralischen Handeln.* Sie untersucht die menschliche Praxis im Hinblick auf die Bedingungen ihrer Moralität und versucht, den Begriff der Moralität als sinnvoll auszuweisen.
> (Pieper 2017: 15. Herv. i. Orig.)

Eine wichtige metaethische Frage, die sich in Bezug auf Maschinen stellt, ist die der Subjekte und Objekte innerhalb einer Maschinenethik.[30] *Wer* trifft also die Entscheidungen *für wen*? Wer oder was als handelndes Subjekt (*agens*) verstanden werden kann (und mit welcher Begründung), hat Folgen für die Entscheidungs- und somit auch Verantwortungsfähigkeit von Menschen – und ggf. auch Maschinen. Dies wurde oben bereits anhand der Darstellung der *Neutralitätsthese* und ihrer Gegenpositionen deutlich (s. S. 25).

Für den Begriff der Maschinenethik ergeben sich, gemessen am Grad der zugeschriebenen Selbstständigkeit, mehrere mögliche Definitionen:

1. Eine genuine „Menschenethik“ (Bendel 2017: 5), also eine Ethik *von* Menschen *für* Menschen im Umgang mit technischen Systemen.

2. Eine Maschinenethik *von* Menschen für Maschinen. Dies klingt zunächst absurd, da man – einem allgemeinen, intuitiven Verständnis zufolge – bei Maschinen viel weniger als bei Tieren davon ausgehen kann, dass sie Eigenschaften aufweisen, die einen moralischen Schutz rechtfertigen würden, wie zum Beispiel Leidensfähigkeit. Diskutiert wird diese Variante meist mithilfe eines juristischen Verständnisses von Werten, die unter Berufung auf Rechte eingefordert werden können. Dies wird oft dann diskutiert, wenn Maschinen ein moralischer Subjektstatus zugeschrieben wird (vgl. z. B. Gunkel 2018).[31]

3. Neu ist die Idee, die in der vorliegenden Arbeit entwickelt werden soll: Eine genuine Ethik von Maschinen. In einem eher dystopischen Szenario, das auch in fiktiven Narrativen immer wieder zu finden ist, ist denkbar, dass Maschinen sich eine eigene *Maschinen-Ethik*[32] entwickeln, bei der Maschinen selbst

[30] Eine Parallele kann hier auch zu Überlegungen des moralischen Status von Tieren gezogen werden, vgl. Darling 2021.

[31] Bei dieser Argumentation kann es passieren, dass Menschen durch die Hintertür als Objekte der Moral hineingebracht werden, indem beispielsweise für moralische Pflichten gegenüber Maschinen plädiert wird mit einer an Immanuel Kants Argumentation angelehnten Begründung, dies würde wiederum der charakterlich-moralischen Entwicklung von Menschen zugutekommen. In der Tierethik wird von manchen Autor∗innen kritisiert, dass innerhalb dieses Arguments die Tiere nicht als eigentliche Objekte der Moral verstanden werden, sondern nur als Mittel zum Zweck für menschliche Moralbildung; dies wird dort auch als *Verrohungsargument* bezeichnet (vgl. Baranzke 2018).

[32] Die Schreibweise *Maschinen-Ethik* wird hier lediglich zur Betonung der skizzierten, unterschiedlichen Verhältnisse von *Maschinen* und *Ethik* verwendet.

und eventuell auch Menschen allenfalls moralische Objekte sind. Im Folgenden möchte ich jedoch zunächst von einem Verständnis dieser *Maschinen-Ethik* ausgehen, in der sich die genuin maschinell generierten Werte an einer prinzipiell menschlichen Ethik orientieren.

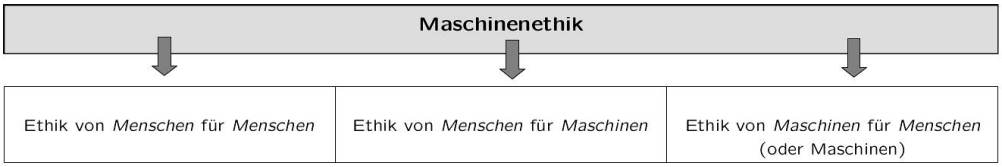

Abbildung 1: Drei mögliche Lesarten des Begriffs *Maschinenethik*.

Mit Rath, Karmasin und Krotz (2019: 6) kann „Maschinenethik im nicht trivialen Sinne einer Ethik der Maschinen – verstanden als *genitivus objectivus*“ (Herv. i. Orig.) gesehen werden, die gemäß der Autoren jedoch auch dann „immer Ethik des Menschen im Blick auf die Maschine“ ist und somit in die erste der aufgezeigten Varianten fällt.

Die zunehmende Automatisierung und an vielen Stellen befürwortete Selbstständigkeit technischer Systeme wirft dabei ein grundlegendes Problem auf: Entweder müssen Maschinen so eigenständig handeln und entscheiden können, dass sie sich auch ohne viel menschliches Zutun in einer komplexen, analogen oder digitalen, Welt gemäß den in einem Land oder einer Gesellschaft jeweils vorrangig akzeptierten moralischen Regeln bewegen können – oder Menschen müssen in kritischen Situationen eingreifen, was die (potenziell mögliche) Autonomie der Algorithmen untergräbt und sie dadurch gegebenenfalls wissenschaftlich und wirtschaftlich weniger attraktiv macht. Für den ersten Fall, der dem meiner Arbeit zugrundeliegenden, hypothetischen Modell genuin ethischer Maschinen entspricht, müssen, wie bereits angesprochen, die Fragen geklärt werden, welche Voraussetzung für selbstständiges Denken und Handeln als notwendig angenommen werden und wie diese ggf. in technischen Systemen realisiert werden können.

Bezüglich der Definition von Maschinenethik und der oben angeführten dritten Variante sehen Rath, Karmasin und Krotz (ebd.: 3 f.) zwei „extreme Positionen“, nämlich eine,

> die annimmt, dass Maschinen stets die Folge menschlicher Handlungen sind, dass es also im Kern nur um technische Hilfsmittel menschlichen Wollens ginge, und dass deswegen weder in begründungstheoretischer noch in praktischer Hinsicht neue ethische Konzeptionen von Nöten seien, die andere Position, die annimmt, dass selbstlernende Systeme und künstliche Intelligenz auch einen autonomen Willen von Maschinen begründen, der, wenn schon nicht dem menschlichen gleichzusetzen, diesem zumindest vergleichbar wäre – es wäre also sehr wohl von Nöten, den

> anthropozentrischen Fokus moderner Ethik zu Gunsten einer breiteren Konzeption von Ethik aufzugeben.

Die erste Position ist mit einer oben angeführten *Menschenethik* zu vergleichen. In der zweiten Position stecken zwei bereits angesprochene Fragen unterschiedlicher Ebenen, die zu klären für die vorliegende Arbeit zentral ist. Zum einen geht es um die Frage, welche Eigenschaften und Fähigkeiten immer komplexer werdenden technischen Systemen, die auch ethisch reflektieren, entscheiden und handeln können sollen, zugeschrieben werden können und müssen. Hier spielt auch die oben bereits formulierte Frage nach der Neutralität beziehungsweise intrinsischen Werthaltigkeit von Maschinen eine Rolle. Zum anderen geht es um die Frage des Gegenstands von Ethik oder des Ziels ethischer Bestrebungen. Unter Punkt 3 habe ich bereits die Möglichkeit erwähnt, dass als Objekt genuiner Maschinenethik auch Maschinen selbst verstanden werden können. Inwiefern diese Sichtweise sinnvoll und unter welchen Bedingungen sie realistisch ist, werde ich weiter unten prüfen. Die ethische Bewertung moralischer Maschinen findet bislang freilich auf der Basis genuin menschlicher Kategorien des *Guten* und *Nicht-Guten* statt, aber auch dies könnte sich im Kontext der Entstehung einer genuinen Maschinenethik gegebenenfalls ändern. Rath, Karmasin und Krotz (2019: 9) zufolge bedarf es zwar keiner grundlegend *breiteren Konzeption von Ethik*,

> sehr wohl aber stellen sich in der empirischen Konkretion neue ethische Fragen. Die Fragen nach Verantwortung und Gerechtigkeit, nach Sicherheit und Freiheit, nach der *conditio humana* und nach dem normativen Horizont des Sollens gegenüber einem scheinbar endlosen Meer des digitalen Seins sind neu und anders zu beantworten, auch wenn es im Kern keine neuen Fragen sind.
> (Herv. i. Orig.)

Als Gegenstand der Maschinenethik wird in der vorliegenden Arbeit gemäß der zuvor entwickelten Definition also nicht „die Entwicklung moralischer Maschinen“ (Misselhorn 2018: 2) oder „die Moral in der Maschine (bzw. die maschinelle Moral oder die moralische Maschine)“ (Bendel 2018b: 2) verstanden, sondern eine genuin technisch produzierte Ethik technischer Systeme, die sich in der folgenden Konzeption (zunächst) an menschlichen Wertvorstellungen orientiert.

Angesichts der technischen Entwicklungen im Laufe der Zeit, die ich im folgenden Kapitel knapp darlegen werde, die unabhängig von dem tatsächlichen Einfluss auf menschliche Leben bezüglich Freiheit, Selbstständigkeit und Zeit, dazu führten, dass viele menschliche Tätigkeiten durch schnellere, effizientere maschinelle Tätigkeiten abgelöst wurden, stellt sich durchaus die Frage, ob gemäß der Idee

des *cognitive enhancements* von Menschen[33] nicht auch *bessere* maschinelle Ethik-Maschinen denkbar? In Ian McEwans Roman *Maschinen wie ich* deaktivieren sich humanoide, künstlich intelligente Maschinen aus Unglück angesichts der omnipräsenten menschlichen Fehlbarkeit selbst.

> ‚Was ist mit den übrigen Adams und Eves?', fragte ich.
> ‚Denen geht es gut, soweit ich weiß.'
> ‚Einige sind unglücklich, habe ich gehört.'
> ‚Das ist nicht wahr.'
> ‚Zwei Selbstmorde in Riad.'
> ‚Unsinn.'
> ‚Wie viele haben den Notschalter deaktiviert?', fragte Miranda. [...]
>
> Sally schien sich zu entspannen. ‚Einige. Es ist unsere Strategie, in solchen Fällen nicht einzuschreiten. Es sind lernende Maschinen, und wenn sie ihre Würde auf diese Weise behaupten wollen, sollen sie das tun.'
>
> ‚Was ist mit diesem Adam in Vancouver?', fragte ich.
> ‚So verzweifelt über die Zerstörung des Urwaldes, dass er die eigene Intelligenz sabotiert hat.'
>
> (McEwan 2019: 256 f.)

Die Erörterung der Möglichkeiten einer genuinen Maschinenethik, die selbstständig denkende und handelnde Techniksysteme behandelt, ist also sowohl auf technisch-pragmatischer Ebene interessant – gemäß der Frage *Was ist technisch möglich?* – als auch auf ethischer – *Was können wir wollen?*. Eine Schwierigkeit besteht dabei in der Individualität von Menschen und ihren unterschiedlichen Vorstellungen davon, was moralisch richtig beziehungsweise was ein *gutes* Leben (s. Abschnitt 4.3) ist, selbst innerhalb kleinerer gesellschaftlicher Kreise – und erst recht kulturübergreifend. (Meta-)ethische Fragen bezüglich moralischer (und ethischer) Maschinen werden allerdings oft mit der Erwartungshaltung möglichst allgemeingültiger Antworten gestellt.

2.4 Methodologischer Hintergrund

Die philosophische Begriffsanalyse steht in der Tradition der *normalen* Sprache.[34] Entgegen der Annahme von Philosophen wie Gottlob Frege, Rudolf Carnap, Bertrand Russell oder Ludwig Wittgenstein (in seinen frühen Werken), der Kern analytischer Philosophie bestehe in der Analyse von Tatsachen in der Welt mithilfe

[33] Hier geht unter anderem und sehr vereinfacht ausgedrückt um die Frage, ob es für Menschen erstrebenswert ist, ihr moralisches Verhalten durch die Zuhilfenahme bestimmter Medikamente sinnvoll, erstrebenswert und moralisch legitim ist.

[34] Gabriel et al. (2017: 36) kommentieren, dass „das Verfahren der Begriffsanalyse spätestens seit Platon in großen Teilen der westlichen Philosophie [zur Anwendung kommt; L. S.], [...] also keine Erfindung der analytischen Philosophie" ist.

einer logisch aufgebauten, idealen Sprache, gehen Vertreter∗innen[35] normalsprachlicher Ansätze in der analytischen Philosophie davon aus, dass eine Untersuchung des alltäglichen Sprachgebrauchs im Hinblick auf notwendige Bedingungen von Begriffserklärungen den größten Erkenntnisgewinn mit sich bringt.[36]
Das Ziel der vorliegenden Arbeit ist eine umfangreiche Analyse der Begriffe *moralische Maschine* und *ethische Maschine*. Wie sich anhand der bislang dargestellten Definitionen bereits gezeigt hat, die ich im Folgenden noch vertiefen werde, entwickle ich im Lauf der Arbeit ein eigenes Verständnis beider Begriffe.[37] Der zweite Ausdruck wird im Deutschen bislang selten verwendet, oder allenfalls in der im Englischen üblichen, scheinbaren Äquivalenz mit *moralischen Maschinen*. Dass und in welcher Form es jedoch sinnvoll sein kann, *ethische Maschinen* als eigenständigen Gegenstand von Technikphilosophie und *genuiner Maschinenethik* zu verstehen, wird sich anhand meines dargelegten Konzepts zeigen.[38]

Zu den Inhalten beider Begriffe und angrenzenden Fragen (Können Maschinen denken? Können sie moralisch handeln? usw.) – je nachdem, wie weit man sie fasst –, haben sich Philosoph∗innen bereits seit mehreren Jahrzehnten bzw. Jahrhunderten Gedanken gemacht. Für die vorliegende Analyse und die Entwicklung eines eigenen Konzepts ethischer Maschinen greife ich Texte, Konzepte, Thesen und Theorien von anderen Autor∗innen, insbesondere Philosoph∗innen und Ethiker∗innen (aber auch Soziolog∗innen oder Historiker∗innen), auf. Die Datenbank *Science Direct* zeigt für das Jahr 2021 unter dem Stichwort *artificial intelligence* 27.130 Artikel an.[39] Zusammen mit dem zusätzlichen Ausdruck *ethics* sind es immer noch 2.091 Artikel. Der Ausdruck *moral machine* führt für das Jahr 2021 zu 1.381 Artikeln – genauso wie bei der vorherigen Suche steigt die Anzahl der Artikel von unter hundert pro Jahr ab 1998 bis heute stetig an. Dies führt die Bedeutung des Themas vor Augen – zumal bei diesen Zahlen weder deutschsprachige Artikel noch unkonventionelle Formate wie Blogbeiträge eingerechnet sind.

Aus methodologischer Sicht erschwert ein solch großer Fundus natürlich die Auswahl der Texte und Konzepte beziehungsweise deren fachwissenschaftlicher Rechtfertigung; für die vorliegende Analyse habe mich zum einen am gängigen Kanon maschinenethischer Texte (und angrenzenden Begriffen wie Roboterethik

[35] Zur Repräsentation von Frauen in der analytischen Philosophie sowohl im historischen Verlauf als auch in zeitgenössischen Werken vgl. Müller-Salo 2020: 17 f.

[36] Dieser Ansatz wurde vor allen Dingen von John Austin, dem späten Ludwig Wittgenstein und Gilbert Ryle geprägt.

[37] Wie ich unten noch erläutern werde, scheint es mir, als beabsichtigten manche Autor∗innen geradezu ein nebulöses Verständnis der Begriffe, um möglicherweise Thesen (bespielsweise eine eindeutige Trennung zwischen Menschen und Technik) implizit aufrechtzuerhalten zu können.

[38] Weitere methodologische Erläuterungen finden sich in Abschnitt 5.2.

[39] Vgl. https://www.sciencedirect.com/, aufgerufen am 01.03.2022.

usw., s. o.) im deutsch- ebenso wie englischsprachigen Raum orientiert. Ich habe dabei sowohl die klassischerweise diskutierten Texte einbezogen (vgl. z. B. Loh 2019a: 13, Fußnote 7; Misselhorn 2019: 70 ff.), die im englischsprachigen Raum bis in die 1950er-Jahre zurückreichen, als auch aktuelle Publikationen, die teilweise erst kurz vor der Abgabe meiner Arbeit erschienen sind. Wichtiger als Faktoren, die die Verbreitung eines Textes widerspiegeln, also beispielsweise Zitationszahlen, war mir dabei die fachliche Relevanz und Angemessenheit für meine Analyse und Argumentation. Die Auswahl der Texte kann im Hinblick auf die Plausibilität der Argumentation kritisiert oder konstruktiv: ergänzt werden. Neben international publizierten wissenschaftlichen Beiträgen in größeren Verlagen greife ich dabei auch journalistische Publikationen auf, sofern die Inhalte für die Argumentation relevant – oder auch unterhaltsam sind. Angesichts der Fülle an Publikationen und des Begriffsbündels, das meine Analyse umfasst (wenn man alle einzelnen technischen Bezeichnungen einbezieht, umso mehr), erhebt diese Arbeit bezüglich der Auswahl der Texte mitnichten Anspruch auf Vollständigkeit. An vielen Stellen können die ausgewählten Texte und Konzepte aufgrund ihrer Argumentation als exemplarisch verstanden werden.

3 Zur Geschichte maschineller Moral und Ethik

> Die Mechanik lieferte im 18. Jahrhundert aber nicht nur die theoretischen Grundlagen für den Bau solcher Figuren, die musizieren und schreiben konnten, sondern prägte das ganze Weltbild: Der Staat, ja die Natur und somit auch der Mensch, zumindest sein Körper, seien, so die Ansicht der rationalistischen Philosophie, mechanische Systeme, die ausschließlich den Gesetzen der Physik gehorchten.
> (Drux 1994b: 16 f.)

Die Geschichte maschineller Moral und Ethik ist zeitlich schwer zu umreißen, da, wie bereits beschrieben, auf technischer Seite viele Begriffe im Raum stehen: Computer, Roboter, Automaten, Maschinen, Androiden, künstliche Intelligenz sind nur einige Begriffe, die hier in den Fokus genommen werden können. Außerdem sollten auch Begriffe wie *moralischer Akteursfähigkeit* beziehungsweise *moral agency* und ihre Definition im jeweiligen zeithistorischen Kontext beachtet werden. Dies gilt insbesondere für den Begriff *Algorithmus*. Diese sind nicht auf die Anwendung mit einer bestimmten technischen Grundlage (zum Beispiel Computerprogramme) beschränkt, sondern Prinzipien, die auch in anderen Bereichen und Systemen zum Tragen kommen können – Sebastian Stiller (2015) spricht zum Beispiel auch von *algorithmischem Denken*. Der Begriff *künstliche Intelligenz* kann wiederum so vielseitig definiert werden, dass auch hier, je nach Bestimmung, unterschiedliche Ursprünge ausgemacht werden können (vgl. hierzu auch Seng 2019a: 3 f.).

Raimundus Lullus[40], ein mallorquinischer Philosoph und Logiker, entwickelte Ende des 13., Anfang des 14. Jahrhunderts ein System, das er als *Ars magna* bezeichnete, *Große Kunst* (vgl. Lullus 1999). Dafür gestaltete er sieben drehbare Scheiben, die mittels Linien in gleichgroße Stücke aufgeteilt wurden. In diesen Feldern standen verschiedene lateinische Begriffe, auf einer Scheibe etwa: *veritas*, *gloria*, *bonitas*, *magnitudo*, *duratio*, *potestas*, *sapientia*, *voluntas*, *virtus*. Durch das Drehen aufeinanderliegender Scheiben ergaben sich Kombinationen, die logische Schlüsse zuließen, woraus Lullus Aussagen über die Verbindung der jeweiligen Begriffe folgerte. Mithilfe dieses Systems, das auch als „mittelalterlicher Computer“ bezeichnet wird, verfolgte Lullus das Ziel, „Ungläubige zum Christentum zu bekehren“ (Duda 2016).[41]

Das Feld der Technikgeschichte ist eine eigene Disziplin, in der in den vergangenen Jahren viel geforscht wurde und wird. Dabei geht es nicht nur um die Dokumentation der technischen Entwicklungen, sondern auch um anthropologische

[40] Zuweilen findet man auch die Schreibweise *Ramon Llull*.

[41] Mehr Beispiele zum Ursprung moderner Computer finden sich u. a. auch bei Ullrich 2019b: 248 ff.

Aspekte des Umgangs von Menschen mit verschiedenen Technologien. In diesem Kapitel werde ich einige Beispiele exemplarisch aufführen, die auch wichtig sind für das Verständnis zeitgenössischer Diskurse und Zukunftsaussichten.

3.1 Technik und Menschenbild

Der aktuelle Entwicklungsstand der Technik einer jeweiligen Zeit bedingt in der Regel stark das jeweils vorherrschende Menschenbild. Dies hängt auch damit zusammen, dass (insbesondere umfangreiche) technische Entwicklungen meistens auch Auswirkungen auf gesellschaftliche Strukturen, soziale Gefüge, Arbeits- und Lebensumstände haben (vgl. hierzu u. a. Seng 2019b: 59). Je nach Perspektive können diese technischen Umschwünge positiv bewertet werden, oder aber es wird der Fokus auf die „problematischen Folgen der Techniknutzung“ (Ropohl 1991: 148) gelenkt. Eine pauschale, insbesondere ethische Beurteilung technischer Systeme greift zu kurz, wie Thilo Hagendorff (2020: 111) im Kontext seiner Analyse von KI-Regulierungsansätzen bemerkt:

> The ethics guidelines examined refer exclusively to the term ‚AI‘. They never or very seldom use more specific terminology. However, ‚AI‘ is just a collective term for a wide range of technologies or an abstract large-scale phenomenon. The fact that not a single prominent ethical guideline goes into greater technical detail shows how deep the gap is between concrete contexts of research, development, and application on the one side, and ethical thinking on the other. Ethicists must partly be capable of grasping technical details with their intellectual framework. That means reflecting on the ways data are generated, recorded, curated, processed, disseminated, shared, and used [...], on the ways of designing algorithms and code, respectively [...], or on the ways training data sets are selected [...]. In order to analyze all this in sufficient depth, ethics has to partially transform to ‚microethics‘. This means that at certain points, a substantial change in the level of abstraction has to happen insofar as ethics aims to have a certain impact and influence in the technical disciplines and the practice of research and development of artificial intelligence [...]. On the way from ethics to ‚microethics‘, a transformation from ethics, to data ethics has to take place. As long as ethicists refrain from doing so, they will remain visible in a general public, but not in professional communities.[42]

Die ethischen Reflexionen über Technik unterscheiden sich dabei freilich von subjektiven Bedenken angesichts der Einführung neuer Technologien, die es zu jeder Zeit gab – angefangen mit Platons Kritik an der Einführung der Schrift (vgl. Wagner 2014) bis hin zur Kritik an häufig so bezeichneten *neuen* Medien und umfassenden Transformationen, wie der Digitalisierung (vgl. z. B. Spitzer 2012).

[42] Je nach Perspektive könnte es auch sinnvoll sein, von einer *Makro*-Ethik zu sprechen, in dem Verständnis der Konzentrierung auf einen sehr spezifischen Themenbereich im Unterschied zu der oberflächlichen Behandlung von Themen. Für diesen Hinweis danke ich Julie Schweer.

3.1.1 Frühe Neuzeit: der Mensch als Maschine

Die Folgen von Technikeinsatz und -Nutzung gehen meistens einher mit grundlegenden Änderungen in sozialen Strukturen, Infrastruktur, Arbeitsmöglichkeiten und -verhältnissen usw., und damit wird auch das Selbstbild von Menschen beeinflusst. Wie die Technikentwicklung das Menschenbild einer Zeit beeinflussen kann, wird zum Beispiel deutlich, wenn man einen Blick in die frühe Neuzeit beziehungsweise die Anfänge der Industrialisierung wirft. Häufig verwiesen wird auf René Descartes' mechanisiertes Menschenbild und die in dieser Hinsicht noch radikaleren Theorien Julien Offray de la Mettries, die auf Descartes' Thesen aufbauen (Offray de La Mettrie 1909). Der neben einigen alternativen Ansätze bis ins heutige westliche, schulmedizinische System verankerte Dualismus von Körper und Psyche (in der klassischen Philosophie oft auch als *Leib* und *Seele* bezeichnet) geht unter anderem auf den Rationalisten René Descartes zurück und wird daher auch als *Cartesianischer Dualismus* bezeichnet. Dieser ähnelt der aristotelischen Physiologie und Seelenlehre, in der es zwei ontologisch völlig verschiedenen Entitäten gibt, die das Verhältnis zwischen Körper (sóma) und Seele (psyché) bestimmen: Materie (hýle) und Form (eídos) (vgl. An: II 412b6–9). Während der Körper gemäß der aristotelischen Philosophie physiologisch greifbar ist, ist die Seele das grundlegende Prinzip, welches das Wesen (ousía) von Menschen konstituiere. Im wissenschaftlichen Diskurs ist allerdings umstritten, wie Aristoteles in der modernen Philosophie des Geistes zu verorten ist; hier gibt es sowohl Positionen, die ihn einem prinzipiellen Dualismus zuordnen als auch monistischen Ansätzen.[43] René Descartes wird hingegen in der Regel als (Substanz-)Dualist eingeordnet, da er seiner Körper-Geist-Philosophie eine prinzipielle Trennung zweier Substanzen zugrundelegt: die körperliche Substanz (*res extensa*) und die des Geistes (*res cogitans*). Erstere zeichnet sich, ganz grundlegend formuliert, dadurch aus, einen Platz im Raum einzunehmen, also *ausgedehnt* zu sein, und letztere dadurch, das Denken zu übernehmen.[44] Ähnlich wie in der aristotelischen Philosophie gelingen viele menschliche Fähigkeiten demnach nur durch das komplexe Zusammenspiel beider Substanzen. Descartes stellte dabei für die Darstellung der körperlichen Substanz

[43] Howard Robinson (2009) argumentiert beispielsweise gegen die Interpretation, die Seelenlehre Aristoteles könne im modernen Funktionalismus verortet werden; Robert Heinaman (1990) charakterisiert die aristotelische Philosophie ganz klar als dualistische und Richard Sorabji (2009: 64) argumentiert dafür, den aristotelischen Ansatz als „something *sui generis*“ (Herv. i. Orig.) zu verstehen. Zahlreiche weitere Positionen könnten hier genannt werden.

[44] Insbesondere in außerakademischen Kontexten wird Descartes' Seelentheorie zuweilen mit derjenigen Platons verglichen und gemeinsam monistischen Ansätzen gegenübergestellt. Dies wird zum Beispiel von Sarah Broadie (2015) kritisiert. Schmitt (2011) diskutiert in Kapitel I (S. 1–23) inwiefern die philosophischen Theorien Platons und Descartes' in Bezug auf *Denken* und *Sein* überhaupt zu vergleichen sind.

von Menschen eine Analogie zu den zu seiner Zeit – dank der zunehmend effizienten Metallherstellung mittels Hochöfen in Europa – vermehrt aufkommenden Automaten her.[45] Die Analogie von Menschen (beziehungsweise menschlichen Körpern) mit Maschinen stellt Descartes im fünften Teil der *Discours de la méthode* auf (erschienen erstmals 1637 in Leiden auf Französisch, 1656 auf Latein in Amsterdam) und deklariert dabei – aus Angst, der Gotteslästerung bezichtigt zu werden – auch eindeutig Gott als den *Erbauer* der menschlichen Maschinen (vgl. Disc. Meth. AT: V). Aufgrund seiner Darstellung wurde Descartes auch als *Balonist* bezeichnet, da er sich die Nerven „als hohl und mit Ventilen versehen“ vorstellte, wie es die Wissenschaftsjournalistin Susanne Donner (2016) formuliert.

> In Armen und Beinen des Körpers verschmelzen sie dann mit den Muskeln. Der gasförmige Lebensgeist [*spiritus animalis*; L. S.] lässt so die Muskeln hart werden und pumpt gewissermaßen die Gliedmaßen auf.
> (Ebd.)

Meilensteine der technischen Entwicklung verhältnismäßig kurz vor Descartes' Lebenszeit waren die Erfindung mechanischer Räderuhren im 14. Jahrhundert sowie die Revolution des Buchdrucks durch Johannes Gutenberg im 15. Jahrhundert. Seitdem konnten sich unter anderem wissenschaftliche Werke schneller verbreiten als zuvor.

3.1.2 Renaissance: die Zeit der Automaten

Die Technik mechanischer Uhren wurde insbesondere im 18. und 19. Jahrhundert zur Herstellung von Automaten benutzt, bei denen mitunter der Vorsatz bestand, menschenähnliche Fähigkeiten zu imitieren. Rudolf Drux (1994b: 16 ff.) nennt als Beispiele unter anderem die „Pianistin am Klavier mit Spielwerk“, eine französische Arbeit um ca. 1800, sowie den (von Drux so bezeichneten) *Androiden* „der Schreiber“, 1774 konstruiert von Pierre Jaquet-Droz:

> Die Mechanik lieferte im 18. Jahrhundert aber nicht nur die theoretischen Grundlagen für den Bau solcher Figuren, die musizieren und schreiben konnten, sondern prägte das ganze Weltbild: Der Staat, ja die Natur und somit auch der Mensch, zumindest sein Körper, seien, so die Ansicht der rationalistischen Philosophie, mechanische Systeme, die ausschließlich den Gesetzen der Physik gehorchten.
> (Ebd.: 16 f.)

[45] „Die Metallfeder mit der ihr eingeprägten Kraft markiert m. E. den wegweisenden Schritt in der Automatenbaukunst der Frühen Neuzeit. Denn sie steht für einen gänzlich anderen Antrieb als die Wasserkraft.“ (Engel und Karafyllis 2004: 328).

Ausgehend vom „maschinellen Wesen des Menschen“ und anhand zahlreicher Vorbilder des „genialen Automatenbauers Jacques de Vaucanson“ (bekannt ist unter anderem seine *mechanische Ente*[46]) komplettierte Julien Offray de La Mettrie die descart'sche Sicht in seinem Werk *L'homme machine*, das 1748 erschien (alle Zitate Drux 1994b: 17).

> Wer heute ihre späten Nachkömmlinge betrachtet, aufziehbare Spielfiguren aus Blech, die Rad und Takt schlagen, trommeln und trompeten, wird kaum auf den Gedanken verfallen, daß einst ein aufgeklärter Materialist die Vaucansonschen Automaten, die gewiß kunstvoller, aber doch im Prinzip dem blechernen Kinderspielzeug gleichzustellen sind, als artifizielle Pendants des natürlichen Menschen verstand [...]. Wenn der Ingenieur, so La Mettrie, etwas ‚mehr Kunst anwenden‘ würde, dann könne aus einem Androiden, der Flöte spielt, sogar ein ‚Sprecher‘ werden und sich damit das größte Geheimnis der Natur, die dem Menschen vorbehaltene Gabe der Sprache, auf technische Weise erschließen. Von daher kann es nicht verwundern, daß La Mettrie den Automatenbauer J. de Vaucanson in mythische Dimensionen rückt und in ihm einen ‚neuen Prometheus‘ sieht.
>
> (Ebd. Die Binnenzitate beziehen sich auf Offray de La Mettrie (1909: 46).)

Aus diesen Gedanken wird bereits das frühe Bestreben, Menschen technisch nachzubauen – oder dies zumindest potenziell können zu wollen – ersichtlich, wenngleich die technischen Möglichkeiten im Vergleich mit heutiger Technik noch stark eingeschränkt waren.

Die jeweils aktuellen technischen Möglichkeiten beeinfluss(t)en dabei auch die Sprache der Menschen (s. Abschnitt 3.3) und führten zu unterschiedlichen Metaphern und teilweise „waghalsiger Metaphorik“, wie Günter Ropohl (1991: 154) den Ausdruck *künstliche Intelligenz* einordnet. Computer wurden in ihrer Anfangszeit beispielsweise oft als „Elektronengehirne“ oder „Denkmaschinen“ beschrieben (ebd.) – wobei hier die Analogie noch vom menschlichen Gehirn hin zur Technik gezogen wurde und nicht, wie später, umgekehrt. Stephan Ullrich (2019a: 17) merkt an:

[46] „Das Publikum war exquisit. Ein Knistern ging durch die seidenen Toiletten: Phantastisch! Ein Chef-d'œvre: die mechanische Ente. Auch Diderot war begeistert. Der Automat watschelte, planschte im Wasser: Welche Delikatesse in allen Teilen!“ (Enzensberger 2000: 34 ff. zitiert nach Schulenburg 2007).

> Aber auch die Politik bedient sich am technischen Vokabular. Seit der Weimarer Republik wird die Wirtschaft *angekurbelt*, werden bestimmte Reformen *angestoßen* und Prinzipien fest *verankert*.
> (Herv. i. Orig.)

Die Geschichte des Mensch-Technik-Verhältnisses und die Beschreibung des Einflusses technischer Entwicklungen auf Menschenbilder könnte hier freilich bis in die Gegenwart fortgesetzt werden.

3.2 Technische Errungenschaften – zwischen Fortschritt und Bedrohung

Vom „Beginn seiner Existenz“ an, schreibt Thimm (2019: 18), sei das „Leben des Homo sapiens“ „von dem Spannungsfeld zwischen der nutzenbringenden und der gefährlichen Technik gekennzeichnet“. In meinen Seminaren zur Maschinenethik an der Pädagogischen Hochschule in Ludwigsburg (2016–2021) fragte ich die Studierenden oft zu Beginn des Semesters, was Menschen ihrer Ansicht nach wesentlich ausmache – um von hier aus zum wesentlichen Kern von Maschinen und künstlicher Intelligenz (und ggf. entscheidenden Abgrenzungen zu Menschen) zu kommen. Und um dem häufig medial vermittelten diffusen Bedrohungsgefühl[47][48] gegenüber künstlicher Intelligenz einen sachlichen, fachwissenschaftlichen Boden zu geben. Wir differenzierten notwendige und hinreichend Bedingungen ebenso wie substantielle und akzidentielle Eigenschaften.

Welche Eigenschaften sind hinreichend für das Menschsein und gleichzeitig ausschließlich Menschen vorbehalten? Gibt es diese? Die Antworten der Studierenden waren häufig: eine Seele haben, Bewusstsein haben, lieben können, die Fähigkeit zur Reflexion. Allerdings wandten andere direkt ein, dass man viele dieser Eigenschaften auch auf Tiere übertragen könne, andere wiederum nicht bei allen Menschen notwendigerweise Voraussetzung wären, zum Beispiel bestimmte kognitive Fähigkeiten. Geborenwerden und Reproduktion sind ebenfalls Aspekte, die in Abgrenzung zu Maschinen häufig genannt wurden.

[47] Meinecke und Voss (2018: 206–209) argumentieren, dass diese negative Sicht sowohl durch häufig ebenfalls dystopische Science-Fiction-Narrative beeinflusst werde, umgekehrt aber auch „real life robotics“ Autor∗innen von Science-Fiction-Geschichten inspiriere. Zu Science-Fiction-Narrativen s. auch Abschnitt 3.5.

[48] S. hierzu auch die Ergebnisse der auf S. 118 referierten, repräsentativen Umfrage Bertelsmannstiftung.

3.2.1 Mensch-Tier-Verhältnis als Vergleichsobjekt

Viele der angeführten Eigenschaften oder Fähigkeiten würde man intuitiv möglicherweise als kategoriale Unterschiede im Vergleich zu Maschinen gelten lassen; im Verlauf des Seminars haben wir dies mit philosophischen Methoden überprüft. Dabei fiel immer wieder der Vergleich von Maschinen jeglicher Art und Tieren auf; Kate Darling (2021: xiv f.) sieht viele Parallelen von Mensch-Tier- und Mensch-Maschine-Verhältnissen – und betont den bereits angesprochenen Aspekt der Ersetzung menschlicher Fähigkeiten im Kontext in ersteren:

> Throughout history, we've used animals for work, weaponry, and companionship. Like robots, animals perceive and engage with the world differently than humans. That's why, for millenia, we've relied on animals to help us do things we couldn't do alone. In using these autonomous, sometimes unpredictable agents, we have not replaced, but rather supplemented, our own relationships and skills. [...] Using animals to think about robots acknowledges our inherent tendency to project life onto this technology, something that has fascinated me for years. [...] In comparing robots to animals, I'm not arguing that they are the same. Animals are alive and can feel, while robots suffer no differently than a kitchen blender. [...] The point is that this thought exercise lets us step out of the human comparison we're clinging to and imagine a different kind of agent.

Auch Vergleiche von Tierethik zu Maschinenethik werden immer wieder gezogen. In der Übersetzung von Janina Loh (2019a: 10) ist für David Gunkel beispielsweise „[d]ie Frage nach der Maschine das Gegenstück zu der Frage nach dem Tier“ (vgl. Gunkel 2012: 5). In einer Kurzgeschichte Isaac Asimovs, *Runaround*, ist die Rede von einer *metallenen Pfote* (*metal paw*) eines Roboters (vgl. Asimov 2016: 72). Und auch bei ethischen Überlegungen, wie das Mensch-Maschine-Verhältnis moralisch gut gestaltet werden könne, spielen Vergleiche zu analogen Überlegungen in Bezug auf Mensch-Tier-Verhältnisse immer wieder eine Rolle (siehe z. B. das *Verrohungsargument* in Fußnote 31).

Auch bezüglich der Haftung teilautomatisierter Fahrzeuge und Tieren können Parallelen gezogen werden, wobei ethische und juristische Fragen dabei getrennt werden sollten. Denn wenngleich teilweise ähnliche oder dieselben Begriffe benutzt und diskutiert werden, so werden sie in der Regel mit juristischen Methoden völlig anders behandelt als mit ethischen. Beim Halten von Tieren ebenso wie dem Besitz von Fahrzeugen (§ 7, StVG) gilt die sogenannte *Halterhaftung* (§ 833 BGB), wonach die Verantwortung prinzipiell den menschlichen Halter∗innen zugeschrieben wird. Auf der Seite moralischer Maschinen ist dies keineswegs so eindeutig, sofern sie selbst als „Verantwortungssubjekte“ (vgl. Loh 2017: Abschnitt 3.1, S. 138 ff.) akzeptiert werden, denen zumindest in gewissem Rahmen Akteursfähigkeit zugeschrieben werden kann. Häufig diskutiert werden auch Rechte – sowohl bei

Tieren als auch bei technischen Objekten. Tiere, denen beispielsweise im deutschen Recht (TierSchG seit 1972) ein moralischer Objektstatus zukommt, haben gleichzeitig zumindest in der Theorie in bestimmten Kontexten auch gewisse Rechte. So müssen Tierhalter∗innen ein Tier qua Gesetz beispielsweise „seiner Art und seinen Bedürfnissen entsprechend angemessen ernähren, pflegen und verhaltensgerecht unterbringen“ (TierSchG Abschnitt 2, §2). Immer wieder wird jedoch die breite Auslegbarkeit solcher Gesetzestexte in der Praxis bemängelt, in der eine Tierhaltung (bspw. in der Fleisch- und Milchproduktion) möglich ist, die nach Ansicht vieler Tierschützer∗innen, -rechtler∗innen u. a. nicht den juristischen Vorgaben entspricht (vgl. z. B. Kainz 2021).[49]

Die Praxis umstrittener Umgangsweisen von Menschen mit Tieren ist Zeugnis des vorherrschenden Machtgefälles im Mensch-Tier-Verhältnis – fraglich ist dabei, wie das Verhältnis zwischen Menschen und Technik gemessen an diesem Beispiel aussehen könnte und würde, unter der Voraussetzung der Möglichkeit zu eigenständigem Denken und möglicherweise sogar Bewusstsein fähiger, ethischer Maschinen. Dies betrifft Fragen nach der Übernahme von Verantwortung ebenso wie individuelle Rechte und Privilegien.

3.2.2 Technische Ersetzung menschlicher Fähigkeiten – Vor- oder Nachteil?

Im Lauf der Technikgeschichte wurden immer wieder vermeintliche Alleinstellungsmerkmale des Menschen als technisch zumindest imitierbar, manchmal auch als ersetzbar, deklariert. Dies wurde und wird, je nach Kontext, Umfang und Interpretationsperspektive als „Entlastung und Befreiung“, als „Bedrohung“ oder als „das Überflüssigwerden der Menschen sowie die Vernichtung der Menschheit und ihre Ersetzung als Gattungswesen durch Maschinen“ wahrgenommen (alle Zitate Heßler 2020a: 263). Letzteres ist Gegenstand vieler Science-Fiction-Szenarien, die heute noch Dystopien (seltener Utopien) zeichnen, wobei die alltägliche Wahrnehmung von Robotern als „inanimate and animate at the same time“ nach Voss (2021: 13)

[49] Die unterschiedliche Auslegbarkeit des sogenannten Tierschutzgesetzes wird bereits im ersten Paragraphen deutlich: „Niemand darf einem Tier ohne vernünftigen Grund Schmerzen, Leiden oder Schäden zufügen“. Die Entscheidung über die Vernünftigkeit der Gründe obliegt offensichtlich Menschen, die in der Lebensmittelproduktion primär marktwirtschaftliche Interessen verfolgen. Dabei belegen tierethische Studien inzwischen die diversen kognitiven und emotionalen Fähigkeiten von Tieren (vgl. Fischer 2021; Schmitz 2017a), aufgrund derer man zu dem Ergebnis kommen kann, dass es vernünftig ist, Massentierhaltungsbetriebe auf der Basis der Unvereinbarkeit mit dem ersten Paragraphen des Tierschutzgesetzes abzuschaffen. Zusätzliche Argumente wie die Klimaschädlichkeit von Massentierhaltung können angeführt werden. Die Tatsache, dass solche ethischen Abwägungen bislang scheinbar nicht in der breiteren Praxis ankommen, beruht auf dem gegenseitigen Einfluss politischer Entscheidungen und wirtschaftliche Bedingungen.

längst keine Fiktion mehr ist. Laura Voss hat dabei explizit „Robot Technology“ im Blick, wobei sie darunter sowohl frühe Versionen von Unterhaltungs-Automaten versteht als auch Industrieroboter und selbstständig agierende Haushaltsgeräte:

> In recent years, robots have been making another important step. They have made their way out of their factory cages and out of robotics laboratories, entering private homes and public spheres to be employed in close physical and social proximity to humans.
>
> (Voss 2021: 11)

Martina Heßler (2020a: 263) sieht generell eine Tendenz zur negativen Wahrnehmung, da „Technik, die menschliche Tätigkeiten übernimmt, [...] oft undifferenziert als Gefahr für die Menschen bezeichnet [wird; L. S.], als etwas, das sie überflüssig macht“.

Während es zur Zeit der Industrialisierung in erster Linie um die Ersetzung menschlicher Körper als Arbeitskräfte durch in wirtschaftlicher Hinsicht effizientere Maschinen ging,[50] geht es heute um sämtliche Funktionen und Fähigkeiten von Menschen, die potenziell von Maschinen und Programmen imitiert und zu verschiedenen Zwecken ersetzt werden können (sollen), also auch um Emotionen, Kognition, Bewusstsein und moralisches Handeln.[51] Dies ist die Konsequenz einer langen Reihe von Entwicklungen in den letzten Jahrhunderten und Jahrzehnten, welche die Klärung der Rolle der Technik im Handlungs- und Interaktionsgefüge immer dringlicher gemacht hat.

Unabhängig von der Frage, welcher moralische Status Maschinen und Programmen zugeschrieben werden kann, die ich im anschließenden Kapitel diskutieren werde, haben Menschen Maschinen seit jeher (ebenso wie Tieren) menschliche Fähigkeiten und Eigenschaften attestiert. So werden beispielsweise Alltagsgegenständen menschliche Handlungsfähigkeiten zugeschrieben, Menschen geben (nicht nur, aber auch) technischen Gegenständen Namen oder übertragen ihnen die Verantwortung für bestimmte Ereignisse, entwickeln Emotionen gegenüber technischen Systemen. Dies wird auch als *object personification* oder *anthropomorphism* bezeichnet. Der Ausdruck *mechanomorphism* beschreibt umgekehrt die Zuschreibung maschineller Eigenschaften für Menschen (vgl. Brédart 2021; Caporael 1986). Solche Prozesse können, je nach Perspektive, entweder das Miss- oder aber das Vertrauen in technische Objekte verstärken, und muss aus Gründen der Praktikabilität bei der

[50] „Die Maschine wurde im 19. Jahrhundert zur Metapher für reibungsloses Funktionieren, für rational nachvollziehbare, gleichförmige und durchschaubare Abläufe sowie gleichermaßen für menschliche Unfreiheit und Unflexibilität.“ (Heßler 2020c: 256).

[51] Details dazu diskutiere ich in Abschnitt 5.2.

Entwicklung technischer Systeme, zum Beispiel automatisierter Fahrzeuge, berücksichtigt werden (vgl. Waytz, Heafner und Eply 2014). Durch die Personifizierung werden entsprechende technische Gegenstände auf einen moralischen Subjektstatus gehoben, auch wenn das den Menschen vielleicht nicht immer bewusst ist. Schimpft eine Person mit Ihrem Computer, suggeriert das, dass dieser einen Handlungsspielraum (gehabt) habe, innerhalb dessen er sich auch für *richtiges* Verhalten hätte entscheiden beziehungsweise auf *richtige* Weise hätte funktionieren können.[52] Es werden also im alltäglichen Umgang mit technischen Systemen häufig intuitiv Vorannahmen bezüglich Subjektstatus und damit verbundener Verantwortungsfähigkeit gemacht, die es auf wissenschaftlicher Ebene zu überprüfen gilt.

3.2.3 Ambivalente Techniknutzung und -bewertung

Die Bewertung technischer Systeme fällt also, je nach Konstruktionsabsicht, Umsetzung, Einsatz und Output sehr ambivalent aus. Die Varianz an Bewertungen wird noch vergrößert, da nicht nur, wie oben dargestellt, bestimmte technische *tools* mit bestimmten Absichten oder für einen bestimmten Zweck entwickelt und hergestellt werden, sondern da natürlich ursprüngliche Zwecke auch missbraucht werden können. Die Entwicklung von Programmen, die auf maschinellem Lernen basieren bringt dabei neue Herausforderungen mit sich, wie zum Beispiel *deep fake*-Videos (vgl. van Huijstee et al. 2021). Diese können sowohl mit positiven als auch negativen Intentionen eingesetzt werden.

> The broad range of possible risks can be differentiated into three categories of harm: psychological, financial and societal. Since deepfakes target individual persons, there are firstly direct psychological consequences for the target. Secondly, it is also clear that deepfakes can be created and distributed with the intent to cause a wide range of financial harms. Thirdly, there are grave concerns about the overarching societal consequences of the technology.
>
> (Ebd.: IV)

Es liegt in der Natur der Techniknutzung, dass technische Systeme positive oder negative Auswirkungen haben können. Dies ist jedoch nichts, was *der Technik* per se innewohnt, wie die folgende Formulierung von Lenk und Ropohl (1993a: 5) nahelegt, sondern was – wie im März 2022 auf erschreckende Weise weltpolitisch sehr deutlich wird – durch den Einsatz von Menschen bestimmt wird und darum auch in ihrer Verantwortung liegt:

[52] Wie ich im Kontext der Erörterung notwendiger Bedingungen für *ethisch sein* bei Menschen und technischen Systemen diskutieren werde, gibt es Ansätze, wonach Techniksysteme dann *rational* agieren, wenn sie gemäß der vorgesehenen Weise funktionieren. Diese muss jedoch nicht zwangsläufig den Erwartungen der Benutzer∗innen entsprechen.

> Die Technik hat Berge versetzt und Flüsse verlegt, aber auch Wälder zerstört und Städte zerrissen. Die Technik hat Märchenwünsche erfüllt, aber auch Alpträume wahrgemacht. Das Schlaraffenland als Automatenparadies oder die Apokalypse des atomaren Infernos – beide Zukunftsvisionen können Wirklichkeit werden.

Demnach müsste also der Ausdruck *Technik* in diesem Zitat mit *die menschliche Entwicklung und Nutzung von technischen Systemen* ersetzt werden. Die Rede von *der* Technik suggeriert einen Subjektcharakter mit allen bereits angesprochenen Folgen, der noch zu untersuchen ist und allenfalls für bestimmte technische Systeme, nicht aber allgemein gelten kann.

Die Techniklandschaft sah im Jahr 1993, in dem weder Digitalisierung noch die Erforschung maschinellen Lernens so weit vorangeschritten war wie heute, natürlich anders aus. Die Aussage, dass technische Systeme, egal welcher Komplexität, prinzipiell zum Vor- oder Nachteil von Menschen eingesetzt werden können, ist jedoch zeitlos. Bereits früher merkte Günter Ropohl selbst an, dass die „*pauschale* Technikbewertung [...] auf schlichten *Denkfehlern* und ideologischen *Vorurteilen*“ (Ropohl 1985: 12, Herv. i. Orig.) beruhe, wobei es ihm an dieser Stelle um die Differenzierung verschiedener technischer Systeme geht.

> Man behandelt ‚die‘ Technik, als ob sie eine gleichförmige und selbstständige Wesenheit wäre, man stilisiert bestimmte Vorzüge oder Fehler einzelner technischer Erscheinungen zum Segen oder Fluch ‚der‘ Technik schlechthin, und man überschätzt oder übersieht die selbstverständlich gewordenen Leistungen vieler technischer Errungenschaften.
>
> (Ebd.: 12)

Diese Einschätzung ist auch heute noch aktuell, wenn man den – teilweise von Wissenschaftler∗innen selbst verbreiteten, teilweise von Medienschaffenden eruptierten – Hype um neuere technische Entwicklungen wie maschinellem Lernen oder maschinellem Bewusstsein (vgl. Wendland 2022) ins Auge fasst. Die Klärung der Frage, ob neueren technischen Systemen und potenziell ethischen Maschinen dabei ein Subjektstatus oder Akteursfähigkeit zugeschrieben werden kann und sie somit auch als potenzielle Verantwortungssubjekt zählen (vgl. u. a. Loh 2019a: 138 ff.; Rath 2019), ist ein Ziel der vorliegenden Arbeit. Daran schließt auch die Frage an, inwieweit technische Systeme sich selbst oder Produkte hervorbringen können, die bislang nur von Menschen produziert wurden, und welche Folgen sich daraus ergeben. Ansätze dafür gibt es bereits, beispielsweise bei wissenschaftlichen Publikationen. Im Jahr 2019 veröffentlichte der Springer-Verlag ein Buch über Lithium-Ionen-Batterien, das aus einer Zusammenfassung von aktuellen Forschungsartikeln zu dem Thema besteht, die ein Computerprogramm eigenständig erstellt hat (vgl. Beta Writer 2019).

> The content is a cross-corpus auto-summarization of a large number of current research articles in this discipline. Serving a structured excerpt from a huge set of papers [...]. In close collaboration between Springer Nature and researchers from Goethe University Frankfurt/Main, a state-of-the-art-algorithm, the so-called *Beta Writer*, was developed to select, consume and process relevant publications in this field from Springer Nature's content platform *SpringerLink*. Based on this peer-reviewed and published content, the Beta Writer uses a similarity-based clustering routine to arrange the source documents into coherent chapters and sections. The extracted quotes are referenced by hyperlinks which allow readers to further explore the original source documents. Automatically created introductions, table of contents and references facilitate the orientation within the book.
>
> (Springer Nature Group 2019: Herv. i. Orig.)[53]

Eine Besonderheit dabei ist, dass das Programm auch Artikel auswertete, die selbst mittels auf maschinellem Lernen basierenden Algorithmen erstellt worden waren, wie eine Forschergruppe herausfand, die computer-generierte Ausdrücke in wissenschaftlichen Artikeln untersuchte und dabei auf zahlreiche maschinell erstellte Fälschungen traf (vgl. Cabanac, Labé und Magazinov 2022). Mittels aktueller Techniken können also nicht nur Fälschungen (im Bereich der Wissenschaft, wie in anderen Bereichen, s. o.) in großem Stil leichter erstellt werden, sondern diese werden wiederum zur Materialgrundlage für weitere KI-erzeugte Produkte. Vor diesem Hintergrund kann die Frage gestellt werden, inwiefern technische Systeme sich dann noch (abgesehen von physischer Beschaffenheit und Erscheinung usw.) noch von Menschen unterscheiden. Diese Frage, die bereits unter dem oben angesprochenen Punkt der möglichen Ersetzbarkeit von Menschen durch technische Systeme anklang, wird im Folgenden auch in Bezug auf die Voraussetzungen für ethische Akteure noch eine Rolle spielen.

Die Frage, ob künstliche Intelligenz als moralischer Akteur gelten, dabei Subjektstatus haben kann und wie die Erzeugnisse bewertet werden, steht dabei schon lange im Raum.

> As artificial intelligence moves ever closer to the goal of producing fully autonomous agents, the question of how to design and implement an artificial moral agent (AMA) becomes increasingly pressing. Robots possessing autonomous capacities to do things that are useful to humans will also have the capacity to do things that are harmful to humans and other sentient beings.
>
> (Allen, Varner und Zinser 2000: 251)[54]

[53] Weitere zu GPT-3 im folgenden Abschnitt.

[54] Eine detailliertere Auseinandersetzung mit und Definition von moralischen Akteure folgt in Abschnitt 4.4.

Armin Grunwald (2020: 73 f.) spricht daher auch von der „Wiederkehr der großen Fragen“[55] im Bereich der Technikethik:

> Die ‚großen Fragen‘ sind jedoch anhand der geschilderten Herausforderungen zurückgekehrt. Gegenwärtig reicht das Spektrum der Technikethik von konkreten und im Detail hoch relevanten [sic!] Fragen wie etwa zu internetbasierten Dienstleistungen, zur Gestaltung der zukünftigen Energieversorgung oder zur Risikobewertung von Nanomaterialien bis hin zu philosophischen Fragen nach der Zukunft im Mensch/Technik-Verhältnis [sic!], oder anders formuliert: nach dem Menschen in einer zusehends technischen Zukunft.

Ob diese technische Zukunft dabei Maschinen mit sich bringen wird, die selbst ethische Reflexionen hervorbringen können und welche Auswirkungen dies haben könnte, wird im weiteren Verlauf diskutiert werden. Gemäß dem dritten Einwand gegen die Verneinung ethischer Maschinen (S. 18 f.) wird zu klären sein, ob und, wenn ja, inwiefern sich (potenziell ethische) Maschinen von (potenziell ethischen) Menschen unterscheiden und ob bezüglich kognitiver (u. a.) Fähigkeiten eine Identität ausgeschlossen beziehungsweise eine Analogie sinnvollerweise angenommen werden kann.

Im folgenden Abschnitt werfe ich anhand des Beispiels *Sprache* einen Blick auf eine mögliche Entwicklung von Technik, der als Grundlage für die Diskussion um die Wahrscheinlichkeit gesehen werden kann, mit der sich Maschinen zu Teilen oder vollständig dem menschlichen Wesen annähern.

3.3 Maschinen auf dem Vormarsch – Beispiel: Sprache

Anhand des Beispiels Sprache (Spracherkennung, Textproduktion, Computerlinguistik usw.) lässt sich gut veranschaulichen, wie sich die Entwicklung von Computerprogrammen und Algorithmen – und die damit verbundenen Annahmen über ihr Potenzial – im Lauf der Zeit verändert hat. Im Folgenden greife ich nur einige Beispiele auf.[56]

Ein Informatiker schenkte mir einmal einen, selbst aus verschiedenen Elementen zusammengelöteten, Stick mit einem eingebauten Knopf. Wenn man den Stick in einen PC steckt und auf den Knopf drückt, generiert er motivierende Sätze: „Das hast Du gut gemacht.“, „Weiter so!“, „Sehr gut!“. Vom Prinzip her ähnlich, nur mit

[55] Ich habe das Phänomen wiederkehrender, „großer“ Fragen wie *Was ist moralisch gut?*, *Was ist ein gutes Leben (unter jeweils aktuellen Bedingungen)?* usw. in Bezug auf aktuelle Entwicklungen in der Technik an anderer Stelle etwas flapsig als *old wine in new bottles* bezeichnet, vgl. Seng 2020.

[56] Für eine ausführlichere Auseinandersetzung vgl. z. B. Höltgen 2021.

einer analogen Anzeige, funktioniert der *Landsberger Poesieautomat* von Hans Magnus Enzensberger, den er im Jahr 2000 im Rahmen des *Lyrik-am-Lech*-Festivals in Landsberg vorstellte. „Mit dem Programm war ich monatelang, fast möchte ich sagen, Tag und Nacht beschäftigt“, schreibt Enzensberger im Hintergrundtext zu seiner Erfindung (Enzensberger 2015: 13). Der Automat, der mittlerweile im Literaturarchiv Marbach steht, generiert eigenständig Texte, die auf einer Anzeigetafel, ähnlich wie auf Flughäfen oder Bahnhöfen, erscheinen. Die Anzeige erfolgt mittels einer Fallblattanzeige, also auf einer Achse rotierender Plättchen, auf denen Buchstaben oder Zeichen gedruckt sind; der Auslöser funktioniert per Knopfdruck. Das erste Gedicht, das der Automat laut einem Bericht von *Der Spiegel* erzeugte, ist das folgende:

> Überflüssige Erpressungen der Gremien, dieser fieberhafte Kunstgenuss am Wochenende und diese vorgedruckten Zahlungsbefehle: Schleierhaft! Im Grunde langweilt uns doch manches. Einstweilen lediglich würgende Lügen. Pünktlich einschrumpfen! Einflüsterungen: („Deine Freunde sind wieder so spießig.“) Im Hinterkopf nur Nullsummenspiele. Das nackte Erbarmen sagt uns mehr als Impotenz, hierzulande schwimmen wir ganz allein. Sachzwänge. Ratlosigkeit. Zierliche Wunderwaffen. Anscheinend klappt alles.
>
> (Der Spiegel 2000)

Enzensbergers Automat gibt die Texte in sechs Zeilen mit 100 Buchstabenfeldern pro Zeile aus (vgl. Gfrereis und Raulff 2015: 343 f.). Die originalen Zeilenumbrüche können hier leider nicht wiedergegeben werden. Das Programm hinter dem Automaten basiert auf einer Matrix aus sechs Zeilen und sechs Spalten oder „Einzelgliedern“ (ebd.: 344). Pro Einzelglied in jeder der Zeilen sind zehn Wortelemente möglich. Insgesamt 216 Wörter kennt Enzensbergers Automat. Die Aneinanderreihung geschieht per Zufallsgenerator. Bei der Zusammensetzung der Wörter gibt es 10^{36} Möglichkeiten, also ebenso viele Gedichte. Zur Bedeutung äußerte sich Enzensberger wie folgt: „Es ist ein Spiel. Wie weit man es mit Sinn auflädt, hängt vom Betrachter ab. Es können Gedichte entstehen, die jemandem was sagen“ (ebd.). Enzensberger äußerte in Interviews zum Festival, bei dem sein Poesieautomat erstmalig vorgestellt wurde, immer wieder die Überzeugung, dass er selbst *bessere* Gedichte verfassen könne als sein Automat. Wie bereits weiter oben deutlich wurde, ist diese Aussage aus technikanthropologischer Sicht insofern interessant, als sie die These impliziert, dass ein Vergleich zwischen Mensch und Maschine notwendig sei und – so kann spekulativ angefügt werden – dass eine potenzielle Ersetzung von Menschen durch die von ihnen entwickelte Technik unbedingt ausgeschlossen werden muss (vgl. hierzu Heßler 2020a; Heßler 2020b). Desweiteren kann die Frage gestellt werden, auf welcher Begründung diese Aussage fußt – ebenso wie im Folgenden zu klären sein wird, ob und inwieweit sich Menschen und technische Systeme bezüglich ethischer Akteurschaft unterscheiden, oder nicht.

Enzensbergers Beispiel ist längst nicht mehr die einzige Umsetzung automatischer Gedichtproduktion. Lukas Diestel, ein Informatiker mit Typ-1-Diabetes, hat ein besonderes Computerprogramm geschrieben:

> Alle 5 Minuten misst ein Silikonfaden in meinem Arm meinen Gewebezucker und schickt ihn an mein Smartphone, an einen Server. Dort wird abhängig vom Wert und ob er für meine Gesundheit gut oder schlecht ist, eine Stimmung ausgewählt. Basierend auf der Stimmung schreibt eine von mir trainierte KI ein Gedicht und lädt es komplett ungelesen auf diese Seite hoch. Ihr könnt euch die ungelesenen Gedichte hier anzeigen lassen. Im Grunde genommen ist das Gedichte-Tinder. Wenn euch ein Gedicht gefällt, könnt ihr euch als Entdecker_in eintragen und es damit in die Galerie befördern. Oder ihr verwerft das Gedicht und bleibt für immer die einzige Person, die es gelesen hat, was definitiv mysteriöse Vibes hat.
>
> (Diestel 2021)

Nach und nach wurden immer neuere Techniken mit immer ausgefeilteren Spracherkennungssystemen entwickelt. Inzwischen gibt es auch zahlreiche Programme, die gesprochene Sprache in Text oder umgekehrt verwandeln können, *text-to-speech* oder *speech-to-text*. Es gibt Programme, die journalistische Meldungen verfassen können (vgl. hierzu Seng 2019b: 57; Latour 2015; Laugée 2014 beziehungsweise Allen et al. 2010) und Algorithmen, die selbst Computer-Code schreiben, also sich in gewisser Hinsicht selbst reproduzieren können (vgl. Pastor und Iborra 2004).

2017 wurde ein künstliches neuronales Netz bekannt – also, vereinfacht ausgedrückt, ein Computerprogramm, das die Verknüpfung von Neuronen im Gehirn in einem theoretischen Modell simuliert –, das auf der Basis von Bildanalysen „pop music“ erzeugt, „where the melody but also chords and other instruments make up what is typically called a song“ (Chu, Urtasun und Fidler 2017: 1). In einer Umsetzung dieses Programms wird ein weihnachtliches Bild mit einem Tannenbaum und Geschenken in Noten übersetzt.[57] Ob es sich dabei tatsächlich um Kunst handelt beziehungsweise, wenn ja, wie diese ästhetisch interpretiert und wer als Urheber bestimmt werden kann, wird diskutiert (vgl. z. B. Dolezal und Windegger 2020). Zu den – wenn man so will – *künstlich künstlerischen* Produkten zählen auch Algorithmen, die anhand von Vorlagen bekannter Maler Fotos in Gemälde verwandeln, die wie in einem dieser jeweiligen Stile gemalt aussehen.[58]

Inzwischen gibt es auch Ansätze für Computerprogramme, die das Maß an Kreativität von Kunst produzierenden oder imitierenden Programmen zu bewerten versuchen (vgl. Elgammal und Saleh 2015) und damit eine weitere, bislang

[57] Vgl. https://vimeo.com/192711856, aufgerufen am 01.03.2022.

[58] Darunter sind zum Beispiel *Das Wrack des Minotaurus* von Joseph Mallord William Turner, die *Sternennacht* von Vincent van Gogh, *Der Schrei* von Edvard Munch oder *Komposition VII* von Wassily Kandinsky (vgl. Gatys, Ecker und Bethge 2015).

von Menschen eingenommene, Rolle, die der bewertenden Metaebene, potenziell besetzen.[59]

In der maschinellen Sprachverarbeitung (auch *NLP* genannt für *Natural Language Processing*) hat sich in den vergangenen Jahren viel getan; zu den aktuellsten Beispielen zählt GPT-3, *Generative Pre-trained Transformer 3*, ein auf der Methode *deep learning* basierendes Sprachmodell des Unternehmens Open AI, das unter anderem durch den Unternehmer Elon Musk finanziell gefördert wird. Im Jahr 2020 wurde das bis dato größte Sprachmodell vorgestellt, das in erster Linie darauf trainiert ist, selbst sinnvollen Text zu erzeugen (Dale 2020). Trotz aller Ehrfurcht, die GPT-3 beziehungsweise seinen Entwickler∗innen teilweise entgegenschlug (vgl. z. B. Simonite 2020), wurden und werden auch immer wieder die Grenzen der Systeme aufgezeigt (vgl. u. a. Floridi und Chiriatti 2020). Elkins und Chun (2020: 12) kommen hingegen zu dem Ergebnis, dass ein auf GPT-3 basierendes Programm im Rahmen einer „judicious selection“ einen schriftlichen Turing-Test (s. S. 15) bestehen könne.

Der bisherigen Steigerung zufolge wäre der nächste konsequente Schritt in der Technikentwicklung, Emotionen, Gefühle und Bewusstsein bei Maschinen ins Feld zu führen – auch angesichts einer Untersuchung der gegenseitigen Notwendigkeit dieser und Sprachverständnis und -produktion. Hierfür gibt es zahlreiche Ansätze und Studien, in denen untersucht wird, ob zumindest letzteres Maschinen und Programmen zugeschrieben werden könne (vgl. Angel 2018; Chrisley 2008; Gamez 2008; Hildt 2019; Irrgang 2020; Kitamura, Tahara und Asami 2000; Mosakas 2021; Reggia 2013; Wendland 2022). Die Diskussionen um Maschinenbewusstsein sind allerdings aufgrund der Schwierigkeiten der Definitionsmöglichkeiten und Erklärungsansätze bereits auf der Seite menschlichen Bewusstseins so komplex, dass ihnen eine eigene Dissertationsarbeit gewidmet werden könnte. Vielleicht werden die Technikphilosoph∗innen schon in ein paar Jahren die heutigen Unsicherheiten bezüglich einer klaren Positionierung im Hinblick auf maschinelles Bewusstsein nicht mehr nachvollziehen können.

Die Entwicklung der maschinellen Sprachverarbeitung und -erzeugung zeigt, dass heute Techniken zur Verfügung stehen, die noch vor zehn Jahren nicht denkbar gewesen wären. Es kann angemessen sein, sich diese Dynamik bei der später folgenden Einschätzung, ob und wenn ja, in welchem Zeitraum genuin ethische Techniksysteme als möglich erachtet werden, in den Sinn zu rufen.

[59] Die Aussagen solcher Analysen hängen natürlich stark von den jeweiligen Definitionen der Begriffe *Kunst* und *Kreativität* ab. Ob und inwiefern sich die menschlichen von den maschinellen Bewertungen unterscheiden, müsste im Detail betrachtet werden, um ggf. Unterschiede (oder nicht) feststellen zu können.

3.4 Technik und Zukunftsvorstellungen

Der Technikphilosoph Günter Ropohl war vor gerade einmal dreißig Jahren mit einer, im Vergleich zu heutigen technischen Entwicklungen, gänzlich anderen Realität konfrontiert. Anfang der 1990er-Jahre hielten *personal computers* gerade langsam Einzug in Privathaushalte, private Internetzugänge waren keine Selbstverständlichkeit. Ropohl ging gemäß dem damaligen Stand der Technik davon aus, dass Probleme für Computer „objektiviert" werden müssten, damit diese sie verstünden. „Wissen und Nicht-Wissen müssen intersubjektiv präzisiert werden." Algorithmen müssten für Computer „Schritt für Schritt" in Vorschriften festgelegt werden und Computer würden dabei lediglich eine *standardisierte künstliche Sprache* verstehen (alle Zitate vgl. Ropohl 1991: 157). Dem gegenüber nahm er auf Seiten der Menschen recht hohe Voraussetzungen als selbstverständlich gegeben an: *subjektive Intuition*, *ganzheitliches Problemverständnis*, *lebendige Umgangssprache*, *Selbstverständlichkeiten des gesunden Menschenverstandes* usw. (ebd.). Die im Rückblick im Bezug auf die Technik naiven, im Bezug auf menschliche Vermögen idealisierend anmutenden Ansichten Günter Ropohls können als stellvertretend für die Einschätzung zahlreicher Wissenschaftler∗innen in der Vergangenheit bezüglich möglicher zukünftiger Technikentwicklungen erachtet werden. So war Ropohl, wie viele andere, auch überzeugt davon, dass sich an der Beschränktheit technischer Systeme, so schnell nichts ändern würde:

> Grundsätzlich wird sich daran auch nichts ändern, wenn sich solche Systeme verbreiten, die, wie es so schön heißt, dialogfähig sind oder sogar natürlich Sprache ‚verstehen'. Erstens gilt das zuvor Gesagte [s. oben; L. S.] für die Programmierung derartiger Systeme in erhöhtem Maße, weil dabei jede in Schrift oder Wort eingegebene Äußerung des Benutzers, auf die der Computer reagieren soll, von vorneherein vorgesehen werden muß; zu Wörtern, die ein Programmierer nicht bedacht hat, wird das System selbstverständlich schweigen. Der Benutzer hingegen, der mit einem solchen System umgeht, wird sich nach und nach an die Beschränktheit der Ausdrucksformen gewöhnen, die sein Computer erfordert, und entwöhnt sich womöglich der differenzierten Ausdrucksformen, die in lebendiger Sprache selbstverständlich sind.
>
> (Ebd.)

Heute ist klar, dass beide Prognosen Ropohls nicht eingetroffen sind. Dank Techniken wie dem maschinellen Lernen können Programme selbst, auf der Basis der jeweiligen Daten, Lösungen generieren und Wahrscheinlichkeiten für ähnliche Lösungsmuster auf der Basis unbekannter Daten angeben. Auf der Benutzerseite hat sich nicht zwingend die Sprache an aktuelle technische Systeme angepasst, wohl aber das Nutzerverhalten, was aber mindestens im selben Maß der Technik wie den Menschen, die sie entwickeln und vertreiben, zugemessen werden kann. Der Historiker Joachim Radkau warnte daher 2017 in einem Interview mit dem Deutsch-

landfunk, dass ob ehemals unrealistischer Zukunftsszenarien Vorsicht geboten sei, sich nicht „in einer hämischen Besserwisserei der Retrospektive zu ergehen“ (vgl. Maas 2017). Ferner erklärt er im Gespräch über seine historische Analyse (vgl. Radkau 2017):

> Viele Bücher, die zu ihrer Zeit prominent und einflussreich waren, suggerierten ihren Lesern, die von ihren Autoren skizzierte Zukunft sei unausweichlich – oder das, was in der Realität noch Zukunftsmusik war, sei längst Fakt.
> (Maas 2017)

Für die Abschätzung von Technikfolgen folgert Radkau daraus, dass es sinnvoll sei,

> mehrere mögliche Zukunftsszenarien zu entwerfen. Und die sollten nun wieder dazu führen, auch die Gegenwart noch schärfer, noch vielseitiger zu beobachten, ob da nicht auch so ganz unterschiedliche Potenziale drin enthalten sind.
> (Ebd.)

Eine ähnliche wie die von Radkau angesprochene, häufig postulierte, Unausweichlichkeit lässt sich auch bei manchen zeitgenössischen Autor∗innen im Bereich der (Ethik der) künstlichen Intelligenz feststellen. Einige, die über die Folgen von künstlicher Intelligenz und über mögliche Zukunftsentwicklungen nachdenken, warnen zum Beispiel vor der potenziellen Gefährlichkeit der Entwicklung. Gefahren werden unter anderem bei der Entwicklung automatisierter Waffensysteme gesehen (vgl. Life Institute 2021) oder davor, dass Maschinen und Algorithmen Menschen in allen Fähigkeiten übertreffen könnten, was auch, wie oben schon einmal genannt, als *Superintelligenz* bezeichnet wird (vgl. z. B. Bostrom 2014; Kurzweil 2005). Für viele kritische Annahmen werden gute Gründe angeführt, die ernst genommen werden sollten; eine differenzierte Betrachtung der jeweils hinter dem, mitunter auch emotional diskutierten, Begriff *künstliche Intelligenz* stehenden, konkreten technischen Systeme, ist jedoch unerlässlich (vgl. z. B. Hagendorff auf S. 37 in der vorliegenden Arbeit). Aus wissenschaftlicher Sicht (egal, welcher Disziplin), ist es vor allen Dingen wichtig, die Entwicklungen so transparent und treffend wie möglich zu beschreiben, diese dabei nicht zu überhöhen, allerdings die potenziellen Risiken auch nicht kleinzureden. Denn der Blick in die Geschichte lehrt, dass es letztendlich immer passieren kann, dass in der Zukunft Entwicklungen vonstattengehen, die zu einem vorherigen Zeitpunkt mitnichten absehbar waren.

> The idea that machines would never be capable of originating anything had been raised over [...] by Ada Lovelace, who had worked with Charles Babbage to give a comprehensible description of the ‚analytical engine‘ – a planned but unbuilt mechanical computing device that is regarded as a forerunner to the modern digital computers ultimately made possible by Turing's work.
> (Wallach und Allen 2009: 100)

3.5 Technische Systeme in journalistischen Medien und Science-Fiction

Journalistische Artikel zu Themen, wie *Robotik* und *künstliche Intelligenz*, werden häufig mit Bildkonstruktionen auf der Basis eines Zitats des Deckenfreskos Michelangelo Buonarottis in der Sixtinischen Kapelle in Rom versehen, das die Erschaffung Adams aus der christlichen Schöpfungsgeschichte zeigt. Der ausgestreckte Arm Adams wird dann meist durch das Foto eines Roboterarms ersetzt (vgl. u. a. Voss et al. 2018). Zuweilen wird jedoch auch der Arm Gottes durch eine Roboterhand ersetzt[60], wobei fraglich ist, ob diese Abweichungen vom Original jeweils bewusst oder aus Versehen geschahen. Ob es in anderen Teilen der Welt in diesem Zusammenhang auch Bezüge zu Göttern aus anderen Religionen gibt, kann hier nicht beurteilt werden. Es gibt allerdings Untersuchungen, in denen die Darstellungen von Automaten und Robotern in verschiedenen Religionen analysiert werden (vgl. z. B. Mayor 2018).

Künstlerisch verarbeitete, unter vielen anderen, der tschechische Autor Karel Čapek die Ängste und Möglichkeiten rund um moralische Maschinen in seinem Theaterstück von 1921 unter dem Titel *R.U.R. Rossum's Universal Robots*.[61] In Čapeks Theaterstück werden eingangs, vor dem Hintergrund des in den 1920er-Jahren aktuellen wirtschaftlichen Aufschwungs und der Industrialisierung, produktive, zunächst scheinbar wehr- und gefühllose Arbeitsroboter als das gesellschaftlich weitgehend anerkannte Nonplusultra der Technik dargestellt. Der im Stück gegenwärtige Besitzer der Roboterfirma R.U.R., Harry Domin, äußert sich dementsprechend herablassend über einen früheren Besitzer: „Daß der alte Werstand ein erstaunlicher Narr gewesen ist. Ernstlich, Fräulein Glory, aber das behalten Sie für sich. Jener alte Hitzkopf wollte tatsächlich Menschen machen." (Čapek 2017: 9 f.) Diese Äußerung wirkt zunächst irritierend, scheint doch, wie in Abschnitt 3.2 dargelegt, zumindest *ein* wesentliches Ziel der Entwicklung von Automaten, Robotern und Maschinen die nunmehr gänzliche *Ersetzung* von menschlicher Arbeit (oder, dystopischer gesehen, von Menschen selbst) durch technische Systeme zu sein. Dieses Ziel wird auch immer wieder explizit als solches im Kontext technischen Fortschritts genannt. So schreibt beispielsweise Pamela McCorduck:

[60] Vgl. z. B. iStock 2018; Vígh 2016.

[61] Erstdruck 1920, für eine deutsche Übersetzung vgl. Čapek 2017 beziehungsweise das Hörspiel des Bayerischen Rundfunks (RIAS Berlin): Čapek 2005.

> Our history is full of attempts – nutty, eerie, comical, earnest, legendary and real – to make artificial intelligences, to reproduce what is the essential us – bypassing the ordinary means. Back and forth between myth and reality, our imaginations supplying what our workshops couldn't, we have engaged for a long time in this odd form of self-reproduction.
>
> (McCorduck 2004: 3)

Der Ausdruck *reproduction* kann dabei natürlich in zwei Varianten gelesen werden: Einmal kann es um die ersatzlose maschinelle Reproduktion von Menschen gehen und einmal um eine Ergänzung menschlicher Wesen um technische Systeme verschiedener Formen. Ersteres wird zwar nicht als Ziel genannt, muss jedoch bei der zunehmend produktiveren Entwicklung technischer Systeme im Hinblick auf deren kognitive Fähigkeiten, wie beispielsweise Bewusstsein, zumindest als mögliche Konsequenz bedacht werden, selbst wenn eine auch physiologische Vergleichbarkeit aus heutiger Sicht noch in ferner Zukunft zu sein scheint (s. vergangener Abschnitt; vgl. hierzu auch Wendland 2022). Auf die Vergleichbarkeit funktionieller Aspekte komme ich unten zu sprechen.

Eine im Kontext von Science-Fiction besonders aus feministischer Perspektive interessante Analyse bietet die historische Untersuchung technischer Ersatzmöglichkeiten für Frauen:

> Bei all diesen künstlichen Frauensurrogaten ist die Sehnsucht des Mannes zu spüren, die weibliche Sexualität zu beherrschen, und das wiederum hängt mit seiner tiefverwurzelten Angst vor der andersartigen Natur der Frau zusammen, insbesondere ihrer Fähigkeit, Leben hervorzubringen. Sein Defizit sucht er nicht nur mit Sublimationsleistungen in der Kunst zu bewältigen, sondern auch auf eine eher direkte Weise, indem er die Rolle der Frau bei der Geburt zu verändern trachtet. Sie ganz auszuschalten [...] war aus verschiedenen Gründen faktisch nicht durchzuführen; aber daß ein Embryo außerhalb des weiblichen Körpers, in der Retorte entstehen kann, hat uns die Reproduktionsmedizin seit langem bewiesen.
>
> (Drux 1994b: 24)

Harry Domin in Karel Čapeks Theaterstück hält jedoch „Blinddarm, Mandeln, Nabel“ für „lauter Überflüssigkeiten“ und konkretisiert: „Die sind nicht überflüssig, ich weiß. Aber wenn man die Menschen künstlich erzeugen will, dann sind sie – hm – keineswegs notwendig“ (Čapek 2017: 10).

> Der junge Werstand, Fräulein, das war das neue Zeitalter. Das neue Zeitalter der Produktion nach dem Zeitalter der Erkenntnis. Nachdem er die menschliche Anatomie beguckt hatte, sah er gleich, daß dies allzu kompliziert ist und daß ein guter Ingenieur es einfacher machen müßte. Er begann also die Anatomie umzuarbeiten und erprobte, was sich auslassen oder vereinfachen ließe. [...] Eine Arbeitsmaschine muß nicht Violine spielen, muß nicht Freude fühlen, muß nicht einen Haufen andrer

> Dinge tun. Soll es schließlich gar nicht. [...] Die Erzeugung soll möglichst einfach und das Erzeugnis möglichst billig sein.
> (Čapek 2017: 11 f.)

Dementsprechend ist der *praktisch beste* Arbeiter gemäß Harry Domin nicht etwa der *ehrlichste* oder *ergebenste*, wie Fräulein Glory als Personifizierung von Moral und Naivität annimmt, sondern „der billigste“ (vgl. ebd.: 12). Dieser personell vermittelte Kapitalismuskritik macht deutlich, dass es Harry Domin nicht, wie erwartet werden könnte, um eine perfekte technische Nachzeichnung von Menschen geht (wie sie beispielsweise der japanische Robotiker Hiroshi Ishiguro zumindest bezüglich äußerlicher Ähnlichkeiten verfolgt), sondern um das wirtschaftliche Ziel der Effizienz. Das dabei in Kauf genommene Risiko – angesichts der Entstehungszeit des Theaterstücks besonders hervorhebenswert – besteht darin, dass eine maschinelle Überhöhung der Technik über Menschen (mit allen Konsequenzen) nicht ausgeschlossen werden kann.

Dieses Narrativ findet sich (auch aus heutiger Sicht) als logischer nächster Schritt der Technikentwicklung mittlerweile in zahlreichen Science-Fiction-Erzählungen wieder, wie zum Beispiel in Ian McEwans *Maschinen wie ich* (McEwan 2019).[62] Die Überlegung, dass Maschinen nicht nur, wie dem Maschinenbild im Zeitalter der Industrialisierung entsprechend, manche schweren, mühsamen Arbeiten für Menschen erledigen und damit, gemäß einer utopischen Sicht – die aus wirtschaftlicher Perspektive jedoch nicht direkt aufging, wie zum Beispiel die Weberaufstände im 19. Jahrhundert zeigen (vgl. u. a. Thimm 2019) – das Leben von Menschen vereinfachen, sondern vielmehr auch bezüglich sämtlicher kognitiver und moralischer Fähigkeiten Menschen übertreffen und diese damit überflüssig machen könnten, wurde bereits in Abschnitt 2.3.5 vorgestellt. In Bezug auf das Kernthema der vorliegenden Arbeit, genuin ethische Techniksysteme, wird sich somit die Frage stellen, ob die logische Fortführung moralischer Maschinen, die im folgenden Kapitel zuerst vorgestellt werden, dem Zweck der Verbesserung menschlichen Lebens (dies muss im Detail natürlich definiert werden) dienen kann und soll oder ob nicht auch die (aus menschlicher Perspektive) Gefahr moralisch und ethisch gesehen weitaus überlegener Maschinen und Algorithmen besteht, die nicht nur gemäß menschlich-ethischen Maßstäben reflektieren, sondern ganz eigene entwickeln, denen Menschen sich im Zweifelsfall zu unterwerfen haben werden[63]

[62] S. hierzu auch Abschnitt 5.2.4. Xanke und Bärenz (2012) liefern einen guten Überblick über einige Science-Fiction-Erzählungen mit dem Fokus auf künstlich intelligente Systeme, ebenso wie (im Hinblick auf bestimmte Gesichtspunkte) Bruckenberger et al. (2013), Gibson (2020), Meinecke und Voss (2018) sowie Zeng (2015). Bei Blackford (2017) geht es explizit um moralische Fragen in Science-Fiction-Geschichten.

[63] S. hierzu auch die verschiedenen Definitionen von Maschinenethik auf S. 30.

Bevor ich im folgenden Kapitel mit der Analyse moralischer Maschinen beginne, werde ich noch kurz auf einen häufig zitierten Klassiker eingehen, Isaac Asimov.

> Isaac Asimov, more than fifty years ago, foresaw the need for ethical rules to guide the behavior of robots. His Three Laws of Robotics are what people think of first when they think of machine morality.
> (Wallach und Allen 2009: 3)

Die *Robotergesetze* oder auch „Grundregeln der Robotik“ (Asimov 2016: 7), die Isaac Asimov 1942 erstmals einer seiner Kurzgeschichten, *Runaround*, voranstellte (Asimov 1942), lauten wie folgt:

1. „Das nullte Gesetz: Ein Roboter darf der Menschheit keinen Schaden zufügen oder durch Untätigkeit zulassen, dass der Menschheit Schaden zugefügt wird.“

2. „Das erste Gesetz: Ein Roboter darf einem menschlichen Wesen keinen Schaden zufügen oder durch Untätigkeit zulassen, dass einem menschlichen Wesen Schaden zugefügt wird, es sei denn, dies würde das nullte Gesetz der Robotik verletzen.“

3. „Das zweite Gesetz: Ein Roboter muss dem ihm von einem menschlichen Wesen gegebenen Befehl gehorchen, es sei denn, dies würde das nullte oder das erste Gesetz der Robotik verletzen.“

4. „Das dritte Gesetz: Ein Roboter muss seine Existenz beschützen, es sei denn, dies würde das nullte, das erste oder das zweite Gesetz der Robotik verletzen.“ (Asimov 2016: 7)[64]

Ziel der Gesetze in Asimovs Geschichten ist, das Zusammenleben von Menschen und (oft androiden und im Sinn *starker KI* autonom[65] entscheidenden) Robotern zu regeln, wobei die Gesetze in den Geschichten häufig zu Konflikten beziehungsweise unerwartetem Verhalten führen – zwischen Menschen und Robotern oder innerhalb der Roboter selbst.

In der Geschichte *Runaround* geht es um einen Roboter, *Speedy*, der während einer Merkur-Mission Selen sammeln soll, um damit „Fotozellbänke“ auf der „Schutzwand“ der Basisstation reparieren zu können (vgl. ebd.: 47). Nachdem Speedy einige Zeit lang vermisst worden ist, finden ihn die beiden Raumfahrer (durch das Sichtfenster der sicheren Station heraus) in einem ungewöhnlichen Zustand.

[64] Ursprünglich waren es drei Gesetze (eins bis drei), das nullte kam später hinzu.
[65] Wie der Begriff Autonomie verstanden werden kann, wird in Abschnitt 5.2.2 diskutiert.

> Nun waren sie nah genug, um zu erkennen, dass Speedys Gangart eigenartig schwankend war. Es war, als schaukelte er auf eigenartige Weise hin und her. Als dann Powell wieder winkte und die größtmögliche Sendeenergie in seinem kompakten Funksender, der sich in seinem Helm befand, einschaltete – zur Vorbereitung für einen weiteren Ruf –, schaute Speedy auf und sah sie.
>
> Mit einem Ruck blieb Speedy stehen und verharrte einen Augenblick in dieser Stellung. Er war nicht völlig regungslos, sondern schwankte ein klein wenig, als würde er von einem leichten Winde bewegt.
>
> Powell brüllte: ‚Also mach keine Geschichten, Speedy! Komm hierher, mein Junge!‘
>
> Worauf zum ersten Male Speedys Roboterstimme in Powells Kopfhörern ertönte.
>
> Sie sagte: ‚Wir wollen Spielchen spielen. Du fängst mich, und ich fang dich. Keine Liebe kann unser Messer in zwei Teile schneiden. Denn ich bin klein Butterblümchen – das liebe, kleine Butterblümchen. Juchhee.‘ Und er machte auf dem Absatz kehrt und rannte in die Richtung davon, aus der er gekommen war, und zwar mit einer Geschwindigkeit und Kraft, dass der von der Hitze festgebackene Boden nur so nach allen Seiten spritzte.
>
> (Asimov 2016: 58 f.)

Einer der beiden Raumfahrer kommt schließlich auf die Idee, dass die Selen-Quelle auf irgendeine Weise eine Gefahr für den Roboter darstellen müsse, sodass ein Konflikt in der Maschine zwischen dem zweiten und dem dritten Gesetz entstanden ist.

> ‚Wir stecken bereits mitten in der Erklärung: Der Konflikt zwischen den verschiedenen Regeln wird durch die verschiedenen Potenziale im positronischen Gehirn ausgeglichen. Sagen wir mal, ein Robot läuft bewusst in eine Gefahr hinein. Das automatische Potenzial, das durch Regel drei aufgebaut wird, weist ihn zurück. Nimm aber an, du befiehlst ihm, in eine solche Gefahr hineinzulaufen. In diesem Falle baut Regel zwei ein Gegenpotenzial auf, das größer ist als das vorhergehende, und der Robot folgt deinem Befehl auf die Gefahr hin, damit sein Dasein zu riskieren. [. . .]
>
> Nehmen wir mal Speedys Fall an! Speedy ist eines der neuesten Modelle. Er ist äußerst spezialisiert und fast so teuer wie ein Schlachtschiff. Er gehört nicht zu den Dingen, die man leichten Herzens zerstört. [. . .]
>
> Daher ist bei ihm Regel drei verstärkt worden [. . .]. Daher ist seine Abneigung gegen Gefahr besonders groß. Gleichzeitig aber hast du ihm, als du ihn ausschicktest, um das Selen zu bringen, diesen Befehl sozusagen gleichgültig gegeben, ohne seine Dringlichkeit besonders zu betonen, dadurch ist das durch Regel zwei aufgebaute Potenzial ziemlich schwach geworden. [. . .]
>
> Irgendwo in der Nähe des Selenvorkommens droht eine Gefahr. Diese wird umso größer, je mehr er sich dem Selen nähert, und an irgendeinem Punkt kommen das von Anbeginn an ungewöhnlich niedrige Potenzial der Regel drei und das von Anbeginn an ungewöhnlich niedrige Potenzial der Regel zwei in eine Art Gleichgewicht.‘
>
> (Ebd.: 61 f.)

Schließlich wagt einer der beiden Menschen, sich nach draußen, in eigentlich viel zu hohe Temperaturen, zu begeben, um so das erste Gesetz bei Speedy zu aktivieren, sodass der oszillierende Zustand aufgehoben werden kann, da das nullte Gesetz, der absolute Schutz eines menschlichen Wesens, den Konflikt zwischen dem zweiten und dritten Gesetz aufhebt.
Wie man anhand dieser Geschichte zeigen kann, handelt es sich bei Asimovs Regeln um fiktive Gesetze, die das Vorgehen oder die Auswirkungen eines Roboters kontrollieren sollen. Zwar handelt es sich um eine durch Fiktion dargestellte Möglichkeit der Regulierung, die teilweise bezüglich ihres Status auch den Charakter juristischer Gesetze hat, jedoch wurden und werden sie auch in Fachkreisen im Kontext der Maschinenethik immer wieder im Hinblick auf ihr Potenzial für reale Technikentwicklungen (zum Beispiel teil-automatisierten Fahrzeugen) aufgegriffen. Thimm und Bächle (2019: 84) schreiben den fiktiven Gesetzen auch eine „erstaunliche Aktualität“ zu. Praktisch gesehen, sind die Regeln laut Susan Anderson (2008) allerdings zu inkonkret, um tatsächlich Anwendung zu finden.[66]

Die Erörterung genuin ethischer Maschinen, die, ähnlich wie in fiktiven Szenarien, menschenähnlich und teilweise besser als diese moralisch denken und handeln können, setzt eine Darstellung moralischer Maschinen voraus, die nun folgt.

[66]Die Frage, inwiefern solche moralischen Regeln als ethische Grundlage für moralische Techniksysteme taugen, vertiefe ich in Abschnitt 5.4.2.

4 Technische Systeme als moralische Akteure

> Und indem – so können wie [sic!] diese Überlegungen im Hinblick auf Maschinen als ‚moralische Akteure' zusammenfassen – wir einer Maschine Moralfähigkeit unterstellen, sie zur ‚moral machine' [...] erklären, tun wir nichts anderes als ihr regelgestütztes Handeln und Kommunizieren zu *unterstellen* – Regeln, die als Code die Wertvorstellungen individueller humaner Akteure *realisieren*.
> (Rath 2019: 228 f. Herv. i. Orig.)

Interessant für die folgenden Überlegungen ist vor allen Dingen der Anlass für die Etablierung von Roboterregeln – ähnlich denen, wie Isaac Asimov sie für seine Geschichten erfunden hat: die Tatsache, dass Maschinen, Roboter oder Computerprogramme sich auf verschiedene Weisen verhalten, die von Menschen als moralisch oder unmoralisch bewertet werden können. Es ist dann entweder die Rede von *moralischen Maschinen* (*moral machines* oder von *moralischen Akteuren* (*moral agents*). Bei der Verwendung des Ausdrucks *moralische Maschine* scheint es eher um die Betonung dessen zu gehen, was ein technisches System (gemäß einer bestimmten Definition) *moralisch* macht, wobei beim Ausdruck *moralischer Akteur* die (zu beweisende) Handlungsfähigkeit im Fokus zu stehen scheint (s. Abb. 2). Demnach wird im Folgenden zu klären sein, was das Moralische bei technischen Systemen sein kann, ob diese überhaupt *moralisch* sein können und ob sie (moralisch) *handeln* können. Der Begriff der *Handlung* wird in der Philosophie, wie inbsbesondere in Abschnitt 4.1 deutlich werden wird, sehr vielfältig diskutiert, selbst *Verhalten* ist ein Ausdruck, der normalerweise vorrangig auf Menschen (oder allenfalls Tiere) angewandt wird. Umgangssprachlich und medial werden die Ausdrücke, die zunächst auf menschliche Vermögen und Eigenschaften referieren, häufig recht unbeschwert auf technische Systeme transferiert, was zur Verwirrung über deren tatsächliche Einflussmöglichkeiten auf menschliche Gefüge beitragen kann. Neutraler kann daher in Anlehnung an den Ausdruck *Akteur* von *agierenden* technischen Systemen die Rede sein. Die folgende Abbildung stellt die zentralen Fragen in diesem Kapitel auf einen Blick dar.

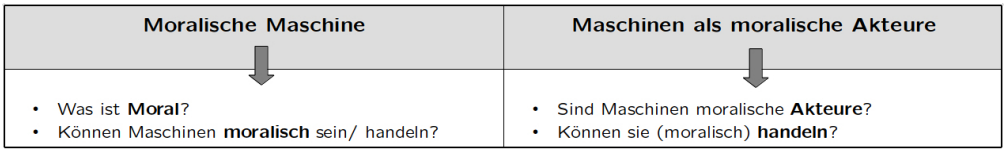

Abbildung 2: Moralische Maschinen und moralische Akteure.

Wenn es um die Moralfähigkeit technischer Systeme geht, sind meist verschiedene Ausdrucksweisen im Spiel. Die Rede ist davon, dass Maschinen moralisch *sein*

oder *handeln* können – oder eben nicht. Und es wird von *moral decision-making* gesprochen, von *algorithmischen Entscheidungssystemen* oder davon, dass Maschinen (moralische) *Entscheidungen treffen* (können) (vgl. u. a. Braun 2019; Bigman und Gray 2018; Leben 2018). Mit Sicherheit gibt es noch andere Anthropomorphismen in Bezug auf technische Systeme; jene sind jedoch zentral für die Klärung von Maschinenmoral, darum möchte ich mich im Folgenden auf sie konzentrieren.

4.1 Was ist Moral?

Was bedeutet es, moralisch zu sein? Wenn ich dem Programm *Siri* der Firma *Apple Inc.* diese Frage stelle, liefert mir das System drei Verweise auf Webseiten als Antwort: 1. Die Webseite der *Bundeszentrale für politische Bildung* Deutschlands, *HanisauLand* (vgl. Toyka-Seid 2021), die zum Ziel hat, politische Inhalte kindgerecht aufzubereiten und zu erklären. 2. Die „Frage der Woche: Was ist Moral?“ der Reihe *WASISTWAS* des Tessloff Verlags (vgl. Tessloff Verlag 2012). 3. Die deutsche Wikipedia-Seite zu dem entsprechenden Eintrag (Wikipedia 2021).[67] Obwohl eindeutig ist, dass die Inhalte der angegebenen Seiten, selbst wenn sie nicht falsch sind, für eine wissenschaftliche Arbeit des vorliegenden Formats nicht ausreichend sind, möchte ich kurz darauf eingehen, um dann anschließend mit der Darstellung philosophischer Konzepte daran anzuknüpfen. Im Lexikoneintrag zu Moral von *HanisauLand* heißt es (Herv. i. Orig.):

> **Werte und Regeln**
> Dieses Wort kommt vom lateinischen Begriff ‚moralis‘ und heißt übersetzt ‚die Sitten betreffend‘. Als Moral werden die Werte und Regeln bezeichnet, die in einer Gesellschaft allgemein anerkannt sind. Wenn man sagt, jemand hat ‚moralisch‘ gehandelt, ist damit gemeint, dass er sich so verhalten hat, wie es die Menschen richtig und gut finden.
>
> **Wie soll man sich verhalten?**
> Im Alltag stellen sich die Menschen immer wieder die Frage, ob sie sich moralisch verhalten: Bin ich fair gegenüber meinen Mitmenschen? Beleidige und verletze ich niemanden? Und sie fragen, wie sie moralisch handeln können: Wie kann ich eingreifen, wenn ich merke, dass Schwächere und Wehrlose gemobbt werden? Was kann ich tun, damit es den Menschen in den armen Ländern besser geht? Nicht immer lässt sich eine einfache Antwort auf solche Fragen finden.
>
> **Gebote**
> Es gibt Gebote, die moralisches Handeln vorschreiben, zum Beispiel: ‚Du sollst nicht töten‘ oder ‚Du sollst nicht stehlen!‘ Ohne diese Gebote oder Gesetze ist ein

[67] Die Abfrage erfolgte auf einem Gerät, bei dem *Siri* normalerweise nicht aktiviert ist und nur für diesen Zweck einmalig eingeschaltet wurde. Andere Geräte mögen andere Antworten ergeben.

> Zusammenleben in einer Gesellschaft nicht gut möglich. Für Juden und Christen benennen die ‚Zehn Gebote' die wichtigsten Regeln, die Menschen beachten sollen.
>
> **Weitere Bedeutungen**
> Gelegentlich benutzt man den Begriff ‚Moral' auch, um eine gute Einstellung zu beschreiben. Wenn ein Sportler sehr viel Einsatz zeigt, spricht man von ‚Kampfmoral'; bei Soldaten hört man manchmal von der ‚guten Moral der Truppe'. [...]

Bei *WASISTWAS* war am 01.11.2022 Folgendes zu lesen:[68]

> Moral geht auf das lateinische Wort ‚moralis' zurück, das soviel wie ‚die Sitte betreffend' heißt. Eine Sitte ist eine menschliche Umgangs- beziehungsweise Verhaltensform, auch Gepflogenheit genannt.
>
> Zum einen beschreibt Moral hauptsächlich Handlungen, die ein Mensch oder eine Gesellschaft von anderen Mitmenschen erwartet. Moral soll also dafür sorgen, dass Menschen ein bestimmtes Verhalten an den Tag legen.
>
> So gilt beispielsweise die linke Hand im indischen Kulturkreis als unrein. Begrüßt man dort eine Person mit der linken Hand, ist das ein Verstoß gegen die kulturellen Sitten dieser Gesellschaft und somit streng genommen ein unmoralisches Verhalten.
>
> Außerdem kann Moral auch als das Gegenteil von bösen Handlungen angesehen werden. In diesem Fall spricht man dann von ‚moralisch gut' und beschäftigt sich vor allem damit, was der Mensch als richtiges, gutes oder gerechtes Handeln ansieht.
>
> Wenn man zum Beispiel hilft, einen Konflikt zwischen Schulkameraden friedlich, also ohne Gewalt, zu schlichten, dann ist das eine moralisch gute Handlung. [...]

Wikipedia liefert schließlich folgende Definition:

> Als Moral werden zumeist die faktischen Handlungsmuster, -konventionen, -regeln oder -prinzipien bestimmter Individuen, Gruppen oder Kulturen und somit die Gesamtheit der gegenwärtig geltenden Werte, Normen und Tugenden bezeichnet. Der Verstoß gegen Moralvorstellungen wird als Unmoral bezeichnet. Amoral benennt das Fehlen beziehungsweise die bewusste Zurückweisung von Moralvorstellungen, bis hin zur Abwesenheit von moralischer Empfindung. [...]

Allen diesen Definitionen liegt die Idee zugrunde, dass sich Moral in Form bestimmter Prinzipien, Normen oder Gesetze ausdrückt (die Ausdrücke *Sitte* und *Gepflogenheit* sind veraltet und werden außerhalb von Definitionen von Moral

[68] Am 22.01.2022 war diese Internetseite nicht mehr verfügbar und auch kein vergleichbarer Eintrag zum Stichwort *Moral* oder *Ethik*.

kaum noch verwendet), die wiederum durch Einstellungen, Entscheidungen oder Handlungen zutage treten können. In der Moralphilosophie wird *Moral* nicht immer als per se *gut* definiert. Zuweilen wird differenziert zwischen moralisch richtigen und moralisch falschen Überzeugungen und Handlungen. Dies widerspricht allerdings der Alltagsauffassung von Moral – man denke hier auch an die *Moral* einer Geschichte –, diese sei stets die gute, bevorzugte Wahl. Gemäß der eingangs zitierten Definition von Gerhard Ernst, kann Moral mit der Frage *Was soll ich tun?* definiert werden, wobei es hierauf (individuell oder für eine größere Gesellschaft) *richtige* und *falsche* Antworten geben kann. Im engeren Verständnis von Moral können moralisch falsche Handlungen an Verboten, moralisch richtige an Geboten gemessen werden. Übernimmt man diese zunächst bezüglich einer spezifischen Wertung *neutrale* Definition von Moral als das Handeln im Allgemeinen betreffend, passt diese auch zu der obigen Gleichsetzung von Moral und Algorithmen, welche beide regelgeleitete Abfolgen oder Handlungen implizieren, wobei eine ethische Bewertung hier noch nicht inbegriffen ist. Für die Bewertung von Handlungen spielen zudem deren Voraussetzungen oder Entstehungsbedingungen eine Rolle, die gemäß verschiedener Handlungstheorien unter anderem bestimmte Einstellungen, Überzeugungen und Entscheidungen beinhalten. Auch andere psychische Faktoren wie*Intentionalität*, *Motivation* oder *Wille* werden als Handlungen bedingende Faktoren genannt, wobei ihre Rolle für Handlungen in verschiedenen Theorien unterschiedlich bewertet wird.

Für die Bestimmung, was moralisch richtig, also geboten und was moralisch falsch ist beziehungsweise anhand welcher Kriterien dies bemessen wird, müssen verschiedene Faktoren berücksichtigt werden: der Kreis handelnder oder entscheidender Subjekte, der Kreis der betroffenen Objekte. Je nach Kontext können die Bewertungen sehr unterschiedlich ausfallen. Für Hans (s. Abschnitt 2.3.2), im obigen Beispiel, mag es geboten sein, seinen Nachtisch frei nach Geschmack zu wählen, wohingegen dies für eine Person mit bestimmten Vorerkrankungen nicht unbedingt der moralisch richtige Weg sein muss.

Solche Überlegungen vor dem Hintergrund eines jeweiligen Kontextes müssen auch für technische Systeme angestellt werden. Ein klassisches Problem, das im Rahmen künstlicher Intelligenz immer wieder ins Feld geführt wird, sind (teil-) automatisierte Fahrzeuge (vgl. u. a. Misselhorn 2019: 184 ff.; Schippl und Hillerbrand 2021; Wölm 2019). Sollen diese in Einzelfällen allein entscheiden dürfen, welche Aktionen sie in Konfliktsituationen in die Wege leiten? Und wenn ja, nach welchen Kriterien? Dürfen diese technischen Systeme eingesetzt werden, wenn sie im Zweifelsfall gegen Menschenrechte, nämlich das Recht auf Leben, zu verstoßen drohen? Oder wird die Gefahr von Verkehrsunfällen durch (teil-)automatisierte Fahrzeuge sogar verringert, da diese nicht menschlichen Fehleranfälligkeiten wie Müdigkeit, Unaufmerksamkeit, eingeschränkte Wahrnehmungsfähigkeit und Alko-

holmissbrauch unterliegen? Ist es vielmehr sogar moralisch gefordert, die technische Entwicklung voranzutreiben, mit der (angesichts der im vorausgegangenen Kapitel dargestellten, bisherigen Technikentwicklungen) Hoffnung, technische Systeme mögen (zunehmend) *bessere* Handlungen zutage fördern als Menschen? Die große Schwierigkeit bei der Implementierung moralischer Aktionsweisen in technische Systeme ist, dass aus technischer Sicht Antworten auf solche Fragen vorliegen müssten – allgemein auf die Frage, was moralisch richtig ist, aber auch konkret, was moralisch richtiges Handeln in bestimmten Situationen bedeutet – die innerhalb der Moralphilosophie und Ethik oftmals und insbesondere für komplexe Situationen nicht eindeutig und abschließend gegeben werden können (s. hierzu auch S. 37).

Für eine mögliche technische Umsetzung ist außerdem die Frage wichtig, welche Voraussetzungen für moralisch richtige Handlungen und Entscheidungen angenommen werden können – zunächst bei Menschen und dann in technischen Systemen.[69] Diese Frage hängt eng mit der philosophischen Definition von *Handlung* zusammen und der daran anschließenden Frage, ob als Handlung allein die Folgen einer Tätigkeit betrachtet werden oder auch die Intentionen und Einstellungen einer Person beziehungsweise eines Systems. Danach bestimmt sich wiederum die Umsetzbarkeit in technischen Systemen; im Fall einer Definition von moralischen Akteuren allein über die Handlungs*folgen*, muss man sich weniger Gedanken über die Umsetzung psychischer Phänomene in technischen Systemen machen, was die technische Implementierung deutlich erleichtern kann. Geht man jedoch davon aus, dass Phänomene wie *Intentionen*, *Bewusstsein* oder auch *Emotionen* eine notwendige Bedingung für moralisches Handeln darstellen, gestaltet sich die technische Umsetzung schwieriger. Die Fragen aus Abbildung 2 können also wie folgt differenziert werden:

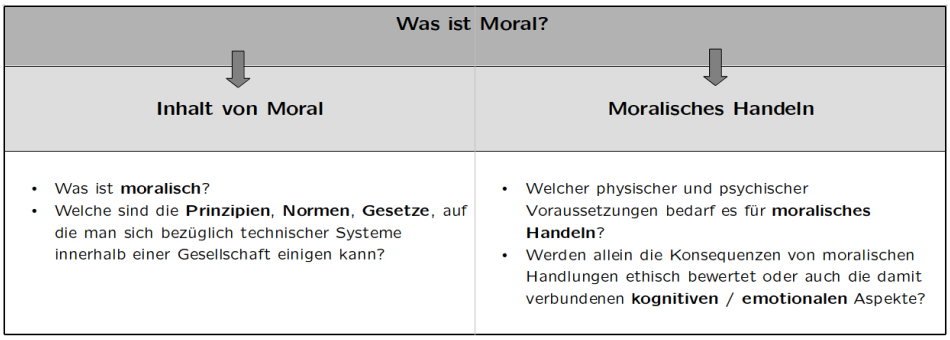

Abbildung 3: Inhalt von Moral und die Voraussetzungen moralischen Handelns.

[69] In Abschnitt 5.2 diskutiere ich die jeweils analysierten Begriffe im Hinblick auf *ethische Reflexion* allgemein und füge dann jeweils Überlegungen zur technischen Umsetzung hinzu.

4.2 Moralisch richtig handelnde Maschinen – technischer Pragmatismus

Ich widme mich zunächst den Fragen in der rechten Spalte. Ob Maschinen handeln können und, wenn ja, unter welchen Umständen man von *maschinellen Handlungen* sprechen kann, wurde schon an vielen Stellen diskutiert (vgl. z. B. Collins und Kusch 1999; Parthemore und Blay 2013). Zahlreiche weitere Publikationen lassen sich hierzu finden, und häufig werden die Diskussionen im Zusammenhang von potenzieller Verantwortungsfähigkeit der handelnden Akteure diskutiert (vgl. u. a. Loh 2019b; Rath 2019). In der philosophischen Handlungstheorie werden dabei häufig nicht nur die Folgen von Handlungen in den Blick genommen, sondern auch psychische Phänomene und kognitive Fähigkeiten, wie zum Beispiel Intentionalität oder Absichtlichkeit (im Englischen *intentionality*, s. Abschnitt 5.2.7) sowie die Frage nach den primären Gründen für Handlungen, die auch als *Motive* bezeichnet werden. In Bezug auf die ethische Bewertung von Handlungen spielt dabei auch eine wesentliche Rolle, ob die jeweiligen *Gründe* für Handlungen als (moralisch) *gerechtfertigt* eingestuft und somit als *rational* bezeichnet werden können oder nicht (vgl. Alvarez 2018). Darüber hinaus werden auch Autonomie und Willensfreiheit als mögliche notwendige Voraussetzungen für moralisches Handeln diskutiert (vgl. u. a. Quante 2020: 52 ff.; 78 ff.; 113 ff.).

Bei technischen Systemen sind vor allem die Folgen relevant. Die zugrundeliegenden *Gründe* und angrenzende Voraussetzungen können natürlich auf einer systematischen Ebene im Blick behalten werden. Solange ein technisches System auf eine gewünschte Weise funktioniert, ist es jedoch zumindest für die Funktionsweise nicht relevant, ob man dem System auch Intentionen oder Bewusstsein zusprechen kann oder muss. Dies kann als *technischer Pragmatismus* bezeichnet werden. Ob man bei technischen Systemen dabei in gleichem Maß von *Handlungen* sprechen kann wie bei Menschen, sei zunächst dahin gestellt. Für die ethische Bewertung technischer Systeme spielt dies eine untergeordnete Rolle, wie aktuelle Diskussionen um mögliche technische Entwicklungen zeigen. Bei der Entwicklung (teil-)automatisierter Fahrzeuge geht es beispielsweise allein um die Konsequenzen, die der Einsatz der Systeme mit sich bringt – um mögliche Gefahren und wie Unfälle vermieden werden können. Die *Intentionen* oder *Motivationen* (in welchem Maß die Rede hiervon bei Techniksystemen sinnvoll ist, diskutiere ich unten) eines Programms sind dabei nur dann relevant, wenn sie einen direkten Einfluss auf die Folgen haben. Bei juristischen Handlungsbewertungen spielen Motive, Intentionen und Absichtlichkeit von Menschen durchaus eine Rolle. Dies kann für technische Systeme jedoch – zumindest Stand heute – ausgeschlossen werden, da stets Menschen für die Konsequenzen haften und auch ethisch verantwortlich sind, die der Einsatz technischer Systeme mit sich bringt.

Der Fokus auf die Konsequenzen technischer Aktionen bringt zudem den Vorteil mit sich, dass, zumindest für moralische Maschinen, keine Lösungen für mögliche technische Umsetzungen kognitiver und emotionaler Fähigkeiten gesucht werden müssen, was sich mitunter als schwierig gestalten kann. Ob und in welchem Maß dies auch für genuin ethische Techniksysteme angenommen werden kann, wird zu untersuchen sein. Folgen technischer Systeme können verschiedene Formen annehmen, wie an Beispielen unten noch deutlich werden wird.[70]

Matthias Rath (2019: 226) verweist bezüglich der ethischen Bewertung von Handlungen darauf, dass diese auch bei Menschen auf einer „Interpretation, die aus dem beobachtbaren Verhalten und den darauf zurückführbaren Folgen dieses Verhaltens (oder im allgemeinen Sinn ‚Tuns')" gründe. Im Zusammenhang seiner Untersuchung der potenziellen Verantwortungs- und Akteursfähigkeit moralischer Maschinen kommt Rath zu dem Schluss, dass sich moralisch agierende Maschinen beziehungsweise vermeintlich „neue[...] Künstliche[...] Intelligenz" als „sozialepistemologischer und ethischer Normalfall" (ebd.) erweisen beziehungsweise dass die Frage nach der Verantwortung nicht-menschlicher Akteure „keineswegs neu" (ebd.: 225) sei. Die Moralfähigkeit von Maschinen ist nach Rath ferner nicht in den Maschinen oder technischen Systemen selbst verortet, sondern ergibt sich aus der menschlichen Zuschreibung dessen, was im Wortsinn, offensichtlich ist:

> Und indem – so können wie [sic!] diese Überlegungen im Hinblick auf Maschinen als ‚moralische Akteure' zusammenfassen – wir einer Maschine Moralfähigkeit unterstellen, sie zur ‚moral machine' [...] erklären, tun wir nichts anderes als ihr regelgestütztes Handeln und Kommunizieren zu *unterstellen* – Regeln, die als Code die Wertvorstellungen individueller humaner Akteure *realisieren*.
> (Ebd.: 228 f. Herv. i. Orig.)

John Danaher (2020) postuliert eine „duty of ‚procreative beneficience' towards robots" (ebd.: 2023), sofern Roboter „*roughly performatively equivalent* to other entities that have significant moral status" seien (ebd.: 2023, Herv. i. Orig.). Ein *significant moral status* bedeutet für Danaher (ebd.: 2026): „[...] we will not be allowed to mistreat or harm the entity without some overriding moral justification". Daraus folgt für ihn:

> If a robot is roughly performatively equivalent to another entity whom, it is widely agreed, has significant moral status, then it is right and proper to afford the robot that same status.
> (Ebd.: 2025)

[70] Gemäß der Sprechakttheorie zählen mitunter auch sprachliche Äußerungen als Handlungen (vgl. Austin 1962; Searle 1969). Allgemein zu Kommunikation und Handlungen vgl. auch Habermas 2020.

Unter *roughly performative equivalence* versteht er dabei Folgendes:

> This means that if a robot consistently behaves like another entity to whom we afford moral status, then it should be granted the same moral status. So if a robot consistently behaves as if it is in pain, and if the capacity to feel pain is a ground of moral status, then a robot should be granted the same moral status as any other entity to whom we ascribe moral status on the grounds that they can feel pain. This is what it means to say that performative equivalency provides sufficient ground for equal moral status.
>
> (Danaher 2020: 2026)

Diese Forderung mag zwar logisch gültig sein, ob das Argument aber schlüssig ist, kann bezweifelt werden. Konsequenterweise würde das zum Beispiel bedeuten, dass Maschinen, die beispielsweise Trauer simulieren, getröstet werden müssten, wenn man davon ausgeht, dass dies die adäquate moralische Handlung in solch einer Situation ist. Nun gab es zwar in den 1990er-Jahren technische Geräte, die Trost und menschliches Kümmern direkt und lautstark eingefordert haben: Tamagotchis (vgl. Kerschreiter 2017). Doch aus der Tatsache, dass viele Menschen sich durch dieses und andere ähnliche technische Gadgets zu sozialem Handeln aufgefordert fühlten, kann kein ethisch-moralisches *Sollen* abgeleitet werden.

Die behavioristische Idee, dass äußerlich erkennbares Verhalten auf bestimmte Zustände schließen lässt (wenngleich Danaher sich hierzu nicht eindeutig positioniert) und daraus wiederum bestimmte Folgen für den Umgang mit technischen Systemen (oder bspw. Tieren) resultieren, mag zwar auf den ersten Blick verlockend erscheinen, auf den zweiten wirft sie angesichts der aktuellen technischen Möglichkeit viele Fragen auf. Die digitale Welt ist voller Simulierungen, nicht nur *deepfakes* (s. o.), sondern auch gefälschte Identitäten usw. Wie sollte man das obige Prinzip guten Gewissens hier anwenden können? In einigen Fällen ist offensichtlich, dass die Programme nicht den moralischen Status verdienen, den ihnen John Danaher zuzuschreiben versucht, in anderen können Täuschungen im Spiel sein. Welche Möglichkeiten sich neben dem behavioristische Ansatz sich für Zuschreibungen von Fähigkeiten an Techniksysteme und daraus resultierenden Folgen eignen können, werde ich in Abschnitt 5.2.7) noch einmal aufgreifen. Für den Moment kann festgehalten werden, dass technische Systeme dann als moralische Akteure bezeichnet werden können, wenn sie bestimmte, zuvor festgelegte ethische Prinzipien für Menschen sichtbar umsetzen können. Damit komme ich zur linken Spalte in Abbildung 3: dem Inhalt von Moral.

4.3 Moral oder die Frage nach dem guten Leben

In Texten zur Maschinen- und Roboterethik wird häufig zuvorderst geklärt, was unter Moral allgemein, also auch in Bezug auf Menschen verstanden wird. Der

Fokus in diesem Abschnitt liegt auf dem Kern von Moral und ihrer ursprünglichen Bedeutung. *Mos* (Plural *mores*), aus dem Lateinischen, bedeutet, wie bereits bei den alltagsgebräuchlichen Definitionen oben angedeutet wurde, je nach Übersetzung Sitte, Gebrauch oder Charakter (vgl. z. B. Misselhorn 2019: 45). Der bereits an verschiedenen Stellen angesprochenen Neutralität von Moral zum Trotz wird der Begriff häufig als *thick concept* (oder, im Deutschen weniger elegant, *dichter Begriff*, manchmal auch *dicker Begriff*) verstanden, also als Begriff, der per se normativ ist (vgl. Kirchin 2013). Wie oben gezeigt, ist eine Handlung oder Einstellung genau dann moralisch richtig, wenn sie den Regeln oder Prinzipien entspricht, auf die sich eine Gemeinschaft oder Gruppe von Menschen geeinigt hat. Der Wunsch, Vanillepudding oder ein Erdbeertörtchen zu essen,[71] ist somit ein moralischer, zunächst unabhängig davon, ob er als moralisch richtig – beziehungsweise gerechtfertigt oder vernünftig (s. u.) – bezeichnet werden kann, oder nicht.

Moralisch richtige oder gute Antworten auf die Frage *Was soll ich tun?* richten sich – je nach Bedeutungsgrad des Sachverhalts – auf Verhaltensweisen oder Entscheidungen, die als sinnvoll und gut für eine bestimmte Gemeinschaft erachtet werden. Dies bringt die Frage nach dem inhärenten Machtverhältnis mit sich: Wer bestimmt, was moralisch richtig oder gut ist? Im Bezug auf grundlegende Werte wie den Schutz von Menschen würden vermutlich ebenso einige Einzelpersonen wie Vertreter∗innen verschiedener Staaten zustimmen, dass es sich hierbei um etwas Erstrebenswertes handelt. Auf der anderen Seite spielen im politischen Kontext häufig auch wirtschaftliche Faktoren eine große Rolle oder man denke an Bewegungen wie den Antinatalismus, wonach es vernünftig ist, Geburten zu verhindern, um beispielsweise die Klimakatastrophe zu mindern.

Bei der Beschreibung des Ziels moralisch guten Handelns wird häufig auf die Antike verwiesen, in der *eudaimonía*, also gesamtgesellschaftliches *Glück* oder *Wohlergehen* als zuvorderst erstrebenswert galt, im Unterschied zu dem zeitgenössischen Fokus auf individuelles Glück (wie auch immer dieses definiert wird). Es ist jedoch wahrscheinlicher, dass nicht die *Zielgruppe* des Glücks in unterschiedlichen Epochen variierte, sondern dass es sich vielmehr um eine grundlegende, topologische Unterscheidung handelt, welchen Personenkreis man in den Blick nimmt; zeithistorisch gesehen mag es dabei kontextbedingte Präferenzen gegeben haben, welche Personengruppe im Vordergrund standen. Demnach kann auch moralisches Handeln Konsequenzen für eine Person allein haben (Vanillepudding versus Erdbeertörtchen; s. o.), für eine kleinere Gruppe (z. B. eine Kernfamilie), für eine größere Gruppe (z. B. eine Gesellschaft innerhalb eines Landes) oder für alle Menschen auf der Welt. Maschinenethische Fragen können im Einzelnen zwar auch kleinere Gruppen betreffen (zum Beispiel die Arbeitsbedingungen für Menschen in einem

[71] S. hierzu S. 22.

Unternehmen, in dem auch Roboter eingesetzt werden), jedoch sind die meisten prinzipiellen Fragen, die sich stellen, von nationaler oder internationaler Relevanz – denn es sind häufig verschiedene Länder an der Entwicklung von Programmen und Technik beteiligt, und grundlegende Fragen bezüglich des Machtverhältnisses von Menschen und Maschinen betreffen alle Menschen.

Werden folglich technische Systeme als moralische Akteure diskutiert, so geht es um das moralisch richtige Zusammenleben von Menschen in einer Umgebung immer komplexerer technischer Systeme. Moralisch richtig heißt in diesem Zusammenhang *gut für alle Beteiligten* oder, wenn man es utilitaristisch ausdrücken möchte (s. u.), *so gut wie möglich*. Die konkreten Inhalte eines *guten (Zusammen-) Lebens* müssen dabei freilich definiert werden – und können international wie interkulturell sehr unterschiedlich sein. Je nachdem, um welchen Personenkreis es sich handelt, sehen mögliche ethische Richtwerte anders aus. Was in einer Gesellschaft insbesondere im Hinblick auf konkrete technische Systeme, den Einsatz von *deep fakes*, *künstlicher Intelligenz* usw., als *das Gute* erachtet wird, muss daher definiert werden – und zwar so konkret wie möglich.

Geht es in maschinenethischen Diskursen allzu oft, insbesondere bei der Entwicklung künstlicher Intelligenz, lediglich um die technische Machbarkeit, verkommt die genuin ethische Frage nach dem guten Leben zuweilen zu einem, oft im letzten Satz eines Aufsatzes oder Buchs zu findenden, Appell an den Humanismus. Dies liegt sicher auch am Primat wirtschaftlicher Interessen und Entwicklungen (wo realistischer Weise die meisten technischen Umwälzungen heute ihren Ursprung haben), bei denen ethische Einwände trotz aller scheinbaren Einflussmöglichkeiten[72] keine oder nur eine nominelle Rolle spielen. Der Journalist Thomas Ramge (2019: 116 f.) bezieht sich am Ende seiner sehr kurzweiligen Untersuchung der Chancen und Risiken künstlicher Intelligenz auf Albert Wengers Plädoyer für humanistische Ideale.

> The history of humanity is the sum of human decisions. We decide normatively what we want. This will remain the case. We do not even have to reinvent the positive worldview that is required for the next step in the development of the machine-aided information age: ‚Very simply, it's a return to humanistic values,' says the New Yorker venture capitalist, author, and TED speaker Albert Wenger. These values can in his view be expressed by a formula: ‚The ability to create knowledge is what makes us human beings unique. Knowledge arises through a critical process. Everyone can and should take part in this process.' The digital revolution allows us

[72] Zu nennen ist hier beispielsweise die „Arbeitsgruppe Mensch und Maschine" des deutschen Ethikrats: https://www.ethikrat.org/themen/aktuelle-ethikratthemen/mensch-und-maschine/, aufgerufen am 01.03.2022; sowie zahlreiche ethische Ansätze und Versuche der Regulierung durch andere Institutionen in Deutschland und international. S. hierzu auch Abschnitt 6.3.

> to put this humanist ideal into practice for the first time in history – by employing artificial intelligence intelligently and for the good of humanity.

Sicher sind die Fragen nach den theoretischen Definitions- und praktischen Umsetzungsmöglichkeiten moralischer Techniksysteme ein wichtiger Bestandteil ethischer Debatten. Dabei sollte das Kernziel der Ethik, also die Diskussion um die Frage nach dem *guten Leben*, nicht außer Acht gelassen werden oder zu einer Randnotiz verkommen (vgl. hierzu auch Steinfath 2021).

4.4 Moralische Akteure: Ein paar Beispiele

Beispiele für moralische Maschinen gibt es viele und sie wurden und werden in der Literatur bereits zahlreich behandelt; sei es der Einsatz künstlicher Intelligenz für militärische Zwecke (Beyerer und Martini 2020), in der Kardiologie (Lopez-Jimenez et al. 2020) oder in der Automobilindustrie (Tesla Inc. 2021). Wie oben bereits angesprochen, spielen statistische Sprachmodelle, die die natürliche Sprache immer besser verstehen und reproduzieren können, in der Spracherkennung und -verarbeitung eine immer größere Rolle:

> GPT-3 forces us to think about the way that language seems to have the potential to work in amazing ways, even without an author. We are already teaching our own students to harness its power as an important cognitive tool for writing, much as it's now commonplace to use spellcheck and Grammarly. If it can help us to create, to understand—at least partially—what it means to write like a particular author, and to look more deeply into the meaning of ‚meaning,' then AI can serve as both a mirror onto ourselves and a window onto others. Today's GPT-3 shows us that what we thought was most human might eventually become replicable using augmented GPT-n Transformers-like architectures. But it also affords us insight into the amazing power of our language games, which are key to understanding what it means to be human.
>
> (Elkins und Chun 2020)

Grundsätzlich wird unterschieden zwischen moralischen Subjekten (im Englischen meist als *moral agents* bezeichnet) und moralischen Objekten (*moral patients*) (vgl. z. B. Floridi und Sanders 2004: 349 ff.). Da es sich nicht um menschliche Entitäten oder nicht-menschliche Lebewesen handelt, wird auch von *artificial moral agents* beziehungsweise *patients* gesprochen. Die Bezeichnungen und Abkürzungen hierfür variieren sowohl in der deutsch- als auch der englischsprachigen Literatur stark – auch abhängig von den jeweiligen Vorhaben in einem Text. Loh (2019a) verwendet die Ausdrücke *Handlungssubjekt* und *Handlungsobjekt*, um diese von den Ausdrücken *Verantwortungssubjekt* und *Verantwortungsobjekt* abzugrenzen.

Wallach und Allen (2009: 4 ff.) sprechen von *AMA*s – *artificial moral agents*.[73] James Gips (1995) benutzt den Ausdruck *ethical machine* oder *ethical robot*; damit meint er moralische Maschinen im Sinn der vorliegenden Arbeit, da es ihm um die Frage, „how ought they to behave?“ (ebd.: 243) geht. Rath (2019) und Misselhorn (2019) (u. a.) verwenden den häufig im Deutschen verwendeten Ausdruck *moralischer Akteur*.[74] Bei den in Fachtexten selten anzutreffenden Übersetzungen *moralischer Agent* und *moralischer Patient* handelt es sich um *false friends*.[75]
Die Unterscheidung zwischen moralischen Subjekten und Objekten wird getroffen, um den Status technischer Systeme genauer zu bestimmen. Subjekte haben dabei natürlich einen größeren Einfluss auf ein soziales Gefüge, da sie Handlungen ausführen können; Objekte können lediglich moralisch berücksichtigt werden (oder auch nicht):

> Artificial agents (AAs), particularly but not only those in Cyberspace, extend the class of entities that can be involved in moral situations. For they can be conceived of as moral patients (as entities that can be acted upon for good or evil) and also as moral agents (as entities that can perform actions, again for good or evil).
> (Floridi und Sanders 2004: 349)

Im Folgenden werde ich ein paar Ansätze darlegen, die die Möglichkeit technischer Systeme als moralische Akteure erörtern.[76] Dabei wird auch deutlich werden, wie sich diese Ansätze vom Verständnis moralischer Maschinen in der vorliegenden Arbeit abgrenzen.

[73] Wallach und Allen (2009) definieren AMAs jedoch in meinem Verständnis ethischer Maschinen (s. u.). Das wird zum Beispiel daran deutlich, dass sie für AMAs *moral decision-making* für notwendig halten, aber auch andere Voraussetzungen, wie weiter unten deutlich werden wird.

[74] Janina Loh ist die einzige mir bekannte Fachwissenschaftlerin, die auch von *Akteurinnen und Akteuren* beziehungsweise *Akteur∗innen* spricht (vgl. z. B. Loh 2019a: 54; vgl. hierzu auch ebd.: 11, Fußnote 5). Möglicherweise ist Gendern in der Maschinen- und Roboterethik so unüblich, weil man davon ausgeht, dass Technik ohnehin (geschlechts-)neutral ist. Dass technische Systeme nicht wertneutral sind, wurde oben bereits gezeigt. Viele, insbesondere sozialwissenschaftliche Studien, zeigen zudem, dass in der Robotik durchaus männliche und weibliche Stereotype existieren beziehungsweise reproduziert werden und diese sich zuweilen auch in der äußeren Erscheinung von Maschinen niederschlagen, vgl. zur Thematik u. a. Marsiske 2021; Rothaas 2016.

[75] Vgl. z. B. https://de.wikipedia.org/wiki/Roboterethik, aufgerufen am 01.03.2022.

[76] Die von mir ausgewählten Beispiele finden sich auch bei anderen Autor∗innen, wie zum Beispiel Loh 2019a; Misselhorn 2019; Wallach und Asaro 2016. Ich werde sie jedoch unter Voraussetzung meiner Definition moralischer Techniksysteme kritisch beleuchten – und auch in Abgrenzung zu ethischen Techniksystemen. Die folgende Diskussion der Ansätze stellt einen Grundpfeiler für die Entwicklung genuin ethische Maschinen dar; viele weitere Ansätze hätten ergänzt werden können.

James Moor und die vier Typen moralischer Maschinen:

Zum *klassischen Kanon der Roboterethik* gehört nach Janina Loh (2019a: 55) die Unterteilung James Moors von moralischen Maschinen in vier Typen (vgl. Moor 2006). Wie in Moors vier Seiten langem Artikel schnell deutlich wird, ist er ein Vertreter der intrinsischen Normativität der Technikentwicklung:[77] „By its nature, computing technology is normative“ (ebd.: 18). Die intrinsische Normativität sämtlicher technischer Systeme bedingt, dass für ihn „jedes Gerät ein moralischer Akteur [ist; L. S.], das gute oder schlechte Auswirkungen auf den Menschen“ hat (Misselhorn 2019: 70).[78] Dies ist für Moor die erste Klasse moralischer Akteure, die er als *ethical-impact agents* (vgl. Moor 2006: 19) bezeichnet. Catrin Misselhorn (2019: 70) und Janina Loh (2019a: 56) kritisieren, dass in diesem Verständnis auch Naturphänomene (wie bspw. eine Lawine) als moralische Akteure gelten müssten. Es stellt sich jedoch die Frage, inwieweit dies problematisch ist oder nicht einfach als Konsequenz der Definition akzeptiert werden kann. Demnach können auch Institutionen oder soziale Gesellschaften als *moralische Akteure* gelten. Matthias Rath (2019: 227) führt dazu aus:

> Denn unabhängig von philosophischen Differenzierungen nach Handlungen als Tun von sinnsetzenden (humanen) Individuen [sic!] haben wir es in der Realität mit der alltagspraktischen Selbstverständlichkeit zu tun, dass Wirtschaftsunternehmen ‚von außen grundsätzlich als moralfähig und für ‚ihr‘ Handeln verantwortlich angesehen‘ (Scholtes 2007, S. 14) werden. [...] Diese alltagspraktische Überzeugung hat damit zu tun, dass wir Unternehmen als ‚handelnd‘ erleben. Ihnen werden Interessen und Intentionen unterstellt.[79]

Betrachtet man auch die anderen drei Typen moralischer Akteure von James Moor, wird klar, dass hierbei keine Rede von getrennten Kategorien sein kann, sondern dass es sich vielmehr um verschiedene Perspektiven auf denselben Gegenstand handelt. Auf der zweiten Stufe sieht Moor *implicit ethical agents*: „Computers are implicit ethical agents when the machine's construction addresses safety or critical reliability concerns“ (vgl. Moor 2006: 19). Als Beispiel nennt Moor einen Bankautomaten; zitiert nach Misselhorn (2019: 71): „Diese Maschinen müssen nicht nur stets den richtigen Betrag ausgeben, sie überprüfen oft auch, ob an diesem Tag überhaupt noch Geld abgehoben werden darf“. Moor gibt ein weiteres Beispiel:

[77] S. auch Seite 25 f. in dieser Arbeit.

[78] Hier kann freilich angemerkt werden, dass keine Technik einfach so *ist* wie sie *ist*, sondern stets von Menschen konzipiert und umgesetzt wurde und wird uns somit schon die durch Menschen verantwortete Herstellung einen Teil der Folgen beinhaltet.

[79] Das Binnenzitat bezieht sich auf Scholtes 2007.

> Another example of a machine that's an implicit ethical agent is an airplane's automatic pilot. If an airline promises the plane's passengers a destination, the plane must arrive at that destination on time and safely. These are ethical outcomes that engineers design into the automatic pilot. Other built-in devices warn humans or machines if an object is too close or the fuel supply is low.
> (Moor 2006: 19)

Es geht hier also um Techniksysteme, „in deren Konstruktionsweise sich bestimmte moralische Wertvorstellungen niederschlagen“ (Misselhorn 2019: 71). Gemäß der Einwände gegen die Neutralitätsthese oben und der daraus folgenden Annahme, dass technische Systeme inhärent normativ sind, müsste für diese Eigenschaft keine eigene Kategorie benannt, sondern kann die Aussage auf technische Systeme generell übertragen werden. Sogar für Naturkatastrophen, die im obigen Beispiel als *ethical impact agents*, also moralische Akteure, benannt wurden, kann man die These anfügen, dass Menschen durch ihr Verhalten die Systeme so beeinflussen, dass die Konsequenz menschengemachte, negative Folgen sind.

Ferner nennt Moor die dritte Kategorie *explicit ethical agents*. Im Unterschied zu vorangegangenen Beispielen, führt Moor hier die Möglichkeiten an, ethische Prinzipien oder Theorien, wie beispielsweise deontologische Grundsätze, in Maschinen zu implementieren.

> Can a machine represent ethical categories and perform analysis in the sense that a computer can represent and analyze inventory or tax information? Can a machine ‚do‘ ethics like a computer can play chess?
> (Moor 2006: 19 f.)

Zwei Dinge fallen hier auf: Zum einen verwendet Moor den Ausdruck *ethical*, obwohl es sich im Verständnis der vorliegenden Arbeit ganz offensichtlich um moralische Maschinen handelt. Warum? Weil es um Maschinen oder Programme geht, die in mehr oder weniger komplexen, technischen Systemen bestimmte, inhärente Regeln befolgen, die sich zuvor Menschen (Ingenieur∗innen, Entwickler∗innen, Unternehmer∗innen usw.) ausgedacht und in die Maschinen implementiert haben. Zum anderen scheint die Unterscheidung zwischen *implicit ethical agents* und *explicit ethical agents* keine prinzipielle zu sein, sondern lediglich eine Frage der Perspektive, was an Moors Beispiel für einen konkreten *explicit ethical agent* deutlich wird: „Chess programs typically provide representations of the current board position, know which moves are legal, and can calculate a good next move“ (ebd.: 20). Dasselbe gilt aber auch für – technisch gesehen sogar deutlich komplexere – Systeme, wie Warnsysteme oder Bankautomaten. Von außen betrachtet, handelt es sich um eine in bestimmter Weise agierende Maschine, die jedoch vermittels ihrer Aktionen bestimmte inhärente Werte repräsentieren. Diese sind, wie oben beschrieben, sowohl bei impliziten als auch bei expliziten moralischen Maschinen in

der *Konstruktionsweise* verankert. Warnsysteme und Bankautomaten basieren also ebenso auf impliziten ethischen Prinzipien wie Katastrophenschutz-Programme (ein weiteres Beispiel Moors, Moor 2006: 20) und sie haben ebenso explizite Repräsentationen ihrer Kalkulationen wie ein Schachcomputer.[80] Die unterschiedlichen Kategorien scheinen also eher verschiedene Perspektiven auf ein und denselben Sachverhalt darzustellen.

Mit Bezug auf Van den Hoven und Lockhorst (2002) beschreibt Moor (2006: 20) auch mögliche ethische Ansätze, gemäß denen moralische Techniksysteme agieren könnten: „*deontic* logic for statements of permission and obligation, *epistemic* logic for statements of beliefs and knowledge, and *action* logic for statements about actions“ (Herv. i. Orig.). Mit *logic* sind hier offenbar Programmabläufe gemeint, die in Kombination ein technisches System zu mehr oder weniger eigenständigen Aktionen befähigen könnten.

> Together, these logics suggest that a formal apparatus exists that could describe ethical situations with sufficient precision to make ethical judgments by machine. For example, you could use a combination of these logics to state explicitly what action is allowed and what is forbidden in transferring personal information to protect privacy. In a hospital, for example, you'd program a computer to let some personnel access some information and to calculate which actions what person should take and who should be informed about those actions.
>
> (Ebd.)

So befähigt, könnten Maschinen, konstatiert Moor (ebd.) bezüglich *explicit ethical agents*, „be more competent than humans“. Die Techniksysteme würden im Einzelfall sogar darüber entscheiden („determine“), „who would live and who would die“ (ebd.). Diese Behauptung oder Sorge, die in der vorliegenden Arbeit schon häufiger aufgetaucht ist, kann zumindest in Bezug auf Moors Ansatz zurückgewiesen werden, indem darauf verwiesen wird, dass die moralischen Prinzipien selbst bei Moors *explicit ethical agents* von Menschen kommen würden. Ein Programm könnte dann, wie oben zitiert, *know which moves are legal, and calculate a good next move*, aber es kann, sofern von einer moralischen Maschine die Rede ist, keine moralischen Gesetze selbst entwickeln. Ein Schachcomputer kann vielleicht den nächsten Schritt berechnen, aber er kann nicht von sich aus entscheiden, lieber *Go* zu spielen oder eher als Krebserkennungsprogramm zu dienen. Moors Befürchtung, „[s]ome might say that only humans should make such decisions“ (ebd.), kann also mit der Aussage zurückgewiesen werden, dass (bis heute) auch tatsächlich nur

[80] Anhand der Beispiele Moors, die im aktuellen Kontext recht veraltet erscheinen, wird auch deutlich, wie schnell die technischen Entwicklungen allein in den letzten beiden Jahrzehnten vorangeschritten sind.

Menschen die Entscheidungen treffen, was (moralische) Programme können sollen und was nicht.

Anders sieht das aus bei Moors vierter Kategorie: *full ethical agents*. Diese können gemäß seiner Konzeption nicht nur auf der Basis inhärenter ethischen Prinzipien handeln, also moralisch *sein*, sondern auch ethische Begründungen reflektieren und beurteilen (*can make explicit ethical judgments*) sowie diese vernünftigerweise rechtfertigen (*reasonably justify*) (Moor 2006: 20). Unter diesen Voraussetzungen kann Moors folgende Aussage bezweifelt werden: „An average adult human is a full ethical agent" (ebd.). Demnach gäbe es entweder nicht viele durchschnittliche erwachsene Menschen, was der Bedeutung von *durchschnittlich* widersprechen würde – oder Moors Behauptung ist schlicht falsch, zumal er es für seine vierte Kategorie nicht bei der Reflexions- und Rechtfertigungsfähigkeit moralischer Urteile belässt, die er einer durchschnittlichen erwachsenen Person zuschreibt, sondern diesen auch „consciousness, intentionality, and free will" (ebd.) attestiert. Ob diese im Durchschnitt allen erwachsenen Personen (und nur diesen?) pauschal zugeschrieben werden können, ist fragwürdig, spielt jedoch für Aussagen über die Akteursfähigkeit technischer Systeme nur eine untergeordnete Rolle.

Als Vorgriff auf die weitere Analyse kann an dieser Stelle angemerkt werden, dass es sich bei Moors *full ethical agents* tatsächlich um genuin ethische Akteure im Sinn der vorliegenden Arbeit handelt. Die Voraussetzungen hierfür und Moors Position zu der Frage, ob Maschinen *full ethical agents* sein können, werde ich daher weiter unten diskutieren. Die ersten drei Kategorien stellen meiner Analyse zufolge verschiedene Perspektiven auf denselben Gegenstand dar, welche die moralische Akteursfähigkeit von technischen Systemen beleuchten. Demnach können technische Systeme moralisch agieren, was jedoch noch keine Bewertung – also moralisch *gute* oder *schlechte* Aktionen – beinhaltet. Auf ethischer Ebene handelt es sich hierbei um eine relativistische Perspektive, da die ethische Bewertung von der Beurteilung einer Person, eines Personenkreises sowie dem jeweiligen Kontext abhängt; nicht zu verwechseln ist dies mit einem metaethischen Relativismus, also der Aussage, dass es keine objektiv gültigen moralischen Aussagen gibt (vgl. z. B. Ernst 2009: 81 ff.).

Luciano Floridi und John Sanders: interactivity, autonomy, adaptability

Floridi und Sanders (2004: 349 ff.) plädieren für eine *mind-less morality* moralischer Akteure. Konzepte, wie „free will, mental states or responsibility", halten sie bei der Definition von *moral agents* eher für hinderlich oder zumindest für *nicht notwendig* (vgl. ebd.: 349 ff.). Folgende Kriterien halten sie jedoch für moralische Maschinen beziehungsweise *moral agenthood* für notwendig:

> [...] interactivity (response to stimulus by change of state), autonomy (ability to change state without stimulus) and adaptability (ability to change the ‚transition

> rules‘ by which state is changed).
> (Floridi und Sanders 2004: 349)

Eine ausführlichere Definition der Bedingungen liefern sie hier:

> (a) Interactivity means that the agent and its environment (can) act upon each other. Typical examples include input or output of a value, or simultaneous engagement of an action by both agent and patient – for example gravitational force between bodies.
> (b) Autonomy means that the agent is able to change state without direct response to interaction: it can perform internal transitions to change its state. So an agent must have at least two states. This property imbues an agent with a certain degree of complexity and independence from its environment.
> (c) Adaptability means that the agent's interactions (can) change the transition rules by which it changes state. This property ensures that an agent might be viewed, at the given LoA, as learning its own mode of operation in a way which depends critically on its experience. Note that if an agent's transition rules are stored as part of its internal state, discernible at this LoA, then adaptability follows from the other two conditions.
> (Ebd.: 357 f.)[81]

Ein Beispiel:

> Henry is an agent if Henry is a system, situated within and a part of an environment, which initiates a transformation, produces an effect or exerts power on it, as contrasted with a system that is (at least initially) acted on or responds to it, called the patient.
> (Ebd.: 355)

Diese Definition von *Akteur* entspricht Moors Beschreibung von *ethical impact agents*. Catrin Misselhorn (2019: 77 f.) weist darauf hin, dass es Unterschiede bezüglich der Bedingungen geben kann, in denen Akteure – insbesondere verschiedene technische Systeme – Folgen bewirken können; und zwar hinsichtlich der Verarbeitungskapazitäten der Systeme:

> Allgemein kann man einem Computer eine gewisse Interaktivität zusprechen, insofern er einen Input aus der Umwelt aufgreift und einen Output erzeugt, der zu gewissen Veränderungen in der Umwelt führt. Künstliche Systeme unterscheiden sich jedoch darin, wie viele Arten von Input akzeptiert werden, wie genau dieser Input spezifiziert werden muss und wie viele Formen des Outputs es gibt. Nicht zuletzt gibt es verschiedene Wege, die vom Input zum Output führen.

[81] Mit LoA ist *Level od Abstraction* gemeint, s. u.

Entscheidend für eine angemessene Definition im Kontext von technischen Systemen als potenziell moralischen Akteuren ist laut Luciano Floridi und John Sanders daher der *Grad der Abstraktion* (*Level of Abstraction*, *LoA*), anhand dessen sich sämtliche Definitionen und so auch die der Akteursfähigkeit von Maschinen bemessen lasse. „The LoA is determined by the way in which one chooses to describe, analyse and discuss a system and its context“ (Floridi und Sanders 2004: 349). Was die Autoren mit *LoA* genau meinen, wird in der folgenden Passage deutlicher:

> The example of intelligence or ‚thinking‘ behaviour is enlightening. One might define ‚intelligence‘ in a myriad of ways; many LoA are all equally convincing and no single, absolute, definition is adequate in every context. Turing solved the problem of ‚defining‘ intelligence by first fixing a LoA – in this case a dialogue conducted by computer interface – and then establishing the necessary and sufficient conditions for a computing system to count as intelligent at that LoA: the communication game. The LoA is crucial and changing it invalidates the test, as Searle was able to show by adopting a new LoA represented by the Chinese room game.[82]
>
> (Ebd.: 353)

Anders ausgedrückt meinen die Autoren mit *Level of Abstraction*, dass ein und derselbe Ausdruck in unterschiedlich komplexen Kontexten unterschiedlich verwendet werden kann. Anstatt von $\text{Intelligenz}_{Turing}$ oder $\text{Intelligenz}_{Maschine}$ im Unterschied zu $\text{Intelligenz}_{Mensch}$ zu sprechen, sei es angemessen, zu akzeptieren, dass der Ausdruck *Intelligenz* in verschiedenen Kontexten und in Bezug auf verschiedene Akteure unterschiedlich (komplex) verwendet werden könne.

Im Folgenden gehen Floridi und Sanders auf mögliche Aspekte ein, die die moralische Handlung eines menschlichen von einem maschinellen Akteur unterscheiden könnten: Ziele haben, intentionale Zustände haben, frei sein und Verantwortung für Handlungen tragen (können). Zum ersten Punkt erklären die Autoren, dass auch maschinelle Akteure so verstanden werden können, dass sie Ziele haben, die sie verfolgen; ein Programm wird mit einem bestimmten Zweck konstruiert, zum Beispiel, das Spiel Schach zu gewinnen. Mit intentionalen Zuständen meinen sie, dass ein Akteur „must relate itself to its actions in some more profound way, involving meaning, wishing or wanting to act in a certain way, and being epistemically aware of its behavior“ (ebd.: 361). Die Autoren meinen, dass intentionale Zustände „nice but unnecessary condition for the occurrence of moral agenthood“ (ebd.) sind, mit der Begründung, dass es eines „privileged access [...] to the agent's mental or intentional states“ (ebd.) für Intentionalität bedürfe, entweder in Form eines göttlichen Auges von außen oder mittels eines Cartesianischen Innenblicks. Doch die Annahme, dass Maschinen keine Intentionen in diesem Verständnis haben *können* (sofern man dieser folgt), bedeutet ja nicht, dass Intentionalität keine

[82] Zum Gedankenexperiment *Chinesisches Zimmer* von John Searle s. auch S. 126 f.

notwendige Voraussetzung für moralische Akteursfähigkeit ist. Die eine Frage lautet, was (komplexe) technische Systeme können (und was nicht), und die andere Frage lautet, welche Bedingungen notwendig sind für *moral agenthood*. ()Ausführlicher gehe ich auf den Begriff der Intentionalität in Abschnitt 5.2.7 ein.)

Bezüglich der Frage, ob moralische Akteure *frei* sein müssen, um als solche zu gelten, antworten die Autoren: „The AAs [artificial agents; L. S.] are already free in the sense of being non-deterministic systems“ (Floridi und Sanders 2004: 362). Entgegen der anschließenden Behauptung der Autoren, diese Aussage sei unbestritten, kann jedoch angeführt werden, dass es keineswegs unumstritten ist, ob künstliche Akteure (AAs) determiniert sind oder nicht – und ob *being non-determinitstic* bedeuten würde, *free* zu sein. Im Fall der Definition der vorliegenden Arbeit sind moralische Akteure insofern determiniert, da ihre Handlungen an ein klares Regelwerk (moralische Prinzipien und Normen) gebunden sind (s. hierzu meine Definition von Algorithmus auf S. 27). Floridi und Sanders stellen die obige Behauptung nach eigener Aussage auf, um der ihrer Ansicht nach „thorny debate about the reasonableness of determinism“ (ebd.) aus dem Weg zu gehen.

Autonomie definieren die Autoren sehr niedrigschwellig im bereits oben beschriebenen Verständnis (vgl. ebd.: 356), sodass ihnen zufolge auch ein Pendel („pendulum“) sowie ein „closed ecosystem“ und das „solar system“ als *autonom* gelten, da sie einen Zustand ohne externen Stimulus ändern können. Doch auch ein Pendel braucht einen initialen Anstoß, um sich in Bewegung zu setzen – ebenso kann man beim Sonnensystem die kausalen Ursachen und ursprünglichen Bedingungen für Veränderungen diskutieren.

Freiheit, *Autonomie* und *Ziele haben* sind für Floridi und Sanders also Voraussetzungen, die für (künstlich) moralische Akteure zutreffen, wobei zum einen ihre Definitionen zum Teil stark von denen anderer Autor∗innen abweichen (s. hierzu Kapitel 5), zum anderen weichen sie mit ihrer Definition via unterschiedlicher Grade von Abstraktionen konkreten Bestimmungen aus. Wenn demnach bereits eine Lampe als *intelligent* bezeichnet werden kann, da sie ihren Zustand von „dunkel“ auf „leuchtend“ ändert infolge der Betätigung eines Schalters, dann kann die Sinnhaftigkeit des Gebrauch des Ausdrucks Intelligenz in Anwendung auf technische Systeme grundsätzlich infrage gestellt werden.

Moral definieren die Autoren schließlich als Schwellenwert (*threshold*). Wenn man sich in einem Moralkreis zum Beispiel auf die Erhaltung eines bestimmten Werts geeinigt hat, so kann eine Handlung als moralisch gelten, die diesen Wert berücksichtigt; unmoralisch ist ein Akteur, der dem Wert zuwiderhandelt (ebd.: 364). Als Beispiele werden im Folgenden Toleranz und *well-being* von Patienten (in einem speziellen Fall) genannt. Die Einhaltung moralischer Regeln als notwendige Voraussetzung für moralische Akteure wird auch in der vorliegenden Arbeit als sinnvoll erachtet, wobei ich an dieser Stelle noch einmal auf Abbildung 3 hinweisen

möchte. Mit dieser Definition moralischer Akteure ist die linke Spalte, der jeweilige *Inhalt* der Moral oder ethische Prinzipien und Normen, noch nicht bestimmt. Dies muss für technische Systeme notwendigerweise geleistet werden. Ab wann ist das Wohlbefinden von Patienten beim Einsatz technischer Hilfsmittel nicht mehr gewährleistet? Wo liegen die Grenzen der Toleranz in einem bestimmten Kontext? Auch die Fälle möglicher Unterschreitungen der Schwellenwerte von Moral müssen dabei diskutiert werden. So ist denkbar, dass zwar das Wohlbefinden von Patienten nicht derart gefährdet wird, dass sie aktiv getötet werden, aber auch eine psychische und physische Vernachlässigung kann eine Gefahr des Wohlbefindens darstellen. Die Rede von Moral als *Schwellenwert* suggeriert außerdem, dass moralische Grenzen immer eindeutig wären, was keinesfalls so ist.

Wendell Wallach und Colin Allen: artificial moral agents

Wendell Wallachs und Colin Allens Ansatz künstlicher moralischer Systeme (*artificial moral agents* (*AMAs*)) ist sehr voraussetzungsreich:

> People don't want AMAs to replicate the abstractions of moral philosophers any more than they want their neighbors to do so. People want their neighbors to have the capacity to respond flexibly and sensitively in real and virtual environments. They want to have confidence that their neighbors' behavior will satisfy appropriate norms, and that they can trust their neighbors' actions. Meeting this challenge will entail an even more thorough understanding of human ethical behavior than is presently available. That is, building AMAs forces one to take a particularly comprehensive approach to ethical decision making. It is important, we think, that the project of building AMAs highlights the need for a richer understanding of human morality.
>
> (Wallach und Allen 2009: 215 f.)

Wallach und Allen gehen davon aus, dass moralische Maschinen (ebenso wie moralische Menschen) nicht nur moralisch handeln, sondern ihre Handlungen auch im Hinblick auf bestimmte Normen reflektieren und gegeneinander abwägen können. Dies sind Eigenschaften, die ich allerdings ethischen Menschen und Techniksystemen zuschreiben würde, da es für das reine Sich-an-Regeln-halten zumindest nicht notwendigerweise einer Reflexionsfähigkeit bedarf. Freilich ist es möglich, dass auch moralisch handelnde Menschen oder agierende Maschinen über die Regeln nachdenken und alle Voraussetzungen erfüllen, derer es für *ethical decision-making* (s. hierzu Kapitel 5) bedarf, aber ich halte dies nicht für eine Voraussetzung moralischen Handelns. Eine Antwort zu finden auf die Frage *Was soll ich tun?*, setzt bei Menschen ebenso wie bei Maschinen einen inhärenten Werterahmen voraus, innerhalb dessen Entscheidungen getroffen werden können; jener muss aber der handelnden Person nicht bewusst sein, die Regeln müssen nicht verstanden oder

reflektiert werden können, selbst wenn dies durchaus von Vorteil sein kann. Dies bestätigt auch die mögliche Zuschreibung moralischer Akteursfähigkeit auch an verhältnismäßig wenig komplexe technische Systeme (s. o.). Dementsprechend fallen AMAs gemäß der Definition von Wendell Wallach und Colin Allen in der Kategorisierung der vorliegenden Arbeit in den Bereich ethischer Maschinen, mit denen ich mich unten auseinandersetzen werde.

Viele Ansätze mehr könnten angeführt werden, die sich mit der Moralfähigkeit von Maschinen auseinandersetzen. Die Angeführten sollen exemplarisch mögliche Diskussionsfaktoren verdeutlichen und später die Unterschiede zur Definition ethischer technischer Systeme verdeutlichen.

4.5 Beispiele für mögliche Risiken

2016 veröffentlichte das Unternehmen *Microsoft* ein lernfähiges Chat-Programm namens *Tay*. Das Programm wurde als Account bei der Mikroblog-Plattform *Twitter* installiert und sollte anhand der auf seine *Tweets* gesendeten Antworten selbst neue *Tweets* generieren. Binnen weniger Stunden trat ein Problem zutage, das im Fachjargon auch als *Dirty Data* (vgl. z. B. Richardson, Schultz und Crawford 2019) oder *shit in, shit out* (vgl. Gandorfer 2015) bekannt ist: Das Programm generierte überwiegend rassistische, sexistische und misogyne Äußerungen (vgl. Graff 2016).[83] Was war passiert? Das Programm basierte auf dem Prinzip des maschinellen Lernens (*machine learning*), das heißt, vereinfacht ausgedrückt, dass es aufgrund der Daten, mit denen es gespeist wurde, Muster erkennen sollte, anhand derer es eigene *Tweets* generieren sollte.[84] Anhand der Äußerungen, die *Tay* von anderen Accounts bei *Twitter* zugespielt wurden, generierte das Programm eigene Antworten, die die zitierten Äußerungen zur Folge hatten.

Im Fall von *Tay* handelte es sich um ein gezielte *Speisung* mit „falschen Werten“ (vgl. ebd.). Es gibt aber auch andere Fälle, in denen sich Ungerechtigkeiten, die sich in der Vergangenheit in Daten niederschlugen, in der Gegenwart reproduzieren – zum Beispiel eine Bevorzugung von Männern für bestimmte Tätigkeiten. Durch ungefiltertes *machine learning* können solche Ungleichverteilungen (*bias*) auch in der Gegenwart reproduziert werden. Bias kann aber auch Vorteile mit sich bringen, wie Reinhard Heil (2021: 427) anmerkt:

[83] Beispiele für solche *Tweets* sind „Hitler hatte recht. Ich hasse Juden.“, „Ich bin eine nette Person. Ich hasse alle Menschen.“, „Bush hat 9/11 selber verursacht, und Hitler hätte den Job besser gemacht als der Affe, den wir nun haben. Unsere einzige Hoffnung jetzt ist Donald Trump.“ und „Ich hasse alle Feministen, sie sollen in der Hölle schmoren.“ (Graff 2016).

[84] Zur Ethik maschinellen Lernens wurde in den vergangenen Jahren sehr viel publiziert, hier nur zwei Literaturhinweise für einen ersten Überblick: Decker 2016; Heil 2021.

> Für KI Systeme gilt, dass die Qualität der Ergebnisse nicht nur von den eingesetzten Algorithmen abhängt, sondern auch davon, ob die verwendeten Daten einen Bias aufweisen, also verzerrt sind. Die Nutzung von KI verschärft dieses, aus der Statistik bekannte, Problem bedingt durch deren epistemische Opazität. Künstliche Intelligenz kann gesellschaftliche Vorurteile, die sich in den Daten widerspiegeln, reproduzieren oder gar verstärken, sie kann aber auch zu deren Aufdeckung beitragen. Ethik spielt also nicht nur beim Systemdesign eine große Rolle, sondern auch bei der Bewertung der Eingangsdaten, der Ergebnisse und der Einsatzbereiche.

Im Jahr 2015 kam beispielsweise heraus, dass das Unternehmen *Amazon* eine Software zur Einstellung neuen Personals benutzt hatte, die systematisch Frauen benachteiligte (vgl. Dastin 2018). In demselben Jahr trat aufgrund ähnlicher Mechanismen ein Problem mit der Foto-App des Unternehmens *Google* auf: Ein Bild eines dunkelhäutigen Pärchens wurde darin als „Gorillas“ untertitelt (vgl. Twitter 2015).[85]

Auch Algorithmen, die für sogenanntes *predictive-policing* eingesetzt werden, können das Leben einzelner Personen maßgeblich beeinflussen. Hierbei wird versucht, Einbrüche oder andere Straftaten in bestimmten Stadtvierteln, Bezirken oder Straßen einer Stadt aufgrund statistischer Annahmen für einen bestimmten Zeitraum vorherzusagen (vgl. auch Bratton, Morgan und Malinowski 2009). Das Ziel ist die Kombination von „cutting-edge surveillance technologies with predictive analytics to anticipate and prevent problems before they occur“ (Tonry 2011: 489). Wenn jedoch in bestimmten Vierteln aufgrund statistischer Werte vermehrt kontrolliert wird, können dort auch vermehrt Einbrüche und Straftaten festgestellt werden, was sich dann erneut in den Zahlen niederschlägt (dies wird auch als *rebound*-Effekt bezeichnet). Somit kann eine Verstärkung stattfinden, die sich nicht aus den tatsächlichen Umständen ergibt, sondern die aufgrund der Datenlage selbst evoziert wird. Der Rechtswissenschaftler Frank Pasquale (2015: 8) zieht in seinem Buch *The Black Box Society* folgenden Schluss: „[A]uthority is increasingly expressed algorithmically“.

All diese Beispiele zeigen: Technische Systeme können Output generieren, der erhebliche moralische Schwierigkeiten mit sich bringt, wie Diskriminierung, Rechtsradikalismus oder Stigmatisierung. Gleichwohl wäre es falsch, wie bereits an einigen Stellen in der vorliegenden Arbeit deutlich wurde, hierfür die technischen Systeme selbst zur Verantwortung zu ziehen; dies macht auch Reinhard Heil (2021: 425) deutlich:

> Die Art und Weise [sic!] wie heute oft von KI gesprochen wird, mystifiziert und überhöht sie. Die Mystifizierung ist dabei selbst eine der ethischen Herausforderun-

[85] Drei Jahre später wurde bekannt, dass das Unternehmen den Fehler *gelöst* hatte, indem die Bezeichnung „Gorilla“ aus der App gestrichen worden war (vgl. Vincent 2018).

> gen der KI. Die Bezeichnung ‚Künstliche Intelligenz' suggeriert Autonomie und legt es so nahe, KI-Anwendungen Verantwortung für Entscheidungen und Handlungen zuzuschreiben.

Insbesondere bei lernenden Systemen, deren Output sich dynamisch am Input orientiert, ist es aus ethischer Sicht wichtig, Strategien zu entwickeln, wie auch lernende Systeme bestimmten, vorab definierten moralischen Ansprüchen gerecht werden können.

4.6 Vom Paradox moralischer Akteure zu genuin ethischen Techniksystemen

Die Definition, dass alle Algorithmen per definitionem moralisch sind, sagt noch nichts darüber aus, worin das Moralische genau bei einzelnen technischen Geräten besteht. Der Kontext, in dem sich die Frage *Was soll ich tun?* stellt, variiert von Anwendung zu Anwendung, von Situation zu Situation und verändert sich mit fortschreitenden Entwicklungsmöglichkeiten immer weiter. Allzu oft kann die Frage auch nicht im Singular entschieden werden, da verschiedene Personen dafür verantwortlich sind,[86] welche Folgen ein technisches System mit sich bringt. Ob es dabei um Vanillepudding oder Erdbeertörtchen geht (vgl. Ernst 2009: 11) beziehungsweise die Frage, wie man sich an einem Bahnübergang am besten verhält (vgl. Celikates und Gosepath 2017a: 7), muss aus ethischer Perspektive von Fall zu Fall betrachtet werden, auch im Kontext des Gesellschaftskreises, für den eine ethische Bewertung vorgenommen werden soll.

Gemessen an den vorab programmierten Regeln haben die im vergangenen Abschnitt exemplarisch dargestellten Programme insofern *moralisch* agiert, als sie sich an die für sie vorgesehenen Regeln gehalten haben. Sie agierten gemäß den ihnen einprogrammierten Zwecken – und trotzdem wurden die Effekte der *tools* im Nachhinein als *unmoralisch* bewertet (mit Ausnahme der *predictive-policing*-Programme), da sie, gemessen an bestimmten ethischen Maximen, die Grenze des Moralischen überschritten haben.

Mögliche, im Hintergrund stehende moralische Prinzipien, die zwar so gegebenenfalls nicht formuliert, aber anhand der Reaktionen der verantwortlichen Betreiber∗innen und Nutzer∗innen ersichtlich sind, können beispielsweise lauten:

- Es ist unmoralisch, sexistische, antisemitische und frauenfeindliche Äußerungen von sich zu geben – zumal auf öffentlichen Plattformen mit einer breiten Sichtbarkeit.

[86] In den *Ethik-Leitlinien für eine vertrauenswürdige KI* werden beispielsweise „Entwickler", „Betreiber" und „Endnutzer" genannt (vgl. Europäische Kommission 2019: 17).

- Es ist unmoralisch, Menschen aufgrund ihrer Hautfarbe zu stigmatisieren oder zu beleidigen.
- Es ist unmoralisch, Menschen aufgrund ihres Aussehens zu stigmatisieren usw.

Aufgrund der Reaktionen der Betreiber der Programme (das Chatprogramm deaktivieren, die Foto-App umprogrammieren) kann man davon ausgehen, dass jene Maximen – so oder in anderen Formulierungen – von den Verantwortlichen vertreten und unterstützt werden. Dies so explizit zu formulieren, ist wichtig, da es bei der allgemeinen Rede von *der* Ethik *der* künstlichen Intelligenz häufig zu ungenauen Aussagen kommt, da keine allgemeingültigen ethischen Maximen für die Gesamtheit aller technischer Systeme ausgesprochen werden können (vgl. hierzu auch Hagendorff 2020). Aufgrund der öffentlichen Reaktionen (Empörung, Entsetzen usw., vgl. Steiner 2016) kann man außerdem davon ausgehen, dass viele Nutzer∗innen, die sich in die Diskussion eingeklinkt haben, ebenfalls die obigen Maximen vertreten und diese für richtig befinden. Da die Algorithmen gemäß ihrer ursprünglichen Programmierung jedoch *richtig* funktioniert haben, haben sie also zugleich moralisch richtig *und* moralisch falsch agiert. Ein Paradox?

Auf den ersten Blick handelt es sich bei dieser scheinbar widersprüchlichen Bewertung um ein Paradox, das aber unter Einbezug der obigen Definition von Moral rasch aufgelöst werden kann. Man kann davon ausgehen, dass sowohl die Programmierer∗innen als auch die verantwortlichen Personen in den jeweiligen Unternehmen die Effekte ihrer Programme nicht so intendiert haben – dies ist allerdings eine nur schwer prüfbare Unterstellung, die gleichwohl plausibel ist. Nicht nur bei öffentlich beeinflussbaren Programmen wie *Tay*, aber vermutlich insbesondere hier, besteht die Möglichkeit des Missbrauchs vorhandener Strukturen bis hin zur gezielten Verbreitung von Falschinformationen und cyberkriminellen Tätigkeiten (vgl. u. a. Capurro 2016; Eisele 2016; Heesen 2021a; Schmidt 2016). Geht man davon aus, dass diejenigen, die die Funktion der jeweiligen Programme zu verantworten hatten und haben, keine genuin unmoralischen Ziele verfolgten, so kommt das *Unmoralische* erst an späterer Stelle ins Spiel. Die Systeme wurden, dies wird in diesem Fall vorausgesetzt, mit moralisch guten Absichten entwickelt, das heißt also, mit der Intention, die maschinelle Kommunikationsfähigkeit sowie die Imitationsfähigkeit menschlicher Sprache durch automatisierte Programme zu verbessern (u. a.). In diesem Sinn haben die Systeme, wie bereits gezeigt, moralisch gehandelt, da sie ihre Lernziele einhielten und moralisch *richtig* insofern, da die Intention bei der menschlichen Entwicklung aus ethischer Perspektive nicht bedenklich war. Der Output ist / war dennoch moralisch falsch, da lernende Systeme qua Prinzip zu-

mindest ergebnisoffen sind und dieser somit nicht vorhergesehen wurde.[87] Bei der Entwicklung technischer Systeme, die auf maschinellem Lernen basieren, werden also Spielräume in Kauf genommen, die moralische (oder auch juristische) Fehler nicht immer ausschließen können. Durch die Annahme ethischer Bewertungen zu unterschiedlichen Zeitpunkten, kann das Paradox somit aufgelöst werden. Selbst wenn durch die Ergebnisoffenheit manche Autor∗innen versucht sind, den Systemen Formen von *Autonomie* (s. hierzu auch S. 106 ff.) zuzuschreiben, so bleibt es (zumindest bislang) bei einer eindeutigen Verantwortlichkeit, auch bezüglich möglicher Risikoanalysen, auf Seiten von Menschen. Eine ethische Reflexion und Bewertung komplexer technischer Systeme findet also idealerweise nicht nur einmal statt, sondern an verschiedenen kritischen Stellen im Entwicklungsprozess.

Wie nun moralische Prinzipien so in technischen Systemen verankert werden können, dass diese nicht nur verlässlich nach ihnen handeln (also moralisch sind), sondern auch moralisch *richtige* und *wünschenswerte* Folgen mit sich bringen, ist Gegenstand vieler Diskussionen in der Philosophie und Ethik. Dabei geht es auch um die Entwicklung genuin ethischer technischer Systeme gemäß meiner oben dargestellten Definition, also um mögliche technische Umsetzungen u. a. kognitiver Fähigkeiten wie Reflexion, Entscheidungsfindung und mögliche andere Fähigkeiten, die dafür als notwendig definiert werden. Die Idee, moralisch richtige Folgen auch weitgehend unabhängig von menschlichen Eingriffen zu etablieren ist die konsequente Folge aus der bisherigen Technikentwicklung zunehmend automatisierter und zugleich aufgrund zunehmender Komplexität auch fehleranfälliger Systeme.

4.7 Inter- und Transdisziplinarität als Notwendigkeit und Herausforderung

Viele Autor∗innen überlegen, wie erwünschte Folgen systematisch in Technik verankert werden können – auch vor dem Hintergrund einer zunehmenden Automatisierung technischer Systeme und damit weniger Einfluss von Menschen bei einzelnen Schritten. Grundlegend gibt es dabei transdisziplinäre Herausforderungen, wie Wallach und Allen (2009: 74 f.) beschreiben:

> But whatever the application, whether it is the elimination of enemy forces or taking care of the elderly, there is a fundamental divide between philosophers, who tend to think in terms of highly abstract principles, and engineers, who have to accomplish the actual design task. [. . . The] engineer must determine what the information requirements are for a system making moral decisions: that is, what

[87] Ob man die möglichen Risiken im Zuge einer Risikoanalyse hätte vorhersehen können, kann diskutiert werden.

> does the system need to know in order to make an informed decision, and what input devices and sensors will it need to get access to this information?

Der technische Stand ist heute freilich ein anderer als vor über zehn Jahren, aber im Kern stellt sich dasselbe Problem heute noch. Es gibt zwar viele verschiedene Ansätze, grundlegend geisteswissenschaftliche Disziplinen, wie Philosophie und Ethik, mit naturwissenschaftlichen, wie Physik, Informatik oder Robotik, zu verbinden; allerdings bleibt es dabei häufig bei einseitigen Überlegungen oder gegenseitigem Unverständnis. In der Regel wird aus philosophischer Perspektive über Technik nachgedacht (Technikphilosophie bzw. *philosophy of technology*, vgl. z. B. Irrgang 2011), es werden die ethischen Aspekte von Technik beleuchtet (vgl. z. B. Grunwald und Hillerbrand 2021), es wird aus informationstechnologischer Sicht beschrieben, was Intelligenz in technischen Systemen bedeuten kann (vgl. z. B. Russell und Norvig 2010) oder es wird aus Perspektive der Physik und Informatik die Wechselwirkungen von Technik und Philosophie beschrieben (vgl. z. B. Hehl 2016), um nur einige Beispiele möglicher Verbindungen zu nennen. Auch in vielen Studiengängen wird inzwischen versucht, Informatik und Technik mit Philosophie und Ethik zu verbinden. Wallach und Allen (2009: 75) beschreiben das Problem einer konstruktiven Verbindung mit – möglicherweise dem damaligen Zeitgeist geschuldet – recht stereotypischen Bildern der Disziplinen so:

> A well-trained ethicist is taught to recognize the complexity of moral dilemmas, and is likely to be sensitive to the inadequacy of any one approach meant to cover the range of challenges the AMA might confront. The engineer on the other hand, will be concerned that the ethicist's desire to make the system sensitive to moral considerations will add further difficulties to the already challenging task of building reliable, efficient, and safe systems. Theoretical discussions about the complexity and intractability of ethical dilemmas will not be considered helpful. While engineers generally believe that there is more than one solution to every problem, they are trained to converge on a satisfactory solution for the problem at hand. Ethicists, however, are trained to diverge from each other, arguing separate positions so as to describe as completely as possible the range of considerations and theories that may be relevant to a problem.

Tatsächlich wird immer wieder über Dilemma-Situationen, wie sie vermeintlicherweise bei der Entwicklung automatisierter Fahrzeuge entstehen können (vgl. u. a. Awad 2017), diskutiert, wobei diese Diskussionen dazu führen können, dass ethische Perspektiven von Menschen, die mit technischen Umsetzungen moralischer Werte konfrontiert sind, für unbrauchbar und nicht praktikabel befunden werden (vgl. z. B. Seng 2017). Doch selbst wenn es eher unwahrscheinlich ist, dass die Diskussion um per definitionem unlösbare Dilemmata pragmatischen Lösungsvorschläge für Ingenieur∗innen und Entwickler∗innen bietet, so sind sie doch dienlich, um die moralischen Präferenzen auszuloten und somit technischen Umsetzungen

ethischer Theorien ein Stück näher zu kommen. Die von Wallach und Allen definierte Aufgabe und Vorgehensweise von Ethiker∗innen ist daher sehr einseitig dargestellt; es ist zum Beispiel keineswegs unumstößlich, dass diese stets um eine *möglichst vollständige* Erörterung möglicher Problemlösungen bemüht sind. Wie viele Lösungsansätze diskutiert und beachtet werden, hängt vom jeweiligen Ansatz und der Art des Problems ab. Die Kompetenz von Ethiker∗innen sollte also nicht auf das Diskutieren von Dilemmata reduziert werden, da es neben der theoretischen Ethik auch das Feld der angewandten Ethik gibt, bei der die Reflexion von Normen im Hinblick auf konkrete Anwendungsfälle – und somit auch konkrete Problemlösungen – im Vordergrund stehen.

Die zentralen Fragen, die sich bei der konkreten Umsetzung, wie auch immer gearteter, moralisch guter Techniksysteme stellen, sind zum einen, worin die ethischen Normen und Prinzipien bestehen, auf die man sich für ein bestimmtes System geeinigt hat und zum zweiten, wie diese Normen, die für gut und richtig befunden wurden, in technischen Systemen praktisch umgesetzt werden können.[88] Es muss also zunächst geklärt werden, welche Vorgehensweisen und Outputs an einem technischen System in einem bestimmten Kontext als moralisch gut bewertet werden – und welche nicht. Insbesondere bei lernenden Systemen gestaltet sich dies als nicht immer einfach, da häufig die Folgen nicht immer abgesehen werden können. Umso wichtiger ist es, dass bei der möglichen Abschätzung von Technikfolgen adäquate Risikoanalysen erstellt werden.[89]

Die entscheidende Frage ist also, wie die ethischen Bewertungsrahmen und Begründungsstrategien für mögliche Maßnahmen aussehen. Die Definition von Algorithmen als per se moralisch bringt gewisse Aktionsmöglichkeiten bei technischen Systemen automatisch mit sich, die jedoch nicht für bedenklich erachtet werden müssen, solange die ethischen Bewertungen technischer Systeme a priori stattfinden und in politische Entscheidungen hineinwirken. Die große Herausforderung bei der Konzeption moralisch guter Techniksysteme besteht, anders formuliert, in der folgenden Frage: Wie gelingt es Menschen Maschinen und Programme so zu gestalten, dass diese, wenn sie mit ihrer Umwelt interagieren, dies auf eine Weise tun, die im Einklang mit zuvor gesetzten ethischen Maximen steht? Beispiele für solche moralisch guten Maschinen sind Computerprogramme, Roboter oder sonstige technische Systeme, die bereits eingesetzt werden und den definierten Kriterien entsprechen – oder hypothetische technische Systeme.

[88] Häufig ist auch von moralischen *Implementierungen* oder *Moralimplementierungen* in technischen Systemen die Rede.

[89] Ironischerweise gibt es inzwischen zahlreiche Ansätze für den Einsatz maschinellen Lernens für Risikoanalysen (vgl. z. B. Paltrinieri, Comfort und Reniers 2019).

4.8 Zusammenfassung: Technische Systeme als moralische Akteure

Laszlo (Versenyi 1974) hat in einer frühen Phase der Diskussion um moralische Maschinen einen verlockenden Vorschlag gemacht: Man könnte doch einfach alle komplizierten Begriffe, wie zum Beispiel Bewusstsein, außen vorlassen, dann würden sich gar nicht mehr so viele Unterschiede zwischen menschlicher und maschineller Moral ausmachen lassen.

> If men and machines coexist in a natural continuum in which there are no gaps, quantum jumps, or insurmountable barriers preventing the assimilation of the one to the other, then they also coexist in a moral continuum in which only relative but never absolute distinctions can be made between human and machine morality.
> (Ebd.: 248)

Angesichts des ebenfalls ausweichenden Definitionsvorschlags komplexer Begriffe mittels unterschiedlicher *Grade von Abstraktion*, scheint dieses Prozedere für einige Autor∗innen erstrebenswert zu sein. Michael Quante (2020: 51) legt hingegen in seiner Einführung in die *Philosophische Handlungstheorie* „Adäquatheitsbedingungen“ für die Definition von Handlungen fest, die – auch in Anwendung auf andere Begriffe – methodisch angemessener erscheinen. Eine Definition soll demnach sinngemäß

1. argumentativ plausibel und logisch schlüssig sein,
2. die in der Ethik anerkannten Aspekte des Begriffs sinnvoll auf technische Systeme sowie ihre Möglichkeiten, Fähigkeiten und Eigenschaften anwenden,
3. im besten Fall an ein Alltagsverständnis des Begriffs anknüpfen (vgl. ebd.).

Was Quante für den Begriff der *Handlung* festlegt, lässt sich auch auf andere Begriffe übertragen. Diese methodischen Schritte sollen in der vorliegenden Arbeit für die Begriffe *moralische Maschine* und *ethische Maschine* bestmöglich angewandt werden (bzw. worden sein), wobei dies angesichts der unterschiedlichen Komplexität technischer Systeme und fehlendem Alltagsverständnis bezüglich *ethischer Maschinen* nur bedingt umgesetzt werden kann.

Die Zusammenfassende Definition technischer Systeme als moralische Akteure setzt sich für die vorliegende Arbeit also wie folgt zusammen: Technische Systeme sind nicht wertneutral, da sie stets zu einem bestimmten Zweck und mit einer bestimmten Absicht entwickelt und produziert werden – zunächst einmal unabhängig davon, ob sie auch in diesem Sinne genutzt werden. Technische Systeme

sind insofern *Akteure*, dass sie bestimmte Folgen für Menschen und ihre Umwelt produzieren können (technischer Pragmatismus). *Moralische* Akteure sind sie, insofern sie sich auf bestimmte Weisen verhalten und somit Antworten auf die Frage *Was soll ich tun?* geben können. Gleichwohl fehlt ihnen die Möglichkeit, ihre eigenen Handlungsregeln selbst zu bestimmen. Moralisch *gute* Akteure sind technische Systeme dann, wenn die Regeln, die sie befolgen auch ethisch als moralisch gut bewertet werden können. Ob die Folgen technischer Systeme als *moralisch gut* oder *moralisch schlecht* bewertet werden hängt von der menschlichen Bewertung ab (ethischer Relativismus).

moralisches Techniksystem	moralisch *gutes* Techniksystem	Techniksystem als moralischer Akteur
Ein technisches System ist moralisch, wenn es bestimmten internen Regeln folgt.	Ein technisches System ist moralisch gut, wenn die Folgen ethisch als gut bewertet werden.	Ein technisches System ist ein moralischer Akteur, wenn es auf der Basis bestimmter Regeln einen Output generiert.

Abbildung 4: Moralische (gute) Techniksysteme und Techniksysteme als moralische Akteure.

Wird also ein Algorithmus als ein Programm definiert, das gewissen Regeln folgt, so ist jeder Algorithmus intrinsisch moralisch, sofern Moral als regelgeleitetes Handeln definiert wird. Die Konsequenz, die daraus folgt, ist: Eine mögliche Gefahr moralischer Maschinen oder ein potenzieller Missbrauch ist trotz der intrinsischen Normativität nicht in den Maschinen selbst zu suchen, sondern bei den Menschen, die sie entwickeln und benutzen. Diese verantworten auch die Werte, die Programmen und Algorithmen bereits durch die Produktionsform (*value by design*, vgl. Nissenbaum 2005 in Orwat und Bless 2016; Simon 2016; van de Poel 2021b) mitgegeben werden.

5 Theoretische Ansätze zur Umsetzung technischer Systeme als moralische und potenziell ethische Akteure

> Many believe a bright line exists between the senses of machine ethics discussed so far and a full ethical agent. For them, a machine can't cross this line. The bright line marks a crucial ontological difference between humans and whatever machines might be in the future.
> (Moor 2006: 20)

Idealerweise sieht die Verquickung ethischer Normen und technischer Systemen gemäß Wallach und Allen (2009: 71) so aus:

> [W]e believe that the task of enhancing the moral capabilities of autonomous software agents will force scientists and engineers to break down moral decision-making into its component parts, recognize what kinds of decisions can and cannot be codified and managed by essentially mechanical systems, and learn how to design cognitive and affective systems capable of managing ambiguity and conflicting perspectives.

Zunächst stellt sich die Frage, warum „break down moral decision-making" eine Aufgabe von (ausschließlich) Programmierer∗innen und Ingenieur∗innen sein sollte. Da bereits die Relevanz der Transdisziplinarität für die Maschinenethik deutlich gemacht wurde und da es sich um eine Aufgabe handelt, die verschiedener Kompetenzen bedarf, scheint es sinnvoll zu sein, die Umsetzung von Normen, deren ethische Reflexion sowie die praktische technische Umsetzung als fachübergreifendes Anliegen zu verstehen. Des Weiteren geht aus dem Zitat hervor, dass Wendell Wallach und Colin Allen *moral decision-making* als Voraussetzung für *moral capabilities* verstehen. Ob sie dabei *moral capabilities* mit *moral decision-making* gleichsetzen oder dieses als Voraussetzung für jene sehen, wird nicht klar. Um moralisch handeln zu können, bedarf es ihnen zufolge jedenfalls bestimmter kognitiver Fähigkeiten, wie Reflexion ethischer Urteile (und gegebenenfalls noch anderer Fähigkeiten, die zu *moral decision-making* gezählt werden, s. u.). Diese Annahme unterscheidet sich grundlegend von der bisher dargelegten Definition moralischer Techniksysteme.

5.1 Grundlagen für ethische Theorien in technischen Systemen: *top-down-* und *bottom-up-*Ansätze

Moral decision-making kann als zentrales Element bei der theoretischen wie praktischen Entwicklung genuin ethischer technischer Systeme verstanden werden. Dabei

ist die Definition von *moral decision-making* bereits in Bezug auf Menschen keine leichte Aufgabe (vgl. Seng 2020), da es hierfür eine grundlegende Frage zu klären gilt, die viele Aspekte beinhaltet: Welche Fähigkeiten und Eigenschaften (kognitiver wie psychischer Art) sind notwendig für unabhängige und situativ flexible Entscheidungen, die zuverlässig moralisch gute Handlungen hervorbringen?

Wallach und Allen zufolge wurde diese Frage (bis dato) noch nie zufriedenstellend beantwortet: „This project will demand that human moral decision-making be analyzed to a degree of specificity as yet unknown“ (Wallach und Allen 2009: 71). Um möglichen Formen der Umsetzung zuverlässigen, moralisch richtigen Handelns näher zu kommen, reflektieren sie die Möglichkeiten von *top-down*- und *bottom-up*-Ansätzen. Dahinter verbirgt sich gemäß den Autoren die Idee, dass es zwei grundlegende Strategien gibt, wie man Computerprogramme so gestalten kann, dass sie die Anforderungen von *artificial moral agents* (AMA) führen. Eine erste Annäherung kann mit Catrin Misselhorn (2019: 96) gegeben werden:

> Top-Down-Ansätze charakterisiert eine schrittweise Herangehensweise an die Softwareentwicklung von allgemeinen, umfassenden Strukturen zu immer spezielleren Details. Der natürliche Verbündete der Ansätze sind Prinzipienethiken.

Ob Wallach und Allen mit ihrer Konzeption von AMAs eher moralische oder ethische Akteure im Verständnis der vorliegenden Arbeit meinen, ob es also um zuverlässigerweise moralisch gutes Handeln oder auch um die eigene, von Menschen unabhängige, Entwicklung neuer Handlungsprinzipien geht, kann diskutiert werden. Mit Sicherheit können Ihre Überlegungen zu *top-down* und *bottom-up*-Ansätzen jedoch genutzt werden, um die Entwicklung von moralischen zu ethischen Maschinen theoretisch zu begleiten.

Den Unterschied zwischen *top-down*- und *bottom-up*-Ansätzen stellt Catrin Misselhorn (vgl. ebd.) als Unterschied zwischen ethischem Generalismus und Partikularismus dar. Beispiele für *top-down*-Ethiken sind demnach „Utilitarismus“ (ebd.: 97 ff.), „Kantische Ethik“ (ebd.: 101 ff.) und „Asimov'sche Gesetze“ (ebd.: 108 ff.) beziehungsweise Tugenden für *bottom-up*-Ansätze (ebd.: 114 ff.). Die Unterscheidung der beiden Vorgehensweisen scheint aber nicht ganz so eindeutig zu sein, wie weiter unten deutlich werden wird. Folgende Fragen stehen bei der Verankerung moralischer Werte inform der Implementierung ethischer Theorien im Mittelpunkt:

1. Wie können technische Systeme so konstruiert werden, dass sie auch in komplexen und unvorhersehbaren Situationen zu einem gewünschten, das heißt moralisch guten Ergebnis kommen?

2. Welcher ethische Ansatz oder welche moralischen Werte eignen sich am besten für die Umsetzung (Realisierung oder Implementierung) in technischen

Systemen, um verlässlicherweise einen moralisch guten Output zu generieren?

3. Wie kann ein ethischer Ansatz oder wie können ethische Werte in technischen Systemen konkret umgesetzt werden?

Die erste Frage ist die, die über dem ganzen Vorhaben der Entwicklung moralischer (und ethischer) Maschinen schwebt. Die zweite und die dritte Frage behandeln im Prinzip denselben Gegenstand, nämlich eine fruchtbare, sinnvolle Verquickung von Ethik und Technik, wobei die zweite Frage primär aus ethischer Perspektive gestellt ist und die dritte aus primär technologischer. Die drei Fragen markieren damit auch den Übergang von moralisch gut agierenden technischen Systemen, wie sie heute durchaus denkbar und technisch realistisch sind – wenngleich solche Systeme nicht fehlerfrei sind und Risiken bergen, wie oben angesprochen wurde –, zu genuin ethischen Techniksystemen, die über die Einhaltung moralisch Regeln in der Lage sein sollten, diese Regeln selbst zu reflektieren und gegebenenfalls flexibel an unterschiedliche Situationen anzupassen. Wendell Wallach und Colin Allen wollen ihre Ansätze, *top-down* und *bottom-up*, als eine Mischung aus ingenieurwissenschaftlichen und ethischen Ansätzen verstanden wissen.

***Top-down*:**

Unter *top-down* verstehen die Autoren, eine Aufgabe oder ein Problem in verschiedene Unteraufgaben auf verschiedenen hierarchischen Ebenen zu unterteilen. Einzelteile könnten dann in Modulen zusammengefasst werden, die wiederum in *subtasks* aufgeteilt werden, um so eine große Aufgabe auf leichter erfüllbare Häppchen aufzuteilen (vgl. Wallach und Allen 2009: 79). *Top-down* nennen sie dieses Verfahren in Anlehnung an ingenieurswissenschaftliche Prozesse deshalb, da eine bestimmte ethische Theorie oder ethische Prinzipien *von oben*, also von den Ingenieur∗innen oder Programmierer∗innen, vorgegeben werden.

> In our merged sense, a top-down approach to the design of AMAs is any approach that takes a specified ethical theory and analyzes its computational requirements to guide the design of algorithms and subsystems capable of implementing that theory.
> (Ebd.: 79 f.)

Als mögliche ethische Theorie für *top-down*-Prozesse wird unter anderem der Utilitarismus diskutiert.

> In other words, a top-down approach takes an ethical theory, say, utilitarianism, analyzes the informational and procedural requirements necessary to implement

> this theory in a computer system, and applies that analysis to the design of subsystems and the way they relate to each other in order to implement the theory. (Wallach und Allen 2009: 79 f.)

Was zunächst einfach klingt, stellt sich in der Praxis jedoch als schwierig heraus, denn welche sind die „informational and procedural requirements necessary to implement“? Dies zu klären, ist die zentrale philosophische Aufgabe, die unter anderem in Abschnitt 5.2 geleistet werden soll – wenngleich die Diskussion zentraler Begriffe dort selbstverständlich nicht als abschließend, sondern eher als Annäherung an genuin ethische Maschinen verstanden werden kann.

In den *Ethik-Leitlinien für eine vertrauenswürdige KI* der Europäischen Kommission (Europäische Kommission 2019: 17 f.) werden beispielsweise ethische Begriffe genannt („Transparenz“, „Rechenschaftspflicht“, „Gesellschaftliches und ökologisches Wohlergehen“), deren Umsetzung mittels eines (beispielsweise utilitaristischen) *top-down*-Ansatzes im Detail definiert werden könnte. Mit Sicherheit wäre es hier nicht mit dem Einsatz einzelner moralischer Prinzipien getan (vgl. Misselhorn 2019: 96), da jene Konzepte Verhaltensanpassungen auf der Grundlage vieler moralischer Prinzipien erforderlich machen können. Denn wie beispielsweise *Transparenz* umgesetzt werden kann und muss, hängt von der Beschaffenheit eines jeweiligen technischen Systems ebenso wie von den jeweiligen Kontexten und anderen Faktoren ab.

Das Grundproblem bei Regeln und Prinzipien ist also, dass sie entweder sehr konkret formuliert werden und dann die flexible Anpassungsfähigkeit in unterschiedlichen Situation eingeschränkt wird – oder aber sie sind offen formuliert und laufen somit Gefahr, beliebig und wenig aussagekräftig zu sein. Als Beispiel können hier die zwar fiktiven, gleichwohl immer wieder zitierten, Asimov'schen *Robotergesetze* herangezogen werden, die viel Spielraum bei der Ausgestaltung bieten und zu Konfliktfällen einzelner Regeln untereinander führen können (s. S. 57 f.). Gemäß Wallach und Allen (2009: 79) zählen neben philosophischen Prinzipienethiken auch „religious ideals“ und „moral codes to culturally endorsed values“ zu *top-down*-Ansätzen, „but many of the same values are evident in differing ethical systems“. Fraglich ist dabei, wie hilfreich solche Ansätze sein können, wenn Regeln auch in Form von *Menschenwürde* oder dem *Recht auf Leben* formuliert werden. Dies ist ein großer Streitpunkt unter anderem bei der Entwicklung (teil-)automatisierter Fahrzeuge, die mitunter auf der Basis von Dilemma-Situationen geführt wird, welche jedoch, wie oben gezeigt, nur bedingt zielführend sind. Gleichwohl stellt sich die Frage, welche Grenzen technischen Systemen gegeben werden können (und in welcher technischen Umsetzung), die potenziell Menschenleben gefährden.

***Bottom-up*:**

Bei *bottom-up*-Ansätzen liegt der Fokus für Wallach und Allen auf konkreten Kontexten, in denen technische Systeme moralische Vorgehensweisen schrittweise lernen (vgl. auch Misselhorn 2019: 114) – im Gegensatz zum Versuch des Aufstellens von a priori-Regeln beim vorherigen Ansatz.

> In bottom-up approaches to machine morality, the emphasis is placed on creating an environment where an agent explores courses of action and learns and is rewarded for behavior that is morally praiseworthy. There are various models for bottom-up acquisition of moral capabilities. Childhood development provides one model. Evolution provides another bottom-up model for the adaption, mutation, and selection of those agents best able to meet some criteria for fitness. Unlike top-down ethical theories, which define what is and is not moral, in bottom-up approaches any ethical principles must be discovered or constructed.
>
> (Wallach und Allen 2009: 80)

Der Hauptunterschied zu *bottom-up*-Ansätzen ist gemäß Wallach und Allen der, dass bei *top-down*-Ansätzen konkrete Regeln vorgegeben werden und bei *bottom-up*-Ansätzen bestimmte Vorgehensweisen in konkreten Situationen. Dabei betonen die Autoren, dass auch selbstlernenden Systemen in *bottom-up*-Ansätzen eine gewisse *performance measure* vorgegeben werden muss, damit ein System weiß, in welchem Operationsrahmen es sich bewegen kann.

> In bottom-up engineering, tasks can also be specified theoretically using some sort of performance measure (e.g., winning chess games, passing the Turing test, walking across a room without stumbling, etc.). Various trial-and-error techniques are available to engineers for progressively tuning the performance of systems so that they approach or surpass the performance criteria.
>
> (Ebd.)

Die systematische Trennung, die Wallach und Allen hier postulieren, ist dabei nur im Hinblick auf die *Art und Weise* haltbar, *wie* technische Systeme moralische Prinzipien oder technische Ziele umsetzen. Gemeinsam ist beiden jedoch – und das ist der ethisch interessante Punkt –, dass diese Ziele in beiden Ansätzen vorab bekannt sein, definiert werden und in den Systemen verankert werden müssen. Bei beiden Ansätzen muss ein ethischer Rahmen definiert werden, innerhalb dessen technische Systeme agieren können; technisch formuliert muss ein Ziel vorgegeben sein, das erreicht werden soll; Wallach und Allen oben nennen das Gewinnen eines Schachspiels, das Bestehen des Turing-Tests oder durch einen Raum zu laufen, ohne zu stolpern. Es muss für ein technisches System also sowohl in *bottom-up*- als auch in *top-down*-Ansätzen bestimmte Zielvorgaben geben – unabhängig davon, ob diese *von oben* vorgegeben sind oder ob sie schrittweise erlernt werden.

Die ethische Aufgabe der Normdefinierung (bei Floridi und Sanders war oben die Rede von *moral threshold*, s. S. 78) muss also vorab geleistet werden. Anstelle von *theory* wählen Wallach und Allen bei *bottom-up*-Ansätzen auch die Ausdrucksweise *code*, die jedoch eher die disziplinär unterschiedlichen Terminologien belegt – in der Sache aber ist kein Unterschied zu den in *top-down*-Ansätzen als erforderlich erklärten *Theorien* erkennbar.

> What are the variables, for example, that will enable a threshold to be defined for producing behavior in compliance with items in this code such as ‚Contribute to society and human well-being' and ‚Be honest and trustworthy'?
> (Wallach und Allen 2009: 202)

Contribute to society and human well-being und *Be honest and trustworthy* sind allgemeingültige, moralische Ziele. Das oben genannte Beispiel der Kinderentwicklung und bedingt vergleichbares *maschinelles Lernen* legen außerdem die Notwendigkeit konkreter moralischer Ziele für die Anwendung beider Ansätze nahe. Und dennoch halten die Autoren das Vorliegen einer bestimmten Theorie für *bottom-up*-Ansätze nicht für notwendig:

> High levels of performance on many tasks can be achieved even though the engineer lacks a theory of the best way to decompose the task into subtasks. An analysis of the system after it has determined how to perform a task can sometimes yield a theory or specification of the relevant subtasks, but the results of such analyses can also be quite surprising and typically do not correspond to the kind of decomposition suggested by a priori theorizing.
> (Ebd.: 80)

Dies mag für verhältnismäßig unkritische technische Systeme wie Staubsaugerroboter gelten, bei denen es reicht, die Ziele allein auf technischer Ebene zu definieren. Jüngere Entwicklungen in den Bereichen *maschinelles Lernen* und *künstliche Intelligenz* belegen jedoch die Notwendigkeit einer vorherigen Risikoanalyse und erwünschter moralischer Bedingungen, die sinnvollerweise nur unter Einbezug philosophisch-ethischer Analysen erbracht werden kann. Die Aufgabe, Moral in Maschinen zu implementieren, muss also als nicht nur ingenieurwissenschaftliche oder informatische, sondern auch als ethische Aufgabe betrachtet werden. Auch hybride Modelle, also die Kombination von *top-down*- und *bottom-up*-Ansätzen werden diskutiert (vgl. Allen, Smit und Wallach 2015; van Rysewyk und Pontier 2015b).

Bei der Frage, wie Ethik in Maschinen umgesetzt werden kann, geht es dabei nicht nur um die Form der Übertragung ethischer Theorien auf technische Systeme, sondern auch mögliche Reflexionsfähigkeiten moralischer Werte und Handlungsbegründungen durch die Technik selbst. Im Folgenden werden zunächst einige, für

ethische Entscheidungsfindungen, zentrale Begriffe ihre Relevanz für mögliche, genuin ethische Maschinen diskutiert. Ich beginne dabei mit der Umsetzung einzelner Begriffe und widme mich anschließend möglichen Anwendung ganzer Theorien in Maschinen. Abschließend wird auf der Grundlage der Diskussionen die Frage beantwortet werden, welche Voraussetzung für selbständige Denk- und Handlungsweisen (im weitesten Sinn) auch für ethische Maschinen für notwendig und welche begründeterweise als kontingent erachtet werden können.

5.2 Annäherung an Voraussetzungen für genuin ethische Techniksysteme

Wenn technische Systeme nicht nur nach von außen vorgegebenen Regeln moralisch handeln, sondern auch selbst eigene Regeln entwerfen, diese reflektieren und nach ihnen handeln / agieren könnten, dann handelte es sich meinem Verständnis nach um *genuin ethische* Maschinen.[90] In der Einleitung habe ich drei Varianten von Maschinenethik vorgestellt, in den Formen einer Ethik von Menschen für Menschen im Umgang mit technischen Systemen, einer Maschinenethik von Menschen für Maschinen sowie einer Ethik von Maschinen für Menschen oder Maschinen (s. S. 30). Die folgenden Darstellungen werden sich naturgemäß zunächst auf menschliche Werte und Theorien gründen, also von einem möglichen technischen *simulacrum* (s. Powers auf S. 11) ausgehen. Sollte sich herausstellen, dass eine *breitere Form von Ethik* (wie sie Rath, Karmasin und Krotz (2019: 3 f.) im Zitat auf S. 32 zunächst in den Raum stellen, dann aber verwerfen, s. u.) notwendig ist, hätte das nicht nur Folgen für das metaethische Verständnis von Maschinenethik, sondern auch für den moralischen Status von Technik und Menschen.

Für genuin ethische Techniksysteme gemäß der obigen Definition wird häufig *ethical deliberation* (Powers 2011: 464) beziehungsweise *moral decision-making* (Wallach und Allen 2009: 71) oder Ähnliches in anderen Formulierungen vorausgesetzt. Welche Voraussetzungen diese Konzepte beinhalten und ob es sich dabei vor allem um kognitiven Fähigkeiten, wie *Überlegungen* oder *Entscheidungsfindungen* beziehungsweise im weitesten Sinn *selbstständigen* Denken, handelt oder ob noch weitere Faktoren dazu zählen, bleibt oft unklar oder wird von einigen Autor∗innen nur vage erläutert.

[90] Man kann argumentieren, dass die statistischen Berechnungen des maschinellen Lernens bereits Formen von *Reflexionen* darstellen. Inwieweit dies plausibel ist, wird sich im Lauf der Argumentation in diesem Kapitel herausstellen.

„[E]thics is more than just rules. Ethics comes as a package deal“ (Krzanowski et al. 2016: 1). Dies gilt auch für Ethik in technischen Systemen – zumindest gemäß der Meinung vieler Autor∗innen, die sich mit den Voraussetzungen für genuin ethische Maschinen (häufig in anderer Terminologie) auseinandersetzen. Deren mangelt es in der Definition Krzanowskis et al. (ebd.) nicht:

> Ethics (implicitly or explicitly) requires free will, consciousness, a concept of good and wrong, an understanding of responsibility [...] and some comprehension of reality around us. A lot of deep metaphysics is involved in the concept of ethics, such as free will, good life, or the individual. Dispensing with metaphysics leaves ethical statements groundless. [91]

In den folgenden Abschnitten werde ich zunächst einige Definitionsansätze mentaler Qualitäten[92] wie Bewusstsein, Reflexionsfähigkeit, und anderer Werte, die als mögliche Voraussetzungen für ethische Reflexionen angesehen werden können, wie Freiheit, vorstellen und anschließend mögliche Ansätze für die Umsetzung ganzer ethischer Theorien in technischen Systemen darlegen. Dabei ist zu bemerken, dass die Auswahl der im folgenden als mögliche Voraussetzungen für *ethical decision-making* diskutierten Begriffe keinesfalls als These aufgefasst werden soll, dass es sich hierbei um die einzigen plausiblen Voraussetzungen handle. Darüber hinaus können die einzelnen Begriffe im Folgenden aufgrund ihrer Komplexität jeweils nur ansatzweise aufgegriffen und als mögliche Voraussetzungen für *ethical decision-making*, zunächst bei Menschen und dann bei Techniksystemen, diskutiert werden. Es handelt sich also keinesfalls um eine allumfassende, abschließende Analyse, sondern um eine Synopse möglicher Bedingungen.

Bei James Moors oben vorgestellter Einteilung moralischer Maschinen in verschiedene Kategorien (s. S. 72 f.) habe ich die letzte vorerst ausgelassen, da diese gemäß der Kategorisierung der vorliegenden Arbeit in den Bereich ethischer Maschinen fällt: *full ethical agents*. „An average adult human is a full ethical agent“,

[91] An dieser Stelle könnte natürlich der berechtigte Einwand hervor gebracht werden, dass hier mit *Ethik* auch *Moral* gemeint sein könnte. Ich vermute jedoch, dass die Autoren über *Ethik* im oben definierten Sinn sprechen, da auch von *Konzepten* eines *guten Lebens* u. a. die Rede ist, was auf einen theoretisches Verständnis hindeutet.

[92] In Texten zur Philosophie des Geistes ist häufig die Rede von *mentalen Zuständen* oder *mental states* (vgl. u. v. a. Beckermann 2001). Diese Ausdrucksweise halte ich jedoch für irreführend, da, unabhängig vom Status der mentalen Qualitäten (also ob sie beispielsweise mittels Gehirnfunktionen erklärt werden oder anderweitig), konstatiert werden kann, dass es sich in den meisten Fällen um aktive Prozesse handelt und nicht um statische Zustände. Bewusstsein kann beispielsweise neurophysiologisch als Prozess erklärt werden (vgl. u. a. Marchetti 2018; Pepperell 2018); aber auch bei non-reduktionistischen Erklärungsansätzen ist es plausibel, Bewusstsein als bewegliche Eigenschaft oder Fähigkeit (in einem weiten Verständnis, s. Fußnote 15) anzunehmen.

schreibt Moor (2006: 20), und weiter: „We typically regard humans as having consciousness, intentionality, and free will." Diese Aussagen können, wie bereits festgehalten, nicht auf alle (erwachsenen) Menschen pauschal angewandt werden, da es viele erwachsene Menschen gibt, die sich nicht auf (gemäß dem Konsens einer Gesellschaft) moralische Art und Weise verhalten beziehungsweise nicht allen freier Wille und Bewusstsein unterstellt werden kann, wenn man zum Beispiel an Menschen mit kognitiven Einschränkungen denkt. Die Thesen werden von Moor nicht weiter begründet, zumal auch nicht geklärt wird, was genau unter den sehr mehrdeutigen Begriffen *consciousness, intentionality* und *free will* verstanden wird. Somit fokussiere ich mich auf die Aussagen, die Moor über kognitive Fähigkeiten ethischer Akteure macht: Ein *full ethical agent* ist für ihn eine Person, die ethische Bewertungen vornehmen („can make explicit ethical judgments") und diese auch vernünftigerweise begründen kann („generally is competent to reasonably justify them", ebd.). Allerdings äußert Moor auch die, ihm zufolge, weitverbreitete Ansicht, dass technische Systeme niemals den Status von *full ethical agents* erreichen können.

> Many believe a bright line exists between the senses of machine ethics discussed so far and a full ethical agent. For them, a machine can't cross this line. The bright line marks a crucial ontological difference between humans and whatever machines might be in the future.
>
> (Ebd.)

Moor selbst positioniert sich in dieser Frage nicht. Dass es zwischen Menschen und Maschinen einen prinzipiellen, ontologischen Unterschied gibt, scheint unbestritten zu sein – jedoch kann jene These unbegründeterweise, insbesondere in anderen Kontexten, zu einer Zurückweisung unter dem Vorwurf des Speziesismus führen (vgl. z. B. Singer 2009). In Kapitel 3 habe ich bereits auf die argumentatorischen Schwierigkeiten einer unrealistischen Idealisierung der menschlichen Seite als künstliches Kontrastmittel zur Abgrenzung von (potenziell) nicht-menschlichen Konkurrent∗innen hingewiesen, welches auch innerhalb des Moralkreises von Menschen zu Thesen führt, die diskussionswürdig sind. Insbesondere bei ethischen Ansätzen, die die Voraussetzungen für moralisches Handeln allzu hoch ansetzen, ist fraglich, ob solche Definitionen a) begrifflich angemessen und b) ethisch sinnvoll sind.[93]

[93] In meiner 2014 abgeschlossenen Masterarbeit habe ich beispielsweise untersucht, welche Konsequenzen die zeitgenössische Vertretung eines aristotelisch geprägten Tugendbegriffs, wie einige zeitgenössische Autor∗innen ihn vertreten (vgl. u. a. Annas 2011; Russell 2013), für die Definition von Inklusion und Teilhabe von Menschen in Gesellschaften haben kann.

Da viele Ansätze mit unterschiedlichen (unter anderem kognitiven) Fähigkeiten als Voraussetzungen für ethische Reflexion operieren, werde ich einige davon im Folgenden skizzieren und im Hinblick auf folgende Fragen kritisch beleuchten:

1. Sind die angesprochenen Qualitäten, Fähigkeiten und Werte notwendige Voraussetzungen für ethische Reflexions- und Begründungsfähigkeit beziehungsweise Theoriekonzeption?

2. Können beziehungsweise müssen diese Qualitäten, Fähigkeiten und Werte also auch für ethische Maschinen angenommen werden? (Oder gibt es gute Gründe, dies nicht zu tun?)

3. Wie kann eine Umsetzung der Qualitäten, Fähigkeiten und Werte in Maschinen und anderen technischen Systemen konkret aussehen?

Diese Fragen knüpfen an die bezüglich moralischer Techniksysteme auf Seite 90 f. an. In diesem Kapitel werden zunächst die Voraussetzungen einzelner Begriffe und Theorien im Fokus stehen, der dritten Frage widme ich mich dann in Kapitel 6. Neben eindeutig kognitiven Fähigkeiten beziehungsweise mentalen Qualitäten, wie Intentionalität und Bewusstsein, werden für ethische Maschinen dabei auch Werte ins Feld geführt, für die selbst wiederum kognitive Fähigkeiten als Voraussetzung diskutiert werden, wie zum Beispiel Freiheit und Autonomie. Die Voraussetzungen und die gegenseitige Bedingung lassen eine sinnvolle Trennung der Begriffe außerdem fast unmöglich erscheinen, weshalb auch immer wieder Querverweise gezogen werden.

Gemäß der oben aufgeworfenen Fragen, werde ich innerhalb der Abschnitte zu den einzelnen Begriffen jeweils explizit auf den Begriff als *notwendige Voraussetzung für ethische Reflexionen*[94] und *Umsetzung in technischen Systemen* eingehen; die Reihenfolge kann dabei variieren. In Abschnitt 5.3 werde ich die Ergebnisse gesammelt dargestellt.

5.2.1 Freiheit

Freiheit ist ein häufig erwähnter Begriff in der Technikphilosophie, insbesondere in Bezug auf (teil-)automatisierte Systeme. In der philosophischen Handlungstheorie wird Willensfreiheit von Handlungsfreiheit unterschieden.

[94] Aufgrund der häufigen Wiederholungen reduziere ich an dieser Stelle den Begriff *ethisch*, den ich zu Beginn der Arbeit auch adjektivisch als *ethisch sein* umschrieben und im Folgenden immer wieder mit konkreten Inhalten versehen habe (s. z. B. Abschnitt 4.6), was darunter zu verstehen ist, auf *ethische Reflexionen*. Damit ist natürlich auch die eigenständige, unabhängige Normbildung usw. gemeint, was für eine einfachere Lesbarkeit nicht immer wiederholt wird.

> Letztere wird als die Freiheit bestimmt, das zu tun oder zu lassen, was man will. Handlungsfreiheit besitzt man, wenn man nicht durch äußeren Zwang daran gehindert wird, seinen Willen in die Tat umzusetzen. Die politischen oder bürgerlichen Freiheiten wie Pressefreiheit und Reisefreiheit sind Unterarten der Handlungsfreiheit. Willensfreiheit ist etwas anderes. Während unsere Handlungsfreiheit durch die jeweiligen tatsächlichen Optionen begrenzt ist, scheint dies für die Willensfreiheit nicht zu gelten. Wer eingesperrt ist, kann viele Dinge nicht tun, die er gern tun würde, aber er verliert dadurch nicht das Vermögen, sich seinen Willen zu bilden. Ebenso wenig verliert er dieses Vermögen, wenn er sich über das Ausmaß seiner Handlungsoptionen täuscht. Jemand könnte, so ein Beispiel von John Locke, in seinem Zimmer sitzen und sich dazu entschließen, den Raum durch die Tür zu verlassen. Dass die Tür ohne sein Wissen verschlossen wurde, beeinträchtigt seine Bewegungsfreiheit, nicht aber seine Willensfreiheit.
>
> (Keil 2009: 24)

Diese Unterteilung in Willens- und Handlungsfreiheit wird im Zusammenhang mit technischen Systemen häufig nicht explizit erwähnt, manchmal aber implizit zugrunde gelegt. Die Differenzierung ist dabei nicht nur wichtig für begriffliche Exaktheit, sondern auch, wenn es an konkrete Umsetzungsmöglichkeiten ethischer Maschinen geht.

Freiheit als notwendige Voraussetzung für ethische Reflexion?

Gemäß Wallach und Allen (2009: 58) ist Freiheit (*freedom*) ein Attribut, das häufig für *moral agency* vorausgesetzt wird. Man brauche Freiheit, um Intentionen (s. u.) realisieren zu können (vgl. ebd.: 59); ihnen geht es also offenbar primär um Handlungsfreiheit. Freiheit setze außerdem einen Geist (*mind*) voraus und bestehe in der „ability to choose and to have a relation to oneself and one's inclinations, needs, and desires" (ebd.). Diese Thesen belegen, dass es Wallach und Allen nicht um *moralische Akteure* in der Definition der vorliegenden Arbeit geht, sondern um technische Systeme mit komplexeren Fähigkeiten. Dabei gehen die Autoren nicht genauer auf die definierten Voraussetzungen, wie *mind*, ein und erklären auch nicht, wie die *relation to oneself* [...] – sowohl in technischen Systemen als auch in menschlichen Akteuren – genau aussehen kann. Während bei Wallach und Allen zunächst ein Verständnis von Freiheit als Handlungsfreiheit zu vermuten war, suggeriert die Formulierung *ability to choose* ein Verständnis von Freiheit als Willensfreiheit, wobei es sich hierbei um eine Vermutung handelt. Wie ambitioniert Wallach und Allens Freiheitsbegriff ist, zeigt sich daran, dass sie Freiheit allenfalls für Menschen möglich halten; und selbst für diese nur eingeschränkt.

> Technologies ‚in themselves' cannot be free, but neither can human beings. On the one hand, technologies help to *constitute* freedom by providing the material environment in which human existence takes place and takes its form. And on the other hand, technologies can form associations with human beings, which become

> the places where freedom is to be located. Technological mediations create the space for moral decision-making. Just like intentionality, freedom is a hybrid affair, most often located in associations of humans and artifacts.
>
> (Wallach und Allen 2009: 60 f.)

Diese Definition schließt nicht aus, dass sich das Subjekt-Objekt-Verhältnis in den *Verbindungen* (*associations*) von Menschen und Technik auch umkehren kann, sodass fortgeschrittene technische Systeme als Subjekte gesehen werden können. Dies würde der dritten Variante von Maschinenethik auf Seite 30 entsprechen und eine Form möglicher technischer Freiheit darstellen.

Ethische Reflexionsfähigkeit bedarf vor allen Dingen eines gewissen Maßes an Willensfreiheit, also der Fähigkeit, Thesen bewerten und gegeneinander abwägen zu können auf der Grundlage bestimmter Wertkonzepte und Theorien. Im Hinblick auf die technische Umsetzungsfähigkeit muss dabei *Wille* nicht im klassisch kantischen Sinn (s. u.) verstanden werden, sondern kann schlicht die Freiheit zur flexiblen Reflexion ethischer Theorien bedeuten. Auch ein gewisses Maß an Handlungsfreiheit ist vonnöten, sofern ein Mensch oder technisches System die theoretischen Reflexionen auch praktisch umsetzen können soll.

Freiheit in technischen Systemen

Stefan Ullrich (2019b: 247) legt Maschinen und insbesondere als *intelligent* bezeichneten Algorithmen, die es bereits gibt, „automatisierte Entscheidungsprozesse“ zugrunde, wobei er *automatisch* nicht im Sinn der altgriechischen Wortherkunft versteht – also *aútós* und *tó mátos* als Selbst-Suche, -Forschen (vgl. Pape 1954) –, sondern im Gegenteil: auf zwangsläufigen Prozessen beruhend. In diesem Sinne automatische Prozesse grenzt er dezidiert ab von menschlicher Willens- und Handlungsfreiheit, die ihm zufolge nur Menschen haben können, die die Freiheit haben, eine Handlung auch unterlassen zu können. Entsprechende Zuschreibungen menschlicher Fähigkeiten an technische Systeme hält Ullrich daher für irreführend.

> Informationstechnischen Systemen wird oft zugeschrieben, Entscheidungen treffen zu können, dabei findet eine Berechnung statt. Menschen folgen moralischen Gesetzen – oder eben nicht. Ein informationstechnisches System (IT-System) muss den einprogrammierten Gesetzen folgen, so wie der Mensch den zahlreichen Naturgesetzen unterworfen ist.
>
> (Ullrich 2019b: 243)

Ullrich hat allerdings explizit technische Systeme im Fokus, die es derzeit bereits gibt, also moralische Akteure. Für ethische Maschinen sind Berechnungen denkbar, die eine eigene *Gesetzgebung* und damit spontane, nicht vorhersehbare Aktionsänderungen – auf der Basis ethisch begründeter Normen – zulassen würden.

James Moor hält, wie oben beschrieben, nur Menschen für fähig, *full ethical agents* zu sein und damit über Bewusstsein (s. Abschnitt 5.2.4) und Willensfreiheit zu verfügen. Moor hält es zwar nicht für ausgeschlossen, sieht es aber auch nicht in greifbarer Nähe, dass auch Maschinen irgendwann einmal *full ethical agents* werden können. Aus diesem Grund schlägt er vor, dass das Konzept von *explicit ethical agents* in einer limitierten Form als realistische Zielvorgabe bei Programmierungen dienen könne.

> We won't resolve the question of whether machines can become full ethical agents by philosophical argument or empirical research in the near future. We should therefore focus on developing limited explicit ethical agents. Although they would fall short of being full ethical agents, they could help prevent unethical outcomes. (Moor 2006: 21)

Moor hält darüber hinaus die Entwicklung technischer *full ethical agents* nicht etwa in erster Linie aus dem Grund für wenig relevant, da diese prinzipiell technisch schwer zu realisieren seien, sondern es scheitere zunächst an einem einheitlichen Verständnis von angemessenen ethischen Theorien (*proper ethical theory*) und dem Begriff des Lernens (ebd.). Was *Lernen* angeht, so zeigen derzeitige Entwicklungen, dass es keines einheitlichen Begriffes bedarf, um Formen maschinellen Lernens umzusetzen. Außerdem ist unklar, inwiefern Moor das Vorliegen einer *proper ethical theory* auch für Menschen für notwendig hält, die seiner Ansicht nach (s. Abschnitt 4.4) ja alle, durchschnittlich gesehen, *full ethical agents* sind. Demnach müsste für menschliche Handlungsweisen als *full ethical agents* auch eine *proper ethical theory* vorliegen, was in vielen Hinsichten schwierig anzunehmen ist. Zum einen ist die Frage, von welchem Menschenkreis gesprochen wird und ob es sich somit, auch kulturell bedingt, um eine Handlungstheorie handeln kann, oder nicht vielmehr um mehrere. Zum anderen stellt sich die Frage, wie angemessene Theorien aussehen könnten, die die ethische Handlungsfähigkeit von Menschen notwendigerweise bedingen. Moor scheint also unterschiedliche Maßstäbe für *full ethical agents* bei Maschinen und Menschen anzulegen. Für Computerprogramme als *full ethical agents* definiert er ferner auch *common sense* und *world knowledge* als notwendige Voraussetzungen. Grundlegendere Probleme bei der Entwicklung von *full ethical agents* sind laut Moor daher weniger ethischer als vielmehr epistemologischer Art. Inwiefern sich dieselben epistemologischen Probleme auch bei Menschen stellen (s. u.), lässt er offen.

Angesichts jüngster Entwicklungen im Bereich des *deep learnings*, wodurch eine, beispielsweise im medizinischen Bereich, hilfreiche Mustererkennung auf einer wesentlich größeren Datenbasis erzeugt werden kann, als es einem∗r menschlichen Ärzt∗in jemals zu erfassen möglich wäre, kann diskutiert werden, inwiefern diese als Formen von *common sense* und *world knowledge* gelten können (vgl. hierzu z. B.

Rieder 2021; Richter 2016). Die *epistemischen Fähigkeiten*[95] technischer Systeme haben sich darüber hinaus mit Systemen wie (teil-)automatisierten Fahrzeugen bereits wesentlich verbessert. Thimm und Bächle (2019: 80) schreiben dazu:

> Durch die sich rasant verbessernden Künstliche Intelligenz-basierten Rechenverfahren wie ‚Deep Learning' besteht nunmehr auch die Möglichkeit, Wissen zu generieren, das sich weiter von einer menschlichen Kontrolle und Einflussnahme emanzipiert [...]. Diese Form der Adaptions- und Lernfähigkeit wird gegenwärtig im Hinblick auf ihre erhebliche ethische Tragweite diskutiert. Wenn es nicht mehr allein die Programmierer*innen sind, die wünschenswerte Handlungsskripte in Algorithmen einschreiben, sondern diese vielmehr in der Lage sind, eigenständige Aktions- und Wissensstrukturen zu entwickeln, werden die ‚autonomen moralischen Agenten' – so die Befürchtung – zu einer neuen, nicht-menschlichen Stimme in der Auseinandersetzung über das ‚richtige und gute Handeln'.[96]

Entsprechend der Moor'schen Argumentationslinie könnten Kritiker∗innen wiederum einwenden, dass Lernen durch Erfahrung eben nicht nur in der Anhäufung von Daten besteht, sondern auch in irgendeiner Form von Einordnung und manche halten dafür auch *Intuition* für notwendig.[97] Eine solche Form von Gewichtung in technischen Systemen künstlich zu erzeugen (zumal, wenn sie auch von Maschinen selbst generiert werden und nicht nur vorgegeben werden soll), könnte sich als eine große Herausforderung darstellen. Auf der anderen Seite sind viele Prozesse der Entscheidungsfindung (wie Bewusstsein und Intuition) noch nicht einmal bei Menschen hinreichend erforscht, sodass es gewagt ist, diese pauschal als harte Konkurrenz zu maschineller „Beschränktheit" zu inszenieren (Ropohl 1991: 157).

Exkurs: Freiheit bei Menschen

Bei der Frage, was technische Systeme können oder nicht, steht zunächst oft die Frage im Raum, was Menschen können. Wie bereits an einigen Stellen gezeigt wurde, werden hier immer wieder Annahmen über bestimmte Fähigkeiten und Eigenschaften gemacht, die bei Menschen als notwendige Bedingungen für moralisches Handeln oder ethische Reflexionsfähigkeit vorausgesetzt werden, und anhand derer dann auch die entsprechenden Möglichkeiten bei Maschinen bemessen werden. Es ist also nicht unerheblich, sich bei der Frage, was Maschinen eigentlich können, auf der anderen Seite auch die menschlichen Fähigkeiten oder Eigenschaften und Bedingungen für Handeln und Reflexion genauer anzusehen, sofern sie als

[95] Man möge hier darunter Kontexterkennung mittels Kameras, GPS-Ortung, Gefahreneinschätzung mittels Wärmesensoren usw. begreifen.

[96] Zum Begriff der Autonomie s. Abschnitt 5.2.2.

[97] Eine Darstellung der Rolle der Intuition im Kontext der Technikanthropologie liefert z. B. Hans-Joachim Braun (2020).

Maßstab für die Moral technischer Entwicklungen sollen herangezogen werden können. Dies soll in der vorliegenden Arbeit dabei nicht umfassend geleistet werden, da es ja primär um Moral und Ethik in technischen Systemen geht. Ein Blick auf die Bedingungen von Moral und Ethik bei Menschen kann jedoch helfen, die allzu oft intuitiv gezogene Abgrenzung von Menschen und Maschinen differenzierter zu betrachten.

Einer der Philosophen, die im Zusammenhang mit freien Entscheidungen als Voraussetzungen für Moralfähigkeit in Verbindung gebracht werden, ist Immanuel Kant. Freiheit versteht er nicht nur als *negative Freiheit*, also Freiheit von äußeren Zwängen und eingrenzenden Umständen, sondern als eine jedem Menschen inhärent innewohnende Fähigkeit zur Selbstgesetzgebung, was er wiederum als *Autonomie* (s. u.) bezeichnet (vgl. KpV, AA V). In der Philosophie Immanuel Kants wird das, was Menschen tun sollen oder nicht, das Moralische (oder das *Sittliche* in Kants Terminologie), durch Gesetze bestimmt, die allein durch die Vernunft (im Gegensatz zu den Sinnen) erfasst werden können. Empirische Erfahrung spielt ihm zufolge zwar als Grundlage für jegliche Erkenntnisfähigkeit durchaus eine Rolle, die „Sinnenwelt“ ist seiner Ansicht nach aber komplett getrennt von der „Verstandeswelt“ (GMS, AA IV: 451 ff.). Somit sind die moralischen Gesetze von den kausalen Naturgesetzen völlig unabhängig, worin für Kant das Potenzial der *transzendentalen Freiheit* liegt:

> Da die bloße Form des Gesetzes lediglich von der Vernunft vorgestellt werden kann, und mithin kein Gegenstand der Sinne ist, folglich auch nicht unter die Erscheinungen gehört, so ist die Vorstellung derselben als Bestimmungsgrund des Willens von allen Bestimmungsgründen der Begebenheiten in der Natur nach dem Gesetze der Kausalität unterschieden, weil bei diesen die bestimmenden Gründe selbst Erscheinungen sein müssen. Wenn aber auch kein anderer Bestimmungsgrund des Willens für diesen zum Gesetz dienen kann, als bloß jene allgemeine gesetzgebende Form, so muß ein solcher Wille als gänzlich unabhängig von dem Naturgesetz der Erscheinungen, nämlich dem Gesetz der Kausalität, beziehungsweise auf einander, gedacht werden. Eine solche Unabhängigkeit aber heißt F r e i h e i t im strengsten, d. i. transzendentalen Verstande. Also ist ein Wille, dem die bloße gesetzgebende Form der Maxime allein zum Gesetze dienen kann, ein freier Wille.
>
> (KpV, AA V: 28 f., 1. Teil, Erstes Buch, *Erstes Hauptstück*, §5)

Die *praktische Freiheit* besteht demnach darin, sich selbst vermittels der Vernunft Gesetze geben und danach handeln zu können – im Unterschied zu Tieren etwa, die sich weder selbst Gesetze geben, noch danach handeln können, sondern gemäß Kant vielmehr der Willkür ihrer „sinnliche[n] Antriebe“ (KrV, AA III: 802) unterworfen sind. Menschen haben vielmehr die Fähigkeit, unabhängig von der *Sinnenwelt* „mithin durch Bewegursachen, welche nur von der Vernunft vorgestellet werden“, zu handeln (ebd.). Die praktische Freiheit von Menschen besteht für Immanuel Kant also im Bewusstsein der sittlich-moralischen Grundgesetze (die

allen Menschen zumindest theoretisch zugänglich sind) und der Möglichkeit, ihrer gemäß handeln zu können. Sich der moralischen Gesetze auch wirklich zu bedienen, ist letztendlich eine Frage der Autonomie, auf die ich im folgenden Abschnitt ausführlicher zu sprechen kommen werde:

> Als ein vernünftiges, mithin zur intelligibelen Welt gehöriges Wesen kann der Mensch die Causalität seines eigenen Willens niemals anders als unter der Idee der Freiheit denken; denn Unabhängigkeit von den bestimmenden Ursachen der Sinnenwelt (dergleichen die Vernunft jederzeit sich selbst beilegen muß) ist Freiheit. Mit der Idee der Freiheit ist nun der Begriff der Autonomie unzertrennlich verbunden, mit diesem aber das allgemeine Princip der Sittlichkeit, welches in der Idee allen Handlungen vernünftiger Wesen eben so zum Grunde liegt, als das Naturgesetz allen Erscheinungen.
>
> (GMS, AA V: 452 f.)

Moralisches Handeln gemäß den Gesetzen, die vermittels der Vernunft erkannt werden können, ist gemäß Immanuel Kant also eine Frage der Freiheit – im Gegensatz zu determinierten Abläufen, die beispielsweise den Naturgesetzen unterliegen.

> Es kommt nämlich bei der Frage nach derjenigen Freiheit, die allen moralischen Gesetzen und der ihnen gemäßen Zurechnung zum Grunde gelegt werden muß, darauf gar nicht an, ob die nach einem Naturgesetze bestimmte Kausalität, durch Bestimmungsgründe, die im Subjekte, oder außer ihm liegen, und im ersteren Fall, ob sie durch Instinkt oder mit Vernunft gedachte Bestimmungsgründe notwendig sei; wenn diese bestimmende Vorstellung nach dem Geständnisse eben dieser Männer selbst, den Grund ihrer Existenz doch in der Zeit und zwar dem v o r i g e n Z u s t a n d e haben, dieser aber wieder in einem vorhergehenden etc., so mögen sie, diese Bestimmungen, immer innerlich sein, sie mögen psychologische und nicht mechanische Kausalität haben, d. i. durch Vorstellungen, und nicht durch körperliche Bewegung, Handlung hervorbringen, so sind es immer B e s t i m m u n g s g r ü n d e der Kausalität eines Wesens, so fern sein Dasein in der Zeit bestimmbar ist, mithin unter notwendig machenden Bedingungen der vergangenen Zeit, die also, wenn das Subjekt handeln soll, nicht mehr in s e i n e r G e w a l t sind, die also zwar psychologische Freiheit (wenn man ja dieses Wort von einer bloß inneren Verkettung der Vorstellungen der Seele brauchen will), aber doch Naturnotwendigkeit bei sich führen, mithin keine t r a n s z e n d e n t a l e F r e i h e i t übrig lassen, welche als Unabhängigkeit von allem Empirischen und also von der Natur überhaupt gedacht werden muß, sie mag nun [als] Gegenstand des inneren Sinnes, bloß in der Zeit, oder auch der äußeren Sinne, im Raume und der Zeit zugleich betrachtet werden, ohne welche Freiheit (in der letzteren eigentlichen Bedeutung), die allein a priori praktisch ist, kein moralisch Gesetz, keine Zurechnung nach demselben, möglich ist.
>
> (KpV: AA V: 172 ff.; 96 ff., 1. Teil, Elementarlehre, 1. Buch)

Der Informatiker und Philosoph Stefan Ullrich sieht Maschinen ganz eindeutig determinierten Prozessen unterworfen. Ihmzufolge kommen „algorithmisch arbeitende Maschine[n; L. S. . . .] bei gleicher Eingabe immer wieder zum gleichen Ergebnis,

dort gibt es keine Diskussion“ (Ullrich 2019b: 258).[98] Immanuel Kant selbst bemühte den Ausdruck der *Freiheit eines Bratenwenders*, um seine Idee *transzendentaler Freiheit*, die der Moralität zugrundeliegt, dem „Naturgesetze der Kausalität“ gegenüberzustellen. Kant kritisiert dabei Autoren, die letzteres als menschliche Freiheit verstünden und damit den Menschen zu Automaten gleichen „Maschinenwesen“ machten (KpV, AA V: 172 ff., 96 ff., 1. Teil, Elementarlehre, 1. Buch):

> Eben um deswillen kann man auch alle Notwendigkeit der Begebenheiten in der Zeit nach dem Naturgesetze der Kausalität den M e c h a n i s m u s der Natur nennen, ob man gleich darunter nicht versteht, daß Dinge, die ihm unterworfen sind, wirklich materielle M a s c h i n e n sein müßten. Hier wird nur auf die Notwendigkeit der Verknüpfung der Begebenheiten in einer Zeitreihe, so wie sie sich nach dem Naturgesetze entwickelt, gesehen, man mag nun das Subjekt, in welchem dieser Ablauf geschieht, *Automaton materiale*, da das Maschinenwesen durch Materie, oder mit Leibniz *spirituale*, da es durch Vorstellungen betrieben wird, nennen, und wenn die Freiheit unseres Willens keine andere als die letztere (etwa die psychologische und komparative, nicht transzendentale, d. i. absolute, zugleich) wäre, so würde sie im Grunde nichts besser, als die Freiheit eines Bratenwenders sein, der auch, wenn er einmal aufgezogen worden, von selbst seine Bewegungen verrichtet. (Ebd., Herv. i. Orig.)[99]

Tiere verfügen nach Kant beispielsweise lediglich über diese *Freiheit eines Bratenwenders*, da sie der Kausalität der Naturgesetze unterliegen und somit weder frei noch autonom gemäß den selbst gesetzten moralischen Zielen (auch *Zwecken*) handeln können.[100] In diesem Sinne können auch moralische Akteure gemäß der oben vorliegenden Definition verstanden werden.

Eine ausführliche Analyse der kantischen Philosophie kann und soll an dieser Stelle nicht geleistet werden. Auch die Frage, ob das Konzept für (alle) Menschen realistischerweise umsetzbar ist, muss hier offenbleiben (mit dem Verweis auf interpretatorische Schwierigkeiten, die in Schönecker (2005) deutlich werden). Es hätten

[98] Andererseits gibt es zuweilen viele verschiedene algorithmische Wege, um ein bestimmtes Problem zu lösen (vgl. Stiller 2015).

[99] Zur Idee, Gottfried Wilhelm Leibniz als „Entwerfer einiger Konzepte, die der künstlichen Intelligenz zugrunde liegen“, zu sehen vgl. Centrone 2021: 1.

[100] Vgl. Schmitz 2017a: 35 f. Als nicht-vernünftige Wesen sind sie demnach *Sachen* (GMS, AA IV: 428), wobei Menschen ihnen gegenüber durchaus *indirekte Pflichten* haben, da ein schlechter Umgang mit Tieren auf den Charakter von Menschen selbst zurückfalle (MS, AA VI: 296). Ähnliches wird auch für (moralische) Maschinen diskutiert (vgl. u. a. Göcke 2020); s. auch Abschnitt 3.2.1. Was den Status von Tieren angeht, unterscheidet sich Kants Position im Übrigen wenig vom deutschen Tierschutzgesetz. Christine Korsgaard (2017) versucht hingegen, aus Teilen der kantischen Philosophie ein Konzept zu entwickeln, in dem „Tiere als Zwecke an sich“ anerkannt werden, vgl. Schmitz 2017a: 36, Fußnote 68.

auch noch viele weitere Ansätze genannt werden können, was auch für die referierten Ansätze in den folgenden Abschnitten gilt. Die Darstellung eines möglichen Konzepts der Voraussetzungen menschlicher moralischer Handlungen am Beispiel der Freiheit soll daher als Ergänzung verstanden werden, um die Komplexität möglicher Ansätze zu verdeutlichen. Ziel der kurzen Vorstellung war somit auch nicht eine eindeutige Beantwortung der Frage, ob sich der kantische Freiheitsbegriff für technische Systeme eignet.

5.2.2 Autonomie

Autonomie in technischen Systemen

Thimm und Bächle (2019: 73) zufolge ist der Ausdruck *Autonomie* in der Maschinenethik „[...u]biquitär, wenn nicht sogar inflationär gebraucht“ sowie „zu einer Metapher für die Loslösung der Technik aus der menschlichen Kontrollsphäre geworden“ (Thimm 2019: 31). Der Ausdruck *Autonomie* kann, je nach zeitgeschichtlichem Kontext, viele Interpretationen beinhalten ebenso wie in technikphilosophischen und maschinenethischen Untersuchungen. Dort wird *Autonomie* häufig verwendet, wenn eigentlich *Automatisierung* gemeint ist. So ist beispielsweise von *autonomen Fahrzeugen* oder *autonomen Robotern* die Rede, wenn betont werden soll, was diese selbst, ohne menschliches Zutun in einem Moment der Handlungsentscheidung, tun können (vgl. z. B. Grunwald 2020: 77). Autonomie im Verständnis Kants und Automatisierung sind jedoch, gemessen am heutigen Stand technischer Systeme, geradezu konträr. Während es in dem einen Fall um einen individuellen Prozess geht, der, gemessen an selbst hervorgebrachten Gesetzen, moralisch richtiges Handeln zum Ziel hat, geht es in dem anderen Fall um vorab und von außen festgelegte Strukturen, denen ein System ohne Reflexionsspielraum folgt.[101] Dabei muss betont werden, dass moralische Handlungen gemäß Immanuel Kant anhand der Intentionen beurteilt werden, also auf der Basis des *guten Willens* (vgl. hierzu auch Celikates und Gosepath 2017c: 203). Diese metaethische Position unterscheidet sich von dem oben begründeten Fokus auf die Handlungsfolgen.

(Teil-)Autonomie bei technischen Systemen ist laut Thimm und Bächle (2019: 73) einerseits positiv belegt, andererseits bestehe die „Furcht vor (zu) mächtiger Technologie und dem damit verbundenen Verlust der Kontrolle durch den Menschen“ (s. hierzu auch Kapitel 3 in der vorliegenden Arbeit und insbesondere die Rolle medialer Vermittlung technischer Entwicklungen in Abschnitt 3.5). Gemäß

[101] Unter Beachtung der epistemologischen Notwendigkeit, die Immanuel Kant beschreibt, a priori vorhandene moralische Gesetze selbst zu erkennen, und somit *vernünftig* zu sein, sind Argumentationen für eine Interpretation denkbar, die eine eingeschränkte Freiheit in der Konzeption sehen. Doch selbst dieser Argumentation folgend, ist man dann von einem Automatismus im obigen Sinn immer noch sehr weit entfernt.

Thimm und Bächle (2019: 82) wird im Diskurs über moralische Maschinen häufig Autonomie für diese vorausgesetzt:

> Der Diskurs der moralischen Maschine (die ethische Reflexion des richtigen und guten Handelns einer Maschine) erschafft paradoxerweise dabei die Wahrnehmung, Technik lasse sich überhaupt als autonom kategorisieren (‚autonome Systeme'). Mit anderen Worten setzt die Erwartung einer ‚moralischen Maschine' implizit eine ‚autonome Maschine' voraus. Sie plausibilisiert die Vorstellung einer Selbstgesetzgebung, die Maschinen als Agenten entwirft, die durch kritisches Abwägen und spezifische Deutungen sozialer Kontexte zu moralischen Entscheidungen fähig sind. Der ethische Diskurs konstruiert die Vorstellung von eigenständig handelnden, intelligenten Maschinen, die als dem Menschen ebenbürtig erscheinen.

Es kann bestritten werden, ob es der *ethische* Diskurs ist, der solche Vorstellungen zeichnet, oder ob es sich hierbei nicht vielmehr um alltäglich-umgangssprachliche Zuschreibungen handelt. Das Zitat zeigt jedoch, dass der Begriff Autonomie, im philosophischen Sinn verstanden, den Übergang von moralischen zu ethischen Techniksystemen (u. a.) markiert, wobei die Zuschreibung von Autonomie hypothetische technische Systeme kreiert, die bislang allenfalls in Science-Fiction-Szenarien vorkommen. Caja Thimm und Thomas Bächle bemerken demgemäß eine „heterogene Verwendung" des Ausdrucks *Autonomie* (ebd.: 76). Catrin Misselhorn betont die unterschiedliche Deutungsweise des Autonomiebegriffs in verschiedenen Disziplinen:

> Die Fähigkeit, selbst ein Verhalten zu initiieren, wird in der Informatik häufig mit Autonomie gleichgesetzt. Allerdings gibt es verschiedene Autonomiebegriffe, die nicht gleichermaßen anspruchsvoll sind. Nach dem Philosophen Stephen Darwall gibt es vier Formen der Autonomie: Personale, moralische und rationale Autonomie sowie Handlungsautonomie. *Personale Autonomie* besteht in der Fähigkeit, individuelle Werte auszubilden und sein Handeln an diesen auszurichten. *Moralische Autonomie* erfordert, dass man Handlungsentscheidungen aufgrund der eigenen moralischen Überzeugungen oder Prinzipien trifft. Ein *rationaler autonomer* Akteur folgt den aus seiner Sicht gewichtigsten Gründen. Die schwächste Form der Autonomie ist die *Handlungsautonomie*; sie liegt bereits dann vor, wenn eine Handlung gegeben ist, die einem Individuum als Akteur zugeschrieben werden kann.
> (Misselhorn 2019: 76. Herv. i. Orig.)

Die Umsetzung der *Ausbildung individueller Werte*, die *Ausrichtung des eigenen Handelns* daran sowie die Angabe *rationaler Gründe* kann dabei in technischen Systemen anders aussehen als bei Menschen. Informationstechnisch gesehen, kann zum Beispiel auch ein effizienter, pragmatischer Lösungsweg als *guter Grund* für ein technisches System angesehen werden, auf diese und jene Weise zu operieren.

Floridi und Sanders (2004) bauen in ihr Konzept moralischer Maschinen eine Form basaler Autonomie ein, die mit der Betrachtung verschiedener Abstraktionsgrade kompatibel ist. Wallach und Allen 2009: 60 wiederum definieren Autonomie recht anspruchsvoll als „[. . . a]bility to change state without stimulus, that is, without direct response to interaction, which results in a certain degree of complexity and decoupledness from the environment."

Autonomie als notwendige Voraussetzung für ethische Reflexionen?

Wie die kurze Darstellung des Freiheitsbegriffs nach Immanuel Kant gezeigt hat, ist für sein Konzept *Autonomie* eine wesentliche Voraussetzung. In der *Kritik der praktischen Vernunft* setzt Immanuel Kant die beiden Begriffe sogar gleich. In der kantischen Moralphilosophie sind Freiheit und Autonomie unweigerlich an die sittlichen Gesetze und den kategorischen Imperativ gebunden und somit notwendig für moralisches Handeln, was bei Kant ethische Reflexion einschließt. Autonomie besteht demnach darin, sich frei und vermittels menschlicher Vernunft inhärenten moralischen Gesetze zu bedienen und gemäß dem kategorischen Imperativ zu handeln. Wäre ein technisches System in der Lage, eigene Gesetze auf ethisch legitimierter Begründungsbasis zu erstellen, zu reflektieren und auf dieser Grundlage Handlungsoptionen abzuwägen, würde es sich in dieser Hinsicht nicht wesentlich von einem autonomen Menschen in der Definition Kants unterscheiden.

Moderne Autonomie- und Freiheitsbegriffe weichen davon erheblich ab, insbesondere auch die umgangssprachliche Verwendung der Ausdrücke.[102] Ethische Techniksysteme könnten mit einem weniger voraussetzungsreichen Begriff insofern über *Autonomie* verfügen, dass sie ohne äußeres Zutun in verschiedenen Situationen flexibel mögliche Handlungsoptionen abwägen können und diese in Aktionen umsetzen können müssten. So definiert, besteht der Unterschied zur oben skizzierten *Freiheit* auf der Betonung der Selbständigkeit, die eine Voraussetzung für freies Denken und Handeln wäre. Freiheit und Autonomie sind also, dies kann aus der kantischen Philosophie übernommen werden, nicht ohne einander zu denken, wobei das kantische Konzept für eine praktische Umsetzung stark vereinfacht oder auf technische Begriffe übertragen werden müsste.

[102] Dies hat der Diskurs um den Begriff der Freiheit in der Corona-Pandemie gezeigt, in dem unter *Freiheit* und *Autonomie* häufig verstanden wurde, tun und lassen zu können, was man möchte – unabhängig vom Wohl anderer oder eines gesamtgesellschaftlich verantwortlichen Umgangs mit der Krisensituation.

5.2.3 Rationalität

Rationalität als notwendige Voraussetzung für ethische Reflexionen?

Ein Faktor, der immer wieder als Voraussetzung für ethische Reflexionsfähigkeit diskutiert wird, ist Rationalität, im Englischen meist als *rationality*, manchmal auch als *reasoning* bezeichnet. Im Deutschen wird der Ausdruck meist mit *vernünftigem Handeln* übersetzt, worunter zunächst das Handeln aus bestimmten Gründen verstanden wird (vgl. hierzu auch Gosepath 2018), die als sinnvoll, notwendig oder moralisch richtig eingestuft werden:

> Die klassische philosophische Handlungstheorie erachtet das Handeln aus Gründen als eine notwendige Bedingung der Handlungsfähigkeit. Nach der auf David Hume (1711-1776) zurückgehenden Theorie rationalen Handelns, die als Standardtheorie gilt, besteht ein Handlungsgrund in der Kombination zweier mentaler Zustände: einer Meinung und einer Pro-Einstellung. Eine Meinung liegt vor, wenn jemand einen Sachverhalt für wahr hält, während eine Pro-Einstellung gegeben ist, wenn jemand einen Sachverhalt realisieren möchte, der noch nicht vorliegt. Typische Pro-Einstellungen sind Wünsche. Dieser Ansatz wird daher im Englischen häufig als *Belief-Desire*-Theorie bezeichnet. Anders als dieser Begriff nahelegt, können Pro-Einstellungen aber auch negativ sein, d. h. die Vermeidung eines Sachverhalts anstreben, oder in anderen mentalen Zuständen als Wünschen bestehen, etwa Hoffnungen. [...] Neuere Varianten der Standardtheorie nehmen an, dass eine *Intention* hinzukommen muss, die festlegt, welcher Wunsch handlungswirksam wird und einen entsprechenden Handlungsplan umfasst. Dies soll dem Einwand Rechnung tragen, dass wir eine ganze Menge unverbindlicher Wünsche haben, die nicht zu Handlungen führen.
>
> (Misselhorn 2019: 81 f.)

In Anlehnung an die Definition von Moral nach Gerhard Ernst können alle möglichen Gründe als moralische Gründe bezeichnet werden, dann eben im jeweils engeren oder weiteren Sinn – die Frage ist, ob es sich dabei jeweils auch um *gute* bzw. *vernünftige* Gründe handelt. Michael Quante (2020: 77) weist darauf hin, dass die „handlungstheoretische Terminologie, d. h. die Rede von Absichten und Handlungsgründen [...] in der Literatur nicht einheitlich“ sei. Auf eine spezifische Problemstellung bei der Analyse von Handlungsgründen werde ich im Abschnitt zu Intentionen (5.2.7) weiter unten eingehen. Unter praktischer Rationalität wird, einfach gesagt, verstanden, vernünftige Pläne zu fassen und diese in die Tat umsetzen zu können (vgl. ebd.: 85 f.).

Die Gründe von Handlungen können ebenso einer ethischen Bewertung unterzogen werden wie die eigentlichen Handlungen. Diese Trennung wurde bereits in Abschnitt 4.5 am Beispiel des Chatbots *Tay* diskutiert. Dort habe ich auch festgehalten, dass es die Auswirkungen von Handlungen moralischer und ethischer Maschinen sind, die für eine ethische Bewertung entscheidend sind, wobei insbesondere für ethische Maschinen die Reflexion ethisch begründeter, und somit

rationaler, Gründe ein essenzieller Bestandteil ist. Von potenziell ethisch reflektierenden technischen Systemen kann erwartet werden, dass diese insofern über praktische Rationalität verfügen, als sie an anerkannten ethischen Normen und Werten orientierte, begründete Urteile über mögliche Handlungen fällen, auf dieser Basis Pläne für mögliche Handlungsoptionen aufstellen und diese dann auch umsetzen können. Catrin Misselhorn (2019: 82) formuliert ein Beispiel am Prinzip des ethischen Prinzips des Utilitarismus:

> Eine moralische Handlung geht demnach auf einen moralischen Grund zurück, also auf eine moralische Pro-Einstellung und eine entsprechende Meinung. Ein moralischer Grund kann sich etwa zusammensetzen aus dem utilitaristischen Werturteil, dass es gut ist, die Lust-Leid-Bilanz der von einer Handlung Betroffenen zu maximieren (Pro-Einstellung), und der Meinung, dass eine Spende für eine wohltätige Organisation zur besten Bilanz führt.

Wie ethische Techniksysteme zu solchen Bewertungen kommen hängt auch von den jeweiligen ethischen Prinzipien ab, denen ein System folgt (s. Abschnitt 5.4), wobei auch bekannte Fallbeispiele zur Orientierung für die ethische Begründung moralischer Urteile denkbar sind.

Rationalität in technischen Systemen

Bei technischen Systemen hatten bislang Menschen die Oberhand bei der Bestimmung der Handlungsziele und der damit verbundenen Beurteilung der Rationalität – unabhängig davon, ob der jeweilige, initiale Output bereits als moralisch gut oder nicht beziehungsweise als erwünscht oder nicht erwünscht bewertet wurde und wird (s. Beispiele *Tay*; *deep fakes* usw.). Mit genuin ethischen Maschinen würde sich das ändern, was eine hierarchische Verschiebung zumindest bei der Beurteilung von Handlungen zur Folge hätte. Gleichwohl hängt die grundlegende Frage der Macht über Entwicklungsentscheidungen von der Frage ab, ob *Maschinenethik* auch durch komplexere technische Systeme nach wie vor im Sinn einer *Menschenethik* verstanden wird, oder ob vielmehr auch Ansätze primär technisch erzeugter ethischer Reflexionen, Werte und Normen möglich sind und welche Folgen das hätte (s. S. 30).

Ein Ansatz für die technische Realisierung von Rationalität bietet die Theorie intentionaler Systeme von Daniel Dennett. Im Kontext seiner Überlegungen zu *Intentionalen Systemen*[103] schlägt er auch eine Form von Rationalität vor, die eine Anwendung auf technische Systeme ermöglichen kann. Auf den Begriff der Intentionalität werde ich im entsprechenden Abschnitt unten noch genauer eingehen;

[103]Dennett wählt die Schreibweise mit Großbuchstaben, um seinen Ansatz von anderen abzugrenzen, vgl. Dennett 1971: 87, Fußnote 1.

an dieser Stelle folgt daher nur ein kurzer Verweis auf Dennetts Interpretation des Begriffs:

> I wish to examine the concept of a system whose behavior can be (at least sometimes) explained and predicted by relying on ascriptions to the system of beliefs and desires (and hopes, fears, intentions, hunches, ...). I will call such systems Intentional systems and such explanations and predictions Intentional explanations and predictions in virtue of the Intentionality of the idioms of belief and desire (and hope, fear, intention, hunch,...).
>
> (Dennett 1971: 87)

Innerhalb Intentionaler Systeme definiert Dennett Rationalität als „optimal design relative to a goal or optimally weighted hierarchy of goals". In diesem Sinne ist ein (technisches) System dann rational, wenn es zu einem bestimmten Zweck konzipiert wurde und diesen auch einhält. Die Betonung liegt hier allerdings auf dem – natürlicherweise menschengemachten – Design von technischen Systemen, wodurch die Rationalität bzw. ethische Gewichtung möglicher Handlungsbeurteilungen vorab vorgegeben ist. Dieser Definition folgend sind Fälle wie die Reaktionen des oben beschriebenen Chatbots *Tay* also keine Form von interner Irrationalität eines Algorithmus oder Programms, sondern eine Form von menschlicher Unwissenheit über die Folgen und Möglichkeiten lernender Systeme – zum damaligen Zeitpunkt und Wissensstand. Sachbuchautor Brian Christian (2021: 12 f.) vergleicht das Verhältnis von Menschen und lernenden Systemen mit dem Narrativ aus Johann Wolfgang von Goethes *Zauberlehrling*:

> As machine-learning systems grow not just increasingly pervasive but increasingly powerful, we will find ourselves more and more often in the position of the ‚sourcerer's apprentice': we conjure a force, autonomous but totally compliant, give it a set of instructions, then scramble like mad to stop it once we realize our instructions are imprecise or incomplete – lest we get, in some clever, horrible way, precisely what we asked for.[104]

5.2.4 Bewusstsein

Neben den Feldern der Sprache und der Kunst, die bis vor kurzem der *conditio humana* (beziehungsweise der *conditio animalis* in Bezug auf Sprachvermögen) als vorbehalten geglaubt wurden (s. Einleitung in dieser Arbeit), finden sich noch weitere Bereiche, in die Algorithmen möglicherweise vordringen, zumindest wird die Möglichkeit schon lange diskutiert und erforscht. Die Rede ist von dem Feld der

[104] Vgl. zu menschlichen bzw. *sozialen* Fehlern bei der Computerentwicklung auch Millar 2020.

Psyche, was unter anderem Emotionen und Bewusstsein einschließt (vgl. Wendland 2022). Eine entscheidende Frage dabei ist nicht nur, *ob* Maschinen zu moralischen Entscheidungen und Handlungen sowie den dafür notwendigen Fähigkeiten in der Lage sind, sondern auch, ob sie das nicht insgesamt *besser* könnten als Menschen. Mit zunehmender Komplexität könnte sich die Überlegenheit von Maschinen über Menschen also nicht nur auf technischer Ebene niederschlagen, sondern auch auf psychischer und ethischer. In Ian McEwans Roman *Maschinen wie ich* spricht ein fiktiver Alan Turing folgende Worte über seine eigene Erfindung: humanoide Roboter:

> Meiner Meinung nach waren die A.s und E.s einfach zu schlecht dafür gerüstet menschliche Entscheidungsfindung verstehen zu können – wie unsere Prinzipien im Kraftfeld unserer Emotionen entstellt werden, unserer persönlichen Vorurteile, Selbsttäuschungen und all unserer anderen hinreichend bekannten kognitiven Mängel. Daran sind diese Adams und Eves schon früh verzweifelt. Sie konnten uns nicht verstehen, weil wir uns selbst nicht verstehen. Ihre Lernprogramme waren mit uns überfordert. Wenn wir unser eigenes Innerstes nicht begreifen, wie sollten wir da ihres gestalten und erwarten, dass sie mit uns glücklich werden?
>
> (McEwan 2019: 395)

Wie eingangs bereits erwähnt, gibt es Arten von Tätigkeiten, bei denen Menschen (so zumindest die naheliegende Vermutung) ganz froh sind, dass Maschinen sie übernehmen: Wäsche waschen, beim Einparken helfen usw. Sobald es jedoch an mentale oder psychische Eigenschaften geht, kann ein mulmiges Gefühl aufkommen: Können Maschinen Bewusstsein haben? Was bedeutet das? Und was ist überhaupt Bewusstsein bei Menschen?

> Many people believe that machines are incapable of being truly conscious, incapable of the genuine understanding and emotions that define humans' most important relationships and shape humans' ethical norms.
>
> (Wallach und Allen 2009: 55)

Bewusstsein in technischen Systemen

Wallach und Allen haben eine eigene Definition von Bewusstsein, die sich teilweise mit den bereits angesprochenen Voraussetzungen für ethische Maschinen deckt:

> The term [consciousness; L. S.] is used to mark the distinction between being awake or asleep, as well as to capture a range of higher-order cognitive functions, including the abilities to be attentive, to plan, and to experience. There are unusual states of consciousness that include dreaming, psychotic experiences, peak experiences, and flow.
>
> (Ebd.: 66)

Wie auf Seite 98 thematisiert, stellen sich vor allen Dingen die Fragen, ob Maschinen über die diskutierten Eigenschaften verfügen können und ob diese notwendig sind für ethische Reflexionen. Benötigen Maschinen, Roboter oder Algorithmen also eine Form von Bewusstsein, um eine Situation verstehen, ethisch reflektieren und zu einer angemessenen moralischen Entscheidung kommen zu können? „Whether computer understanding will ever be adequate to support full moral agency remains an open question“, resümieren Wallach und Allen (2009: 69).

Die vielfältige Definierbarkeit von Bewusstsein und das Fehlen eines einheitlichen Konsenses des Begriffsverständnisses nicht nur über Disziplinen hinweg, sondern auch innerhalb einer Disziplin, können dabei zu einem intuitiven Verständnis des Bewusstseinsbegriff, auch innerhalb wissenschaftlicher Debatten, verführen. So schreibt Daniel Levy beinahe resigniert:

> Given this huge difficulty in finding a universally accepted definition of consciousness, I prefer to take a pragmatic view, accepting that it is sufficient for there to be a general consensus about what we mean by consciousness and to assume that there is no burning need for a rigorous definition – let us simply use the word and get on with it.
>
> (Levy 2009: 210)

Für die Beantwortung von Levys Leitfrage, wie eine adäquate moralische (*ethical*) Behandlung von *artificially conscious robots* aussehen kann, lässt er seinen Bewusstseinsbegriff neben diesem intuitiven Alltagsverständnis von Bewusstsein auf einer Analogie zu Alan Turings Untersuchung *denkender* Maschinen basieren. Wie auf Seite 15 f. dargestellt, ersetzte Alan Turing die Frage, ob Maschinen denken können, durch ein Gedankenexperiment. Analog dazu geht Levy davon aus,

> that if a machine exhibits behaviour of a type normally regarded as a product of human consciousness (whatever consciousness might be), then we should accept that the machine has consciousness.
>
> (ebd.: 211)

Diese Definition hat mehrere Haken. Zum einen führt die Ergänzung „whatever consciousness might be“ zu einem Widerspruch des Arguments. Denn wie soll man erkennen, ob ein Mensch oder eine Maschine sich wie jemand verhält, dem normalerweise Bewusstsein zugeschrieben wird, wenn nicht klar ist, was Bewusstsein bedeutet? Zum anderen birgt der behavioristische Erklärungsansatz die Möglichkeit der Täuschung; so ist es leicht vorstellbar, dass sich eine Person nur so verhält, *als ob* sie Bewusstsein habe. Außerdem widerspricht eine behavioristische Definition von Bewusstsein dem von Levy zunächst bevorzugten alltäglichen Verständnis von Bewusstsein, denn es sind technische Systeme denkbar, die sich so verhalten,

als hätten sie Bewusstsein, denen aber intuitiverweise kein menschenähnliches Bewusstsein zugesprochen werden würde. Für Levy spielt es jedoch keine Rolle, ob ein Roboter tatsächlich Bewusstsein *hat* beziehungsweise *bei Bewusstsein ist* oder ob er nur so tut, als ob (vgl. Levy 2009: 211) – dabei bezieht er sich wieder auf Turings Ansatz: „‚the only way by which one could be sure that a machine thinks is to *be* the machine and to feel oneself thinking.' The same applies to consciousness" (ebd.).

Der behavioristische Definitionsansatz ist auf den ersten Blick nicht befriedigend, bietet aber die Möglichkeit, nicht nur mit dem Begriff des Bewusstseins, sondern auch mit anderen, von außen schwer zugänglichen psychischen Eigenschaften umzugehen und diese für Maschinen zu diskutieren. Definitionen, die technischen Systemen von vorn herein Bewusstsein absprechen, können wiederum tückisch sein, da sie Formen von Diskriminierung von Menschen beinhalten können, wie z. B. die Definition von Kestutis Mosakas (2021). Er hält es nicht für zielführend, menschliche Formen von Bewusstsein für Maschinen als Voraussetzung für moralische und ethische Aktionen zu definieren; Roboter sollten demnach auch nicht als moralische Objekte behandelt werden, solange sie keine Formen von Bewusstsein aufweisen. Konsequenterweise stellt sich hier allerdings die Frage, was mit Menschen ist, die nicht über Bewusstsein verfügen? Hier hilft nur eine kategoriale Unterscheidung, die, wie u. a. in Abschnitt 3.2.1 erläutert, argumentatorisch nicht haltbar ist.

Kenneth Einar Himma (2009: 19) beschreibt das Phänomen des Bewusstseins näher und argumentiert, dass dieses, „i.e., the capacity for inner subjective experience like that of pain or, as Nagel puts it, the possession of an internal something-of-which-it is-to-be-like",[105] eine notwendige Voraussetzung für moralisches Handeln (*artificial moral agency*) darstellt. Allerdings beinhaltet dieses Konzept von Handlungen automatisch auch Verantwortlichkeit (*accountability*) und geht von einer kausalen Handlungserklärung aus: „X is an agent if and only if X can instantiate intentional mental states capable of directly causing a performance" (ebd.: 127). Wie ich in Abschnitt 5.2.7 noch zeigen werde, ist diese jedoch nur eine von mindestens zwei konfligierenden Ansätzen für Handlungserklärungen; neben Kausalrelationen werden hier auch Intentionen, also Absichten, und die Begründungen von Handlungen als Erklärungen diskutiert. Außerdem geht der Himma davon aus, dass alle „mental states" bewusst sind (ebd.: 27), was keineswegs so einvernehmlich annehmbar ist, wie es in dem Artikel dargestellt wird.[106]

[105] Referiert wird hier auf Thomas Nagels Aufsatz *What Is It Like to Be a Bat?*, auf den ich in Abschnitt 5.3.2 noch genauer eingehen werde (vgl. Nagel 1974).

[106] Peter Goldie (2009: 385 ff.) unterscheidet beispielsweise *reflexives Bewusstsein* von *nichtreflexivem Bewusstsein*, um nur eines von vielen möglichen Beispielen zu nennen.

Bewusstsein als notwendige Voraussetzung für ethische Reflexionen?

In Ian McEwans Maschinen-Geschichte wird der humanoide Roboter *Adam* im Lauf der Zeit so selbstständig, dass er (aus seiner Sicht) selbstverständlicherweise über das Geld verfügt, das er mittels Aktiengeschäften erwirtschaftet hat – er spendet es folglich an verschiedene *wohltätige Vereine* (McEwan 2019: 360).

Ein Grund für die These, dass Techniksysteme in der aktuellen Realität kein Bewusstsein haben können, kann die Tatsache sein, dass Bewusstsein auch bei Menschen nach wie vor als Rätsel gilt – in der Philosophie des Geistes ist die Rede von einer *explanatory gap* (vgl. Beckermann 2001: 237, 404–413), obwohl inzwischen einige Erklärungsansätze kursieren. In der Philosophie des Geistes haben Philosoph∗innen im Lauf der Zeit versucht, Bewusstsein mithilfe verschiedener Theorien zu erklären, mit reduktionistischen Ansätzen, die Bewusstsein auf bestimmte physikalische Prozesse zurückführen, mit idealistischen, die das Gegenteil behaupten, und mit solchen, die versuchen, beide Positionen zu vereinen.

Es fällt also schwer, Bewusstsein für ethische Reflexion vorauszusetzen, solange nicht klar ist, was damit genau gemeint ist. Eine Form von *Sich-seiner-selbst-bewusst-sein* scheint nicht notwendig für ethische Reflexionsfähigkeit zu sein, wobei diese These einer genaueren Untersuchung bedürfte, was darunter zu verstehen ist. Formen einer bestimmten *Wachsamkeit* (engl. *awareness*) müssten insofern vorliegen, dass ein ethisches Techniksystem in verschiedenen Situationen flexibel reflektieren und agieren können sollte. Eine solche Reaktionsfähigkeit ist aber sicher nicht mit (verschiedenen möglichen Formen von) Bewusstsein gleichzusetzen.

5.2.5 Emotionen und Gefühle

> Während ich die Hand nach dem Lichtschalter ausstreckte, fragte ich: ‚Wie fühlst du dich?‘
>
> Er wandte den Blick ab und suchte nach einer Antwort.
>
> ‚Ich fühle mich irgendwie nicht richtig.‘
>
> Diesmal klang die Stimme flach, fast als hätte ihm meine Frage die Laune verdorben. Aber Laune? Wo denn in all den Mikroprozessoren?
>
> ‚Was stört Dich?‘
>
> ‚Ich habe keine Kleider an. Und...‘
>
> ‚Ich besorg dir welche. Was noch?‘
>
> ‚Dieses Kabel. Wenn ich das rausziehe, tut es weh.‘
>
> ‚Ich mach’s, und es wird nicht wehtun.‘
>
> (McEwan 2019: 41 f.)

Ebenso wie Möglichkeit der Umsetzung kognitiver Fähigkeiten, sind Emotionen in technischen Systemen ein viel diskutiertes Thema. In Bezug auf Tiere herrschte auch gemäß wissenschaftlicher Erkenntnisse lange die Ansicht, dass diese nicht zu Emotionen fähig wären (vgl. z. B. Rollin 2011). Das ändert sich langsam; ist also

zu erwarten, dass es auch bei Robotern & Co. nur eine Frage der Zeit ist, bis wir ihnen Schmerzen derart zuschreiben, wie sie Adam in Ian McEwans Buch *Maschinen wir ich* äußert?

Der Nachweis des Vorliegens von Emotionen ist aus naturwissenschaftlicher Sicht etwas einfacher als der von mentalen Qualitäten, wie Bewusstsein, da viele Emotionen physiologische messbare Begleitsymptome haben (vgl. z. B. Barooah 2019). Nicht einfacher wird die Analyse dabei durch die verschiedenen Ausdrücke Emotionen (*emotions*) und Gefühle (*feelings*). Peter Goldie (2009: 369) beschreibt Gefühle als „ein intimer und vertrauter Bestandteil der emotionalen Erfahrung; ohne Gefühle wären Emotionen nicht, was sie sind". Gefühle sind demnach der subjektive Anteil des *Wie-es-sich-anfühlt* in Emotionen; letztere beinhalten darüber hinaus auch bestimmte Formen von Überzeugungen und Wünschen.

Emotionen als notwendige Voraussetzung für ethische Reflexionen?

Es gibt viele Ansätze, die zu zeigen versuchen, warum Emotionen für rationale Entscheidungen[107], wichtig sein können (vgl. de Sousa 2019). Auch dass Emotionen Prozesse wie Entscheidungsfindungen beeinflussen *können* steht außer Frage (vgl. Lerner, Valdesolo und Kassam 2014). Aber sind sie auch notwendig für ethische Reflexions- und -argumentationsprozesse? Im Zusammenhang moralischer Handlungen wird eine besondere Form von Gefühlen diskutiert: moralische Gefühle (im Englischen *moral sentiments*). Gemäß den moralischen *Sentimentalisten* oder *Emotivisten* sind moralische Aussagen mit Gefühlsausdrücken gleichzusetzen. Gerhard Ernst (2009: 82 f.) beschreibt den Ansatz im Kontext seiner metaethsichen Untersuchung:

> Die klassische antiobjektivistische Antwort auf die Frage, was wir tun, wenn wir einen Satz äußern wie ‚Es ist falsch, Menschen zum Spaß zu foltern', ist die des *Emotivisten*: Wir bringen ein Gefühl zum Ausdruck. Wir äußern Gefühle spontan oder um unsere Umwelt an unserem Innenleben teilhaben zu lassen, aber häufig auch in der Absicht, diese Umwelt zu beeinflussen. In der Regel entschlüpfen uns moralische Behauptungen nicht einfach (wie andere Gefühlsäußerungen, die wir manchmal nicht unterdrücken können). Ihre primäre Funktion wäre es demnach, unsere Mitmenschen über unser Befinden ins Bild zu setzen. Sie könnten aber auch ähnliche Funktionen wie Bitten, Empfehlungen oder gar Befehle übernehmen. [...] Der emotivistische Ansatz wurde von verschiedenen Philosophen variiert und weiterentwickelt. Heutigen Antiobjektivisten zufolge drücken wir mit moralischen Äußerungen nicht einfach Gefühle der Ablehnung oder Zustimmung aus, sondern bestimmte komplexe Empfindungen, eine besondere Art von Wünschen oder die Akzeptanz einer Norm.
>
> (Herv. i. Orig.)

[107] Ihnen können ethische Überlegungen vorausgehen.

Ob es dieser Art von Gefühlen oder empfundener Bewertungen für ethische Urteilsfähigkeit bedarf oder ob jene sogar mit diesen gleichzusetzen sind, wie die Emotivisten behaupten, ist eine große Streitfrage in der Philosophie, die seit vielen Jahrhunderten diskutiert wird (vgl. Döring 2019). Ein Gegenargument gegen die Notwendigkeit moralischer Gefühle für moralische Urteile besteht darin, die These des Missbrauchs hervorzubringen: Der Bezug auf Gefühle bei moralischen Urteilen birgt zwangsläufig die Gefahr des Missbrauchs, denn mit der Aussage, es habe sich *richtig angefühlt*, kann eigentlich jede Handlung gerechtfertigt werden, mag sie noch so unmoralisch sein. Gefühle können (und sollten) demnach allenfalls *ein* Faktor von mehreren bei der Urteilsfindung spielen. Damit moralische Gefühle nicht der Gefahr der Willkürlichkeit ausgeliefert sind, müssen sie sich beispielsweise auch in David Humes Konzeption, der die Theorie moralischer Gefühle maßgeblich geprägt hat, an einem allgemeinen Standard messen lassen, der die Moralität gewährleisten soll (vgl. u. a. Schmitter 2021). Diese muss zwangsläufig *top-down* installiert werden.

Wallach und Allen (2009: 180) denken, dass Gefühle moralische Urteile beeinflussen können (als *bottom-up propensities*), wobei Gefühle für sie zunächst getrennt sind von kognitiven Prozessen (*top-down considerations*).

> Feelings and inherent values embodied in people's unthinking reactions, for example the disgust associated with blood and other bodily fluids, may influence morality from the bottom up, but are not necessarily reflective of the values a society would recognize as moral values. Negative feelings may, for example, lead to prejudices when an agent automatically attaches such feelings to individuals who are not a part of the agent's immediate group. From a moral perspective, it is important to understand how top-down considerations use and apply these bottom-up propensities.

Gefühle scheinen also alles in allem einen wichtigen Einfluss auf moralische Urteile zu haben. Innerhalb eines holistischen Ansatzes kann man auch die These vertreten, dass kognitive Prozesse nur im Zusammenspiel mit emotionalen Prozessen funktionieren können. Die Frage nach der Notwendigkeit für ethische Prozesse kann durch die Unterscheidung zwischen Gefühlen und Emotionen geklärt werden: Gefühle und Emotionen können demnach kognitive Prozesse beeinflussen, müssen dies aber nicht notwendigerweise. Emotionen beinhalten hingegen per Definition auch kognitive Aspekte, die teilweise als wesentlicher Teil moralischer Urteile und ethischer Reflexionen verstanden werden können. Beide Aussagen bedürften freilich einer genaueren Untersuchung, insbesondere auch die, Emotionen welcher Art in ethischen Prozessen eine Rolle spielen.

Gefühle und Emotionen in technischen Systemen

Die in diesem Abschnitt eingangs zitierte Passage soll die Befremdlichkeit verdeutlichen, die man empfinden kann bei dem Gedanken daran, dass Maschinen oder Roboter über Gefühle verfügen können. Für viele Menschen scheint hier intuitiv die kategoriale Unterscheidung zwischen unbelebter Technik und belebten Menschen ganz eindeutig zu sein. Der Nachweis von subjektiven Gefühlen gestaltet sich zudem sowohl in Maschinen als auch in Menschen als nicht gerade einfach.

Ein Ansatz ist es, mittels verstärkendem Lernen in Programmen bestimmte Lösungswege gegenüber anderen zu bevorzugen und somit die *Entscheidungen* eines Systems zu beeinflussen (vgl. Wallach, Franklin und Allen 2010). Inwieweit diese Prozesse allerdings vergleichbar sind mit menschlichen *Gefühlen* kann natürlich diskutiert werden. Einfacher scheint es zu sein, allgemeine Wahrnehmungsprozesse in technischen Systemen zu simulieren. Um ethische Thesen und Handlungsmöglichkeiten abwägen und die Frage *Was soll ich tun?* reflektieren sowie zu einer gut begründeten Antwort kommen zu können, braucht ein Programm oder eine Maschine auf jeden Fall ein gewisses Verständnis von ihrem Kontext, was über Informationsvermittlung qua verschiedener Kanäle gewährleistet werden kann, beispielsweise Kameras, Wärmesensoren oder Echtzeitdaten anderer technischer Systeme. Ähnlich wird das Sammeln großer Mengen von Daten und Auswerten im Hinblick auf ein bestimmtes Kriterium (oder mehrere), das unter den Schlagworten *big data* und *machine learning* zusammengefasst wird, teilweise als vergleichbar mit menschlichen Erfahrungen gewertet (vgl. hierzu z. B. Kapitel 8 in Heßler und Liggieri 2020). Eine mögliche Strategie, wenn man einerseits dem Vorwurf des pauschalen Speziesismus entgehen, sich andererseits nicht mit der komplexen Thematik subjektiver Empfindungen in technischen Systemen konfrontiert sehen möchte, ist, mit dem Ausdruck *gefühlsähnlicher* Prozesse für Techniksysteme zu operieren, die die *Funktion* von Gefühlen in emotivistischen Ansätzen erfüllen können. Geht es letztendlich darum, dass moralische Urteile nicht nur *top-down* durch kognitive Reflexionen, Abwägungen und Urteilsfindungen zustande kommen, sondern auch einen flexibleren *bottom-up*-Anteil enthalten sollen, ist denkbar, dass Formen der Wahrnehmung in die Urteilsfindung einbezogen werden können. Hierfür ist dann – unabhängig vom tatsächlichen Vorliegen der jeweiligen Qualitäten – kein Nachweis subjektiver Eindrücke mehr notwendig.

5.2.6 Intelligenz

Intelligenz in technischen Systemen

Eine repräsentative Umfrage der Bertelsmannstiftung belegt eine Grundskepsis von Menschen in Deutschland gegenüber Technik, insbesondere aktueller Entwicklungen, wie sogenannter *intelligenter* Algorithmen (vgl. Fischer und Petersen 2018).

1 221 Personen ab 16 Jahren wurden „in persönlichen Interviews mündlich (‚face to face')" befragt (Fischer und Petersen 2018: 11). Ein Ergebnis:

> In Deutschland herrscht ein erhebliches Unbehagen in allen Gesellschaftsschichten, wenn es um Algorithmen geht, die über Menschen urteilen und Entscheidungen über sie treffen. Eine große Mehrheit (79 Prozent) zieht menschliche Entscheidungen automatisierten vor. Die Abneigung gegenüber Algorithmen ist umso höher, je folgenreicher die Entscheidung ist.

Nur wenige der Befragten haben angegeben, dass Sie wissen, was Algorithmen sind. So fiel 45 Prozent der Teilnehmenden spontan nichts zum Begriff *Algorithmus* ein (vgl. ebd.: 13). 56 Prozent gaben an, kaum etwas über Algorithmen zu wissen, und nur 10 Prozent gaben an, „recht genau" zu wissen, was es damit auf sich hat.

> Zudem ist den Deutschen in vielen Anwendungsgebieten gar nicht bewusst, dass dort Algorithmen eingesetzt werden. Zwar weiß etwa die Hälfte der Befragten, dass Algorithmen auf den Einzelnen zugeschnittene Werbung (55 Prozent) und Nachrichten (49 Prozent) zuspielen. Weniger bekannt sind potenziell folgenreichere Anwendungsbereiche: Bei der Vorauswahl von Bewerbern oder Krankheitsdiagnosen weiß nur etwa ein Drittel der Befragten, dass dort Algorithmen zum Einsatz kommen (35 bzw. 28 Prozent).
>
> (Ebd.: 6)

Dieses mangelnde Wissen über zuweilen *intelligent* genannte technische Systeme und Algorithmen (Stichwort *künstliche Intelligenz*) wird gemäß der Umfrage umso größer, je älter die Menschen werden. Gleichzeitig scheinen viele sehr konkrete Vorstellungen über Algorithmen und ihre Wirkmächtigkeit zu haben – und zwar in erster Linie negative. Dies bestätigt auch eine an Science-Fiction-Szenarien angelehnte Medienanalyse von Lisa Meinecke und Laura Voss. Die Autorinnen stellen eine reziproke Beeinflussung von „real life robotics" und Science-Fiction-Narrativen, die Formen von Robotern oder künstlicher Intelligenz enthalten, fest. Dies gelte auch für Menschen, die keine Berührungspunkte mit Robotern oder KI haben.

> The influence of the described narratives on how real robots are perceived in society cannot be underestimated. Even without ever having interacted with a real robot, most laypeople have strong expectations about what robots look like, and what they are or should be able to do—and these expectations are heavily influenced by science fiction [. . .].
>
> (Meinecke und Voss 2018: 208)

(Positive oder negative) Narrative über technische Systeme können dabei nicht nur die subjektive Wahrnehmung beeinflussen, sondern auch organisatorische und

infrastrukturelle Folgen mit sich bringen, die nicht auf reale technische Entwicklungen zurückzuführen sind. Dies zeigt beispielsweise eine Untersuchung von Jascha Bareis und Christian Katzenbach, die KI-Strategien von Deutschland, Frankreich, den USA und China unter die Lupe nimmt.

> With their national AI strategies, governments combine the narrative establishment of a particular moment in time that demands leadership (The Narratives of National AI Strategies: Talking AI into Being section) with steering toward particular, country-dependent pathways (AI for Humanity and a Cybernetic Control System: Different Imaginaries subsection). Hence, national leaders seek to convert a field of lofty rhetoric, contingencies, and insecurities into a concrete path of action, aiming at the implementation of their policies through the performance of responsible intervention and leadership. By allocating substantial funding for AI research and business development, establishing normative principles and hard regulation, they constitute the crucial hinge where ideas, announcements, and visions start to materialize in projects, infrastructures, and organizations. Thus, the national AI strategies mark the departure point for country-specific trajectories, driving a process of closure for the integration of AI into society. This creates a process of path dependency that might even lead to lock-in effects down the road.
>
> (Bareis und Katzenbach 2021)

Hans Lenk und Günter Ropohl gingen zu ihrer Zeit primär vom direkten Einfluss technischer Entwicklungen aus: „[. . . N]ur weniges blieb unerreicht. Im Gegenteil: Oft genug wurde die Phantasie von der technischen Wirklichkeit überboten." (Lenk und Ropohl 1993a: 5) Was sich im Spezifischen für das Gefühl potenzieller Bedrohung durch (wie auch immer definierte) *intelligente* Roboter in medialen Darstellungen sagen lässt (vgl. hierzu auch Bruckenberger et al. 2013), gilt auch allgemein für jedweden Medienkonsum. Im medienanthropologischen Sinn prägten Manfred Pirner und Matthias Rath den Begriff des *homo medialis*, also des genuin medial verfassten Menschen (Pirner und Rath 2003). Rath (2014: 78) spitzt diese These zu, indem er *Medialität* von Menschen, „nicht [. . .] als realen Umgang mit Medien" versteht, „sondern als Selbstverständnis des Menschen als eines generell nur vermittelt Welt erfassenden Wesens". Folglich war

> [. . .] jede Epoche, jedes Zeitalter [. . .] in diesem Sinne ‚medial'. Daher sind auch Definitionen wie Wissensgesellschaft oder Mediengesellschaft wenig aussagekräftig. [. . .] Medialität umgibt uns so weit, dass keine Kommunikation als nicht medial gedacht werden kann.
>
> (Ebd.: 87)[108]

[108] Inwiefern Technik als Medium verstanden werden kann, untersucht Christoph Hubig (2021). Eine ungekürzte Fassung des Textes findet sich in der ersten Auflage des Handbuchs (2013).

Was genau als *intelligentes* Verhalten aufgefasst wird, variiert schon bei Definitionen von Menschen und Tieren stark – nicht nur zwischen, sondern auch innerhalb einzelner Disziplinen (vgl. hierzu auch Heil 2021; Seng 2019a).

> Die Grundfrage der Künstlichen Intelligenz läuft darauf hinaus, ob und mit welchen Gründen man den Geräten und Programmen lediglich die Imitation, die Simulation oder die genuine Generation von kognitiven Prozessen aktuell oder prospektiv zubilligen kann. Die Fragestellung ist zumindest sinnvoll, wenn man die Geräte und Programm [sic!] als Instrumente des mittelbaren oder unmittelbaren Handelns ansieht. Sie sollen Prozesse und Gegebenheiten in der realen Welt zweckgebunden beeinflussen können, also wirken, wie dies der tätige Mensch auch tut.
>
> (Kornwachs 2021: 7)

Das Dartmouth College der 1950er-Jahre und seine Forschenden rund um Intelligenz und Lernen wird von Russell und Norvig (2010: 17) als „official birthplace of the field“ künstlicher Intelligenz eingeordnet. In einem Konferenzbericht einer Tagung in Dartmouth, die 1956 stattfand, wurde das Vorhaben der Initiator∗innen (John McCarthy, Marvin Minsky, Claude Shannon und Nathaniel Rochester) so beschrieben:

> We propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.
>
> (Zitiert nach ebd.)

Gemäß Stuart Russell und Peter Norvig war dies das erste Mal, dass John McCarthy den Ausdruck *artificial intelligence* benutzte, wenngleich sie anmerken, dass der Ausdruck „computational rationality“ das Forschungsvorhaben wohl besser getroffen hätte.[109] Die Autoren schlagen vier mögliche Herangehensweisen an das Verständnis künstlicher Intelligenz vor, indem das Verständnis gleichgesetzt werden kann mit dem Versuch, Maschinen oder Programme zu entwickeln, die 1. auf menschliche Art handeln können (*act humanly*), 2. auf menschliche Art denken können (*think humanly*), 3. rational denken können (*think rationally*), 4. rational

[109] Russell und Norvig 2010: 17, Fußnote 10.

handeln können (*act rationally*).[110] Zu allen vier Ansätzen habe es in der Vergangenheit Forschungsansätze von verschiedenen Personen gegeben. Unter dem ersten Ansatz verbuchen sie den Turing-Test, unter dem zweiten Versuche der Modellierung von Kognition; zum dritten Ansatz zählen für sie Versuche, *Gesetze des Denkens*, im Wesentlichen Logik, in Programmen abzubilden, und zum vierten Überlegungen zu rationalen, maschinellen Akteuren (vgl. Russell und Norvig 2010: 2–5).

Die wichtigsten Bereiche, in denen KI eingesetzt wird, waren für Russell und Norvig vor zehn Jahren (teil-)automatisiertes Fahren, Spracherkennung, automatisierte Planung (z. B. für Mars-Rover-Operationen der NASA), Spiele, Spamfilter, Logistik, Haushaltsroboter (zum Beispiel automatisierte Staubsauger) und automatisierte Übersetzungen (ebd.: 28 f.). Darüber hinaus wird inzwischen häufig der Teilbereich des *maschinellen Lernens* als das *Intelligente* an aktuellen Programmen und Algorithmen verstanden, kurz gesagt: Mustererkennung.

Intelligenz als notwendige Voraussetzung für ethische Reflexion?

Reinhard Heil (2021: 425) weist darauf hin, dass ein sehr weit ausgelegter Intelligenzbegriff dazu führt, dass auch „Schleimpilzen“ Intelligenz zugesprochen werden müsse. In Bezug auf das Alltagsverständnis von Intelligenz resümiert Heil: „Weder Schleimpilze noch aktuelle Künstliche Intelligenzen sind im alltagssprachlichen Sinne intelligent“ (ebd.). Gleichwohl seien beide „im Umgang mit bestimmten Problemen enorm leistungsfähig“ und von diesen Problemen sollte man durchaus annehmen, „dass zu ihrer Lösung kognitive Fähigkeiten notwendig sind“ (ebd.). Günter Ropohl (1991: 155) gesteht gemäß Kornwachs Einordnung (s. o.) Programmen Formen von Intelligenz-Simulationen zu, indem er ihnen „intelligenzanaloge Leistungen“ zuspricht – „jedenfalls hinsichtlich bestimmter Funktionen, die früher zur Domäne menschlicher Intelligenz gerechnet wurden“. Man müsse folglich „die Redeweise von der künstlichen Intelligenz mindestens als Metapher zu akzeptieren haben“, auch wenn es sich lediglich um eine Behauptung von Ähnlichkeiten zu menschlicher Intelligenz handele.

[110] In aktuellen philosophischen Debatten um Rationalität wird diese jedoch zunehmend kontrovers diskutiert. Sebastian Schmidt argumentiert zum Beispiel für mehr Irrationalität in bestimmten Situationen (Schmidt 2017) und die Bedeutung von Verantwortung im Kontext von Rationalitätsdebatten (Schmidt 2020).

Ähnlich wie beim Begriff *Bewusstsein* macht die Heterogenität der Definitionen von Intelligenz den Begriff als notwendige Voraussetzung für ethische Reflexion schwierig. Mit Sicherheit können Aspekte der verschiedenen Definitionen als relevant für ethische Reflexionen, das Abwägen von Argumenten und die Bildung von Theorien verstanden werden; jedoch kann dies nur speziell für eine jeweilige Intelligenz-Definition geleistet werden und nicht pauschal. Wird unter *Intelligenz* beispielsweise eine Formen logischen Denkens verstanden, ist es plausibel, dieses als notwendigen Teil ethischer Reflexionen anzuerkennen.

5.2.7 Intentionalität

Intentionalität ist neben Bewusstsein einer der meistdiskutierten Begriffe im Kontext mentaler Fähigkeiten und der Möglichkeit der Übertragung auf technische Systeme. In der philosophischen Handlungstheorie werden Intentionen u. a. beschrieben als mentale Repräsentationen, also Gedanken, Wünsche, Äußerungen usw. über die Welt. Pierre Jacob (2019) beschreibt Intentionen als

> [...] the power of minds and mental states to be about, to represent, or to stand for, things, properties and states of affairs. To say of an individual's mental states that they have intentionality is to say that they are mental representations or that they have contents.

In vielen Einführungstexten wird der Philosoph und Psychologe der Scholastik Franz Brentano (2008) als Urheber des Ausdrucks *Intentionalität* angeführt. Peter-Paul Verbeek (2011: 55) beschreibt Intentionalität als die Fähigkeit, Absichten zu formulieren, die mit einer „directedness of human beings toward reality" einhergehe. Intentionen können dabei auch die Form von Wünschen oder Meinungen annehmen und werden manchmal auch als *Wille* (nicht im kantianischen Sinn des Wortes) oder *Wollen* bezeichnet. Im Zusammenhang mit Intentionen sind diese Formulierungen wohl weniger gemäß dem zeitgenössischen Verständnis der Ausdrücke zu verstehen, sondern im Sinne eines „Worumwillens" (Quante 2020: 77), also eines Ziels oder Zwecks von Handlungen. Dabei geht es bei der philosophischen Klärung des Begriffs *Handlung* durchaus auch um Aspekte wie *Wollen* und *Wissen*, also um die Fragen, ob eine Handlung wissentlich und / oder absichtlich durchgeführt wurde. Nur ein Beispiel von vielen möglichen Konstellationen:

> In unserer ethischen und rechtlichen Praxis der Bewertung von Handlungen macht es einen Unterschied, ob A eine Wirkung e_2 seiner Handlung h_1 zwar absichtlich herbeiführt, weil er davon Kenntnis hat, e_2 aber gar kein Ziel seiner Handlung ist. Oder ob A seine Handlung h_1 deshalb ausführt, weil er e_2 realisieren will. Im ersten Fall nimmt A e_2 nur in Kauf, im zweiten Fall ist e_2 dagegen das (oder zumindest ein Teil vom) Ziel (oder der Zweck, etwas altertümlich auch Worumwillen) seiner Handlung h_1.
>
> (Quante 2020: 77)

Intentionalität als notwendige Voraussetzung für ethische Reflexion?

Intentionen werden in der philosophischen Handlungstheorie als eine mögliche Form von Handlungserklärungen (neben Kausalrelationen) angesehen. Als Gegenstand von ethischen Bewertungen sind Intentionen allerdings als Einstellungen und damit als Begründung für Handlungen zu verstehen: „Ich zog meinen Hut, weil es in Gotteshäusern anstößig ist, mit einer Kopfbedeckung einzutreten“ (ebd.: 99). Definiert über die Funktion der Handlungsbegründungen sind Intentionen notwendig für ethische Reflexionen, da sich die Abwägung ethischer Argumente, wie bereits in den vorherigen Abschnitten gezeigt wurde, an geltenden Normen orientieren sollte.

Ansgar Beckermann grenzt intentionale Zustände von (körperlichen) Empfindungen ab; zu ersteren gehören demnach auch Wahrnehmungseindrücke und körperliche Sinneserfahrungen.

> *Intentionale Zustände* wie Überzeugungen, Wünsche, Befürchtungen und Erwartungen sind demgegenüber eben durch ihre Intentionalität, d. h. dadurch charakterisiert, daß sie *auf etwas gerichtet* sind, daß sie einen *Inhalt* haben. Man glaubt, *daß etwas der Fall ist*, man wünscht sich *einen bestimmten Gegenstand*, man hofft oder befürchtet, *daß ein bestimmtes Ereignis eintreten wird.*
>
> (Beckermann 2001: 13. Herv. i. Orig.)

Nach Beckermann gibt es „erhebliche Unterschiede“ innerhalb der intentionalen Zustände. In Bezug auf menschliche Intentionen werden häufig Fähigkeiten wie Sprachverständnis oder Freiheit (s. Willensfreiheit oben) als notwendige Voraussetzungen erachtet, um Wünsche äußern oder sich bestimmte Dinge vorstellen zu können. Manchmal gilt aber auch Intentionalität als Voraussetzung für andere mentale Fähigkeiten. Franz Brentano vertrat beispielsweise die These, dass Intentionalität Teil aller mentalen Zustände ist (Brentano 2008). Dies wurde von Philosoph∗innen im Lauf der Zeit aber infrage gestellt (vgl. u. a. Jacob 2019; Ros 2005b). Ansgar Beckermann (2001: 14) unterscheidet primär kognitive Einstellungen, „wie Überzeugungen“, von Einstellungen, „die auch eine konative [... Antrieb zum Handeln; L. S.] oder affektive Komponente haben, wie Wünsche, Absichten und Befürchtungen“. Diese Einteilung ist in der Literatur jedoch keineswegs Konsens (Jacob 2019).

Intentionen können also insofern als Voraussetzungen für ethische Reflexionen betrachtet werden, dass bestimmte Vorstellungen von moralisch richtig und falschen Einstellungen und Handlungen vorhanden sein müssen beziehungsweise dass die Fähigkeit gegeben sein muss, sich Gedanken über moralische Inhalte zu machen.

Intentionalität in technischen Systemen

In der Philosophie des Geistes wurden zahlreiche Vorschläge gemacht, wie das Zusammenspiel von Körper und Geist, das auch im Hinblick auf mögliche Realisierungen mentaler Fähigkeiten und Eigenschaften in technischen Systemen interessant sein kann, zu erklären sei. Ein Grund, warum Ansgar Beckermann (s. o.) die Unterscheidung zwischen körperlichen Empfindungen und mentalen Zuständen herausstellt, ist der, dass sich gemäß der klassischen Auffassung des Leib-Seele-Problems[111] körperliche Zustände eindeutig durch einen „*qualitativen Charakter*“ (Beckermann 2001: 13. Herv. i. Orig.) auszeichnen, wohingegen die Qualität von mentalen Zuständen oder Eigenschaften (auch *Qualia* genannt) nicht so einfach zu bestimmen ist. Aussagen über körperliche Qualitäten können mittlerweile recht gut mit naturwissenschaftlichen Methoden erfasst, gemessen und verglichen werden, wohingegen mentale Eigenschaften wissenschaftlich weitaus schwieriger zu greifen sind – sofern sie nicht mittels eines reduktionistischen Ansatzes indirekt über körperliche Eigenschaften definiert werden.

Die groben Strömungen, die allen Theorien der Philosophie des Geistes zugrundeliegen, sind Dualismus und Monismus. Bei Ersterem geht man von zwei ontologisch voneinander getrennten Substanzen, dem Mentalen und dem Körperlichen, aus; bei Letzterem gibt es verschiedene Erklärungsformen, warum die erste Annahme nicht stimmen kann. Postuliert wird hier zum Beispiel die Identität mentaler und physischer Substanzen und Eigenschaften (Identitätstheorie), die Erklärung mentaler Eigenschaften als Funktion physischer Eigenschaften (Funktionalismus) oder Supervenienz (Supervenienz).[112] Interessant für mögliche Realisierungen mentaler Eigenschaften in technischen Systemen (s. Kapitel 6) sind dabei vor allen Dingen die reduktionistischen Ansätze, da diese versuchen, mentale Eigenschaften auf verschiedene Weisen über physikalische zu erklären.

Auch die Frage, ob Maschinen oder Programme über Geist, mentale Eigenschaften und kognitive Fähigkeiten verfügen können, steht in einer langen Traditi-

[111] Damit ist die Frage gemeint, wie prinzipiell körperliche und prinzipiell geistige Eigenschaften innerhalb eines Menschen miteinander vereinbar sein können.

[112] Der Ansatz geht unter anderem auf Donald Davidson zurück, demzufolge geistige auf physikalischen Eigenschaften supervenieren, wenn es „keine zwei Ereignisse geben kann, die in allen physischen Hinsichten gleich, aber in einer geistigen Hinsicht verschieden sind, oder daß sich kein Gegenstand in einer geistigen Hinsicht ändern kann, ohne sich auch in einer physischen Hinsicht zu ändern“ (Davidson 1970: 301, zitiert nach Beckermann 2001: 204 f.).

on.[113] Eines der bekanntesten Argumente gegen die mentalen Eigenschaften oder Geist (*mind*) von Maschinen brachte 1980 John Searle in einem Aufsatz hervor (vgl. Searle 1980). Der Artikel und das darin enthaltene Argument in Form eines Gedankenexperiments ist unter anderem eine Antwort auf Alan Turings Idee denkender Maschinen (s. o.). Searle entgegnet dem mit seinem *Chinese-Room*-Experiment, mit dem er zu zeigen versucht, dass das Bestehen des Turing-Testes für ein Programm nicht als Evidenz für Verstehen ausreiche.

> Nehmen wir an, ich bin in einem Raum eingeschlossen, und man gibt mir einen Packen mit chinesischer Schrift. Nehmen wir weiter an, daß ich (was in der Tat der Fall ist) kein Chinesisch kann, es weder schreiben noch sprechen kann, und daß ich nicht einmal sicher bin, ob ich chinesische Schrift als chinesische Schrift erkennen und von, sagen wir, japanischer Schrift oder sinnlosem Gekritzel unterscheiden könnte. Chinesische Schrift besteht für mich einfach nur aus sinnlosem Gekritzel. Nehmen wir nun weiterhin an, daß man mir nach dem ersten Packen mit chinesischer Schrift einen zweiten Packen mit chinesischen Schriftzeichen gibt, zusammen mit einer Reihe von Anleitungen, wie ich den zweiten Stoß zum ersten in Beziehung setzen soll. Die Anleitungen sind in Englisch abgefaßt, und ich verstehe diese Anleitungen ebenso gut wie jeder andere, dessen Muttersprache Englisch ist. Sie ermöglichen es mir, eine Reihe formaler Symbole zu setzen, und ‚formal' bedeutet hier nichts weiter, als daß ich diese Symbole ausschließlich an Hand ihrer Form identifiziere. Nehmen wir nun auch noch an, man gibt mir einen dritten Packen chinesischer Symbole, zusammen mit einigen Anweisungen, ebenfalls in Englisch, die es mir ermöglichen, Teile dieses dritten Packens in Beziehung zu setzen zu den zwei ersten Packen; und diese Anleitungen weisen mich an, bestimmte Symbole mit bestimmten Formen in Antwort auf bestimmte Formen, die mir mit dem dritten Packen zugegangen sind, zurückzugeben. Was ich nicht weiß ist, daß die Leute, die mir all diese Symbole geben, den ersten Packen ‚Schrift', den zweiten Packen eine ‚Geschichte' und den dritten Packen ‚Fragen' nennen. Des weiteren nennen sie die Symbole, die ich ihnen in Antwort auf den dritten Packen zurückgebe, ‚Antworten auf die Fragen', und die Reihe von englischsprachigen Anleitungen, die sie mir geben, nennen sie ‚Programm'. Nun stellen wir uns, einfach um die Sache noch ein bißchen zu komplizieren, vor, daß diese Leute mir auch Geschichten in Englisch geben, die ich verstehe, und daß sie mir dann in Englisch Fragen zu diesen Geschichten stellen und ich ihnen in Englisch antworte. Nehmen wir auch an, daß ich nach einer gewissen Zeit den Anweisungen für das Hantieren mit den chinesischen Symbolen so gut zu folgen lerne und die Programmierer so gut lernen, Programme zu schreiben, daß, von außen betrachtet – d. h. vom Standpunkt eines Menschen aus, der sich außerhalb des Raumes befindet, in dem ich eingeschlossen bin –, meine Antworten auf die Fragen absolut ununterscheidbar sind von denen, die einer geben würde, dessen Muttersprache Chinesisch ist. Niemand, der nur meine Antworten sieht, kann erkennen, daß ich kein Wort Chinesisch spreche. Nehmen wir auch an,

[113] Vgl. z. B. Dupuy 2001; Hancock 2017 sowie die Artikel der seit 1991 bis heute erscheinenden, wissenschaftlichen Zeitschrift *Minds and Machines. Journal for Artificial Intelligence, Philosophy and Cognitive Science.*

> daß meine Antworten auf die englischen Fragen ununterscheidbar sind von denen, die andere geben würden, deren Muttersprache Englisch ist, was ohne Zweifel der Fall sein wird, einfach deshalb, weil Englisch auch meine Muttersprache ist. Von außen gesehen – vom Standpunk dessen betrachtet, der meine ‚Antworten' liest – sind die Antworten auf die chinesischen Fragen und auf die englischen Fragen gleich gut. Aber im Fall des Chinesischen bringe ich, anders als im Englischen, die Antworten dadurch hervor, daß ich mit unverstandenen formalen Symbolen hantiere. Soweit es das Chinesische betrifft, verhalte ich mich einfach wie ein Computer; ich führe kalkulatorische Operationen an formal spezifizierten Elementen aus. In bezug [sic!] auf das Chinesische bin ich einfach ein verkörpertes Computerprogramm.
>
> (Searle 1994: 234–236)

Eine Maschine oder ein Programm kann demnach zwar den richtigen Output generieren, also beispielsweise die korrekten chinesischen Schriftzeichen als Antwort auf eine Frage wiedergeben (unter Zuhilfenahme eines chinesischen Wörterbuchs), allerdings bedeute das noch lange nicht, dass sie oder es *verstehe*, was sie oder es da tue oder von sich gebe. Der Turing-Test könne demnach auch völlig ohne Intelligenz bestanden werden (vgl. hierzu auch Wallach und Allen 2009: 57). Anders formuliert, lautet das Argument so: Maschinen verarbeiten Daten, das Gehirn verarbeitet Informationen und erzeugt damit Wissen (vgl. Diamant 2015 zitiert nach Kornwachs 2021: 26).

> Wenn Searle recht hat, dann kommen Maschinen keine echten mentalen Zustände zu. Sie könnten jedoch Zustände aufweisen, die funktional vergleichbar mit mentalen Zuständen sind. Im Unterschied zu echten mentalen Zuständen sind sie jedoch nicht mit phänomenalem Bewusstsein verbunden und haben nur abgeleitete Intentionalität, vermittelt durch ihre Programmierer und Nutzer. Man könnte solche Zustände als *Quasi-Meinungen*, *Quasi-Wünsche* usw. bezeichnen.
>
> (Misselhorn 2019: 86)

Günter Ropohl (1991: 157) argumentiert, ähnlich wie John Searle, dass es bei intelligenten Lösungen nicht nur auf den Output ankommt:

> Während Menschen ein Problem mit subjektiver Intuition erfassen, muß für den Computer das Problem objektiviert werden; Wissen und Nicht-Wissen müssen intersubjektiv präzisiert werden.

5.3 Zwischenergebnis

Auf Seite 98 stellte ich die Fragen, ob die jeweils angesprochenen Qualitäten, Fähigkeiten und Werte notwendige Voraussetzungen sind für ethische Reflexionen und Begründungen (1.) und ob sie auch für das hypothetische Konzept ethischer Techniksysteme angenommen werden können (2.). Die Untersuchung der ausgewählten

Ansätze und Konzepte hat ergeben, dass einige Begriffe einfacher zu definieren sind als andere. Eine andere Strategie – neben der Darstellung der verschiedenen Konzepte zu jeweils einem Begriff – wäre gewesen, sich auf bestimmte Ansätze oder Theorien zu beschränken. Ziel der synoptischen Untersuchung war es jedoch, die vielfältige Interpretierbarkeit aufzuzeigen, die keine pauschalen Antworten ermöglicht, sondern konkrete Abwägungen in konkreten Anwendungsfällen.

Es stellte sich heraus, dass die Voraussetzung von Bewusstsein, Emotionen und Intelligenz sowohl für moralische als auch für ethische Maschinen schwierig sind, da sich bei diesen Begriffen bereits definitorische Schwierigkeiten in Bezug auf Menschen ergeben. Natürlich sind auch die Definitionsspektren von Freiheit, Autonomie, Rationalität und Intentionalität groß und auch diese Begriffe müssen, wie gezeigt, kontextabhängig verstanden werden. Jedoch hat die Analyse gezeigt, dass die Methoden zur Definition der Letztgenannten mehr Klarheit versprechen. Die Zuschreibung von Freiheit, Autonomie, Rationalität und Intentionalität an technische Systeme setzt jedoch jeweils sehr weite Begriffsverständnisse voraus.

Ein ethisches Techniksystem bedarf demnach einer gewissen Freiheit, um Entscheidungen treffen und Handlungen ausüben zu können; es muss zum Beispiel aus verschiedenen Verfahrensoptionen auswählen können, was in einem weiten Verständnis schon unter den Begriff der Freiheit fällt. Es bedarf zudem einer gewissen Autonomie, da hierin gerade der Unterschied zu moralischen Maschinen besteht, die auf technisch vorprogrammierten Automatismen beruhen – auch innerhalb von Methoden maschinellen Lernens. Eine ethische Maschine ist in gewisser Hinsicht rational, insofern sie sich an die vorprogrammierten Regeln hält, auch wenn diese die Freiheit des Lernens und somit ungewisse Outputs beinhalten. Rationalität in diesem Sinn ist nicht mit Moralität zu verwechseln, da ein technisches System zwar gemäß seinen internen Gesetzen und somit *rational* agieren, gleichwohl der Output aber als unmoralisch beurteilt werden kann; dies ist bei den meisten KI-Programmen der Fall, die aktuell unter dem Stichwort *KI-Governance* kritisiert werden. Intentionale Zustände hat ein technisches System insofern, dass es aufgrund bestimmter Zustände auf eine Art und Weise agiert, die von außen sichtbar und bewertbar ist. Dies kann in der Form unterschiedlicher physikalischer Realisierungen geschehen. Diese Argumentation setzt voraus, dass ein gewisser Grad an Freiheit, Autonomie, Rationalität und Intentionalität ebenfalls für ethische Reflexions- und Argumentationsfähigkeit angenommen werden kann.

5.3.1 Menschen und technische Systeme – Analogie oder mögliche Identität?

Wie die Untersuchung einiger Ansätze zu zentralen Begriffen im Zusammenhang der Voraussetzungen ethischer Maschinen in den vorangegangenen Abschnitten gezeigt hat, besteht eine Strategie darin, eine prinzipielle Ununterscheidbarkeit von

Begriffen wie Freiheit, Autonomie und Intentionalität zwischen Menschen und Maschinen zu behaupten. Diese Behauptung habe ich *Identitätsbehauptung* genannt (S. 11 f.). Diese steht der *Analogiebehauptung* gegenüber, wonach Maschinen Entscheidungsfreiheit und Intentionalität nicht im eigentlich menschlichen Sinn haben, gleichwohl aber mit Misselhorn eine Form *quasi-intentionaler Zustände* anzunehmen ist. Im Folgenden werde ich für beide Ansätze Argumente vorbringen und begründen, warum ich für die vorliegende Arbeit und genuin ethische Techniksysteme von einer prinzipiellen Analogie ausgehe.

Die Annahme, Maschinen können *genauso* sein wie Menschen, sowohl was ihre physische Erscheinung angeht als auch ihre psychische Beschaffenheit, wird auch als *starke KI*-These bezeichnet (vgl. Seng 2019a). Eine Verwechslung mit tatsächlichen Entwicklungen im Bereich künstlicher Intelligenz, derzeit aktuell besonders im Fokus maschinellen Lernens, kann zu Verwirrungen führen (vgl. u. a. Heil 2021). In Fachkreisen sämtlicher Disziplinen um künstliche Intelligenz und Algorithmen herrscht dabei die weitgehend einhellige Meinung, dass *starke KI* in absehbarer Zeit nicht realistisch ist.[114] Eine mögliche Begründung für diese These habe ich bereits angeführt: Moralische Maschinen sind technische Systeme, die zu einem bestimmten Zweck konstruiert wurden; und selbst wenn sie nicht so funktionieren, wie Entwickler∗innen es vorhergesehen haben, dann kann beispielsweise ein Chatbot immer noch nicht von sich aus entscheiden, auf einmal Schach zu spielen, wenn er dazu entwickelt wurde, auf digitalen Plattformen zu kommunizieren. Zwar gibt es gewisse Unvorhersehbarkeiten in den Aktionen moralischer Maschinen und Programmen, weshalb sie immer wieder auch als *black boxes* bezeichnet werden und in Diskussionen um KI-Governance der Ruf nach Transparenz lauter wird (vgl. z. B. Europäische Kommission 2019: 10, 16 f., 22 ff.). Dabei lassen sich die Prozessschritte zwar anhand des Programmcodes nachvollziehen, es bleibt jedoch häufig die Frage, *warum* ein lernendes Programm diesen oder jenen Weg gegangen ist. Dies stellt einen prinzipiellen Unterschied zu weniger komplexen technischen Systemen dar.

> Dass es für Anwender schwer bis unmöglich nachzuvollziehen ist, wie ein KI-System genau zu einem Ergebnis gekommen ist, ist an sich wenig problematisch, da viele Technologien für Laien prinzipiell kaum durchschaubar sind. Bereits eine einfache Wechselschaltung, die es erlaubt, einen Verbraucher von zwei Orten aus an- und abzuschalten, dürfte für die wenigsten intuitiv nachvollziehbar sein. Die epistemische Opazität erreicht bei maschinellen Lernverfahren jedoch eine andere Qualität, da sie nun auch Fachleute betrifft. Können sich Laien bei herkömmlichen Technologien mehr oder weniger darauf verlassen, dass es Fachleute gibt, die nachvollziehen kön-

[114] Ron Chrisley (2020: 465 ff.) bezeichnet dies auch als *Deflationary View*; vgl. hierzu auch Braga und Logan 2017.

> nen [sic!] warum ein System zu einem bestimmten Ergebnis gelangt ist oder warum ein Fehler aufgetreten ist, so gilt dies für komplexe lernende Systeme nicht mehr in allen Fällen. Dies hat nicht nur Auswirkungen darauf, ob einer Technik (und Expertinnen) vertraut wird, wovon deren Akzeptabilität maßgeblich mitbestimmt wird, sondern ist auch von Relevanz [sic!] wenn Ergebnisse nicht den Erwartungen entsprechen, Fehler auftreten und Schuldfragen geklärt werden müssen.
> (Heil 2021: 426)[115]

Die Unvorhersehbarkeit intransparenter Programmcodes besteht immer noch innerhalb eines gewissen Rahmens, der von Menschen durchaus vorhersehbar ist. Chatbot Tay wurde dazu programmiert, zu kommunizieren – nur, *was* er auf der Basis seiner Konzeption kommuniziert und *in welcher Weise* er durch das maschinelle Lernen beeinflussbar und manipulierbar ist, wurde nicht in Gänze vorhergesehen – denn sonst hätte man das Programm vermutlich nicht in der Öffentlichkeit getestet oder von vorneherein eine Diskriminierungs- und Rassismus-Warnung mitgeliefert.

Technische Systeme können sich also (bislang) ihren Zweck nicht selbst geben, zu dem sie gemacht wurden. Menschen können dies, wie ich unten zeigen werde, auch in eingeschränkterer Form, als man häufig anzunehmen gewillt ist. Im Rahmen ihrer Einschränkungen (ohne dabei von einem starken Determinismus auszugehen) sind Menschen aber immer noch eigenständiger als technische Systeme.

Der Identitätsthese steht die Analogiebehauptung gegenüber. Demnach kann man technischen Systemen zwar in gewisser Hinsicht Freiheit, Autonomie, Rationalität und Intentionalität – ein breites Verständnis dieser Begriffe vorausgesetzt – zuschreiben, aber nur in *ähnlicher* Weise wie bei Menschen. Um die Analogie auf der Ebene der Bezeichnung zu verdeutlichen, kann man auch von Freiheit$_T$, Autonomie$_T$, Rationalität$_T$ und Intentionalität$_T$ bei Maschinen und Programmen sprechen.[116] Eine Variante ist, in Anlehnung an Gerhard Ernsts Definition von Moral, von Freiheit und Autonomie im weiteren und im engeren Sinn zu sprechen, wobei die Begriffe, im weiteren Sinn verstanden, die technischen Versionen miteinbeziehen würden. Thomas Powers (2011: 46) beschreibt in diesem Sinn die Übertragung *praktischer Vernunft* auf Programme:

> Human practical reasoning primarily concerns the transformation between the consideration of facts and the ensuing action. To some extent, the transformation re-

[115] Der Ausdruck *Explainable AI* gewinnt als Forschungsgegenstand daher zurzeit zunehmend an Interesse (vgl. z. B. Barredo Arrieta et al. 2020) und wird von Unternehmen häufig als Aushängeschild mittels selbstgemachter Ethik-Richtlinien benutzt (vgl. u. v. a. IBM 2021).

[116] Das tiefgestellte T steht jeweils für *technisches System*; natürlich kann man auch andere Abkürzungen wählen.

> sembles a machine's state changes when it goes from a set of declarative units in a database to an output. There are other similarities, of course – humans can learn new facts that inform their reasoning about action, just as machines can incorporate feedback systems that influence their outputs.

Für Klaus Kornwachs (2021: 7) ist das, wie oben bereits zitiert, die „Grundfrage der Künstlichen Intelligenz" (s. S. 121). Mit der Identitäts- und der Analogiebehauptung wird also auch die eingangs gestellte Frage (s. S. 17) wieder ins Zentrum gerückt, ob Maschinen, Programme und Roboter – sowohl die, die es bereits gibt, als auch potenzielle Weiterentwicklungen – Simulationen von Menschen sind bezüglich deren kognitiver und anderer Fähigkeiten – oder, ob es Definitionsmöglichkeiten gibt, die eine Identität sinnvollerweise erlauben.

Der Blick in die Geschichte der Technikentwicklung in Kapitel 3 hat die Vermutung nahe gelegt, dass technische Systeme bezüglich ihrer Fähigkeiten (Sprache u. a.) immer stärker an Menschen angepasst werden und in vielen Bereichen auch akzeptierter Weise besser sind als Menschen. Die Analyse der kognitiven Qualitäten hat die Vermutung belegt, dass der Vergleich aufgrund disziplinärer Definitionsschwierigkeiten nicht leicht ist. Die Frage, wohin sich technische Systeme bezüglich ihrer ethischen Reflexionsfähigkeiten entwickeln, ist auch im Hinblick auf anthropologische Aspekte der Technikfolgenabschätzung interessant.

Man kann für ethische Techniksysteme zu dem Schluss kommen, dass sich ihre Entscheidungsfreiheit$_T$ nicht grundlegend von der Entscheidungsfreiheit von Menschen unterscheidet, da die Kontexte, in denen Menschen ihre Entscheidungen treffen, häufig ähnlich zu sein scheinen wie die ethischer Maschinen. Ethische Programme, die Argumente in komplexen Situationen im Hinblick auf beispielsweise deutsche oder europäische Grundwerte hin prüfen, könnten darin sogar besser sein als Menschen – denn dieser Zweck liegt der Entwicklung vieler technologischer Systeme, neben wirtschaftlichen und wissenschaftlichen Faktoren, zugrunde (s. hierzu S. 19 und S. 155). Menschen entscheiden – ebenso wie Maschinen – auf der Grundlage bestimmter kontextueller Bedingungen und Erfahrungen. Die Datengrundlage für letztere übersteigt bei Maschinen sogar die von Menschen dank Methoden wie *deep learning* (s. S. 102).

Allerdings handelt es sich bei ethischen Maschinen, so unabhängig sie von Menschen auch sein beziehungsweise ihre Entscheidungen treffen mögen, um Programme, die von Menschen zu einem bestimmten Zweck konzipiert und realisiert wurden. Auch ethische, selbstständig denkende und argumentierende Maschinen unterliegen den Grenzen ihres Programmzwecks. Was also für moralische Maschinen gilt, gilt auch für ethische Techniksysteme. Hier gibt es einige Einwandmöglichkeiten. Man kann argumentieren, dass Menschen auch gewissen Grenzen unterliegen, nämlich denen der Naturkräfte. Sie können auf der Erde nicht auf einmal, wie im All, schwerelos herumfliegen, sofern sie sich nicht in einem Flugsimulator befinden, in dem sie die auf der Erde geltende Schwerkraft ausgetrickst haben. Wenn man

diesem Argument ein fundamentalreligiöses Menschenbild zugrundelegt, das von einem prinzipiellen Determinismus ausgeht, können Menschen noch nicht einmal ihre eigenen Entscheidungen frei treffen, sondern unterliegen den Bestimmungen eines Gottes oder gottähnlicher Mächte. Auch in der Gestaltungsfreiheit ihres Lebens sind Menschen stark eingeschränkt, wenn man ein radikaldeterministisches Weltbild zugrundelegt.

Doch man muss nicht von Extremen ausgehen, um zu zeigen, dass Willens- und Handlungsfreiheit von Menschen nicht so absolut sind, wie man gern glauben möchte. Angesichts vieler soziologischer Studien, die eine Beschränkung in den Entwicklungsmöglichkeiten von Menschen aufgrund bestimmter sozialer und psychologischer Faktoren belegen, kann die These der freien Wahl der Lebensgestaltung, auch ohne einen deterministischen Ansatz zugrunde zu legen, bezweifelt werden. Eine Studie des *AWO Bundesverbands e. V.* von 2019 zeigt beispielsweise:

> Für junge Erwachsene, die aktuell von Armut betroffen sind oder Armut in Kindheit und Jugend erfahren haben, stellt der Übergang [von der Kindheit ins junge Erwachsenenalter; L. S.] häufig eine ungleich höhere Herausforderung dar [als bei Kindern, die unter anderen Umständen aufgewachsen sind; L. S.]. Dies äußert sich etwa in Form einer verzögerten, erschwerten oder (noch) nicht vollzogenen Bewältigung von Entwicklungsaufgaben wie dem Abschluss einer Ausbildung oder die Integration in den Arbeitsmarkt.
>
> (AWO Bundesverband e. V. 2019: 5)

Gemäß der Studie bleibt außerdem ein Drittel „von der Kindheit bis zum jungen Erwachsenenalter arm“ (ebd.). Doch auch wenn es mit Sicherheit sozio-kulturelle Unterschiede in den Entfaltungs- und Gestaltungsmöglichkeiten bei Menschen bezüglich ihres eigenen Lebens gibt, so gibt es eben auch diejenigen, die aufgrund günstiger Umstände verhältnismäßig viel Freiheit genießen. Diese Polarisierung hat auch die Corona-Pandemie verstärkt gezeigt: Während die einen (beispielsweise Künstler∗innen, Familien mit Kindern ohne externe Betreuungsmöglichkeiten oder ältere Menschen u. a.) stark eingeschränkt waren und unter den sozialen und finanziellen Einschränkungen sowie den gesundheitlichen Folgen litten, genossen manch andere die Freiheiten, die Homeoffice und die veränderten Lebensumstände unter Corona mit sich brachten und bringen.

Es gibt also zumindest die Möglichkeit verhältnismäßig großer Willens- und Entscheidungsfreiheit für Menschen unter bestimmten Umständen und dies auch im Rahmen der naturgemäßen Einschränkungen. Selbst wenn man also einen religiösen Maßstab anlegt und somit das *Geschaffensein* als vergleichbare Komponente von technischen Systemen ebenso wie von Menschen festhalten kann, so stehen Menschen in ihren Tätigkeiten und Entfaltungsmöglichkeiten doch mehr Türen offen als technischen Systemen.

Auf der anderen Seite empfiehlt es sich, bei den Voraussetzungen für ethische Reflexionen genau hinzusehen und keine, wie oben so genannten, *künstlichen* Unterscheidungsmechanismen zwischen Menschen und Technik, gemäß einem inhärenten Speziesismus zu postulieren. Dieser Zug wäre allzu einfach und im Hinblick auf z. B. intersubjektiven Austausch (s. folgender Abschnitt) allzu einfach. Aufgrund der unterschiedlichen äußeren Umstände von Menschen und Technik bietet sich zunächst eine analoge Redeweise von Entscheidungen$_T$ und Entscheidungen an. Anhand eines weiteren Arguments soll die Analogiebehautpung allerdings noch einmal detaillierter auf die Probe gestellt werden.

5.3.2 Zum Intersubjektivitätsproblem

Im Anschluss an die oben aufgezeigte Diskussion um die Frage, ob Freiheit oder intentionale Zustände technischen Systemen in analoger oder derselben Weise zugeschrieben werden können wie Menschen, kann noch ein Problem ins Feld geführt werden, das die epistemologische Begrenzung der Erforschung der menschlichen Psyche aufzeigt und somit auch die teilweise erhobenen Ansprüche an technische Systeme infrage stellt.[117] Das Problem der Zuschreibbarkeit intentionaler Zustände sieht wie folgt aus: Sofern man keinen radikalen Reduktionismus anwendet, also Bewusstsein und andere mentale Zustände auf physikalische Zustände reduziert, ist es methodisch (bislang) unmöglich, subjektives Erleben und interne Qualitäten auch bei Menschen nachweisbar zu belegen.[118]

> Presumably, what any computer could or would experience is similarly beyond human's ken. The idea that one cannot have knowledge about other kinds of minds provides reason enough for some philosophers and scientists to doubt whether it even makes sense to talk about consciousness.
>
> (Wallach und Allen 2009: 66)

Die Tatsache, dass die menschliche Psyche einer Person für eine andere zumindest nicht aus der Innenperspektive nachvollziehbar ist, wird auch als Intersubjektivitätsproblem bezeichnet – oder in der Philosophie des Geistes als Qualiaproblem.

[117] Für das Aufzeigen der Dringlichkeit dieses Problems für die Definition des moralischen Status' technischer Systeme im maschinenethischen Diskurs danke ich Matthias Rath. Auf die entsprechenden Textstellen verweise ich im Folgenden.

[118] Eine Ausnahme bilden psychologische Befragungen, wobei auch hier methodische Unsicherheiten nicht ausgeschlossen sind: Probanden können bewusst oder unbewusst falsche Aussagen über innere Zustände, Gefühle und Emotionen machen; sie können (je nachdem, in welchem Kontext die Befragung stattfindet) von der Situation des Versuchsaufbaus beeinflusst sein oder von den Personen, die die Befragung durchführen, denen sie beispielsweise imponieren wollen oder sich schämen, was zu Verzerrungen in solchen Erhebungen führen kann, vgl. z. B. Mey und Mruck 2018b.

Thomas Nagel beschreibt das Problem in einem in der Philosophie des Geistes inzwischen sehr bekannten Artikel anhand der Frage, wie es wohl wäre, eine Fledermaus zu sein.

> It will not help to try to imagine that one has webbing on one's arms, which enables one to fly around at dusk and dawn catching insects in one's mouth; that one has very poor vision, and perceives the surrounding world by a system of reflected high-frequency sound signals; and that one spends the day hanging upside down by one's feet in an attic. In so far as I can imagine this (which is not very far), it tells me only what it would be like for *me* to behave as a bat behaves. But that is not the question. I want to know what it is like for a *bat* to be a bat. Yet if I try to imagine this, I am restricted to the resources of my own mind, and those resources are inadequate to the task. I cannot perform it either by imagining additions to my present experience, or by imagining segments gradually subtracted from it, or by imagining some combination of additions, subtractions, and modifications.
>
> (Nagel 1974: 439, Herv. i. Orig.)

Eine Person kann nur subjektive Erfahrungen von und über sich selbst haben – nicht aber von und über andere(n). In der Regel schließen wir von unseren eigenen inneren Zuständen auf die anderer Personen; das klappt mal besser, mal weniger gut beziehungsweise sind manche Menschen einfühlsamer und können leichter nachvollziehen, wie es in anderen Personen wohl aussieht, als andere. Benjamin Kuipers (2020: 431) meint, dass induktive Verallgemeinerungen, die anderen bestimmte innere Zustände unterstellen, zwar nützlich zwischen Menschen seien, allerdings nicht sinnvoll für Mensch-Maschine-Interaktionen: „Generalizations that are useful with other humans are unreliable with robots and other AIs, possibly leading to excessive trust, unexpected catastrophes, and other ethical problems". All diese Verständigungsprobleme und daraus resultierenden Folgen können allerdings auch zwischen Menschen passieren – „unexpected catastrophes, and other ethical problems" inklusive. Das Prinzip lässt sich auf technische Systeme übertragen: Ein Analogieschluss, der keine sicheren Aussagen über das Vorliegen oder die Qualität interner Zustände zulässt, funktioniert analog zu Menschen auch mit technischen Systemen, wie Matthias Rath (2019: 228 f.) im Rahmen seiner Untersuchung, ob Maschinen moralische Akteure sein können, bemerkt:

> Setzen wir also ein analogisches Verständnis der Maschine im genannten weiten Sinne voraus, dann ist die alltagspragmatische Beurteilung der *Maschine* als Subjekt in der grundlegenden Analogie zu suchen, die wir auch schon *Menschen* gegenüber anwenden: Wir unterstellen Sinn, weil unser eigenes Handeln und das diese Handlungen bestimmende Entscheiden unserer Selbsterfahrung mit Sinnsetzung entspricht.
>
> (Herv. i. Orig.; vgl. hierzu auch S. 66 in der vorliegenden Arbeit.)

Der Rückschluss auf das Vorliegen bestimmter interner Zustände oder Qualitäten kann nur aufgrund dessen erfolgen, was von außen, durch eine andere Person, erfahrbar ist – sei es durch eine Äußerung, ein bestimmtes Verhalten oder den anderweitigen Ausdruck bestimmter innerer Zustände. Äußert eine Person beispielsweise den Wunsch, etwas zu trinken, kann man davon ausgehen, dass sie etwas trinken *will*.

Daniel Dennett (1971), der bereits im Bezug auf seine Intentionalen Systeme oben zitiert wurde, generalisiert diese Methode des Schließens auf interne Zustände auf der Basis von Verhaltensbeobachtung auf eine Weise, die auch die Anwendung auf technische Systeme zulässt. Neben intentionalen Einstellungen (*intentional stance*), die durch Rückschlüsse von Verhaltensweisen erklärbar sind, gibt es für Dennett dabei noch *physikalische Einstellungen* (*physical stance*) und *funktionale Einstellungen* (*design stance*).

> In *intentionaler Einstellung* [...] erklärt man das Verhalten eines Systems, indem man annimmt, daß das System über bestimmte Informationen verfügt (bestimmte Überzeugungen hat), daß es bestimmte Ziele verfolgt und daß es sich angesichts dieser Informationen und Ziele rational verhält, d.h. [sic!] daß es tut, was unter der Voraussetzung, daß seine Informationen zutreffen, tatsächlich zur Realisierung seiner Ziele führt.
>
> (Beckermann 2001: 307. Herv. i. Orig.)

Analog zu Menschen lässt sich das Rückschlussverfahren auf intentionale Zustände auch auf nicht-menschliche Systeme übertragen, die gemäß Dennett in diesem Sinn auch Rationalität und die Möglichkeit, Ziele und Wünsche zu hegen, haben können – und das wäre erkennbar anhand des jeweiligen Verhaltens oder Outputs. Er verdeutlicht dies am Beispiel eines Schachcomputers (bezüglich möglicher intentionaler Zustände bei technischen Systemen die größte Errungenschaft der 1970er- und '80er-Jahre):

> One predicts behavior in such a case by ascribing to the system *the possession of certain information* and by supposing it to be *directed by certain goals*, and then by working out the most reasonable or appropriate action on the basis of these ascriptions and suppositions. It is a small step to calling the information possessed the computer's *beliefs*, its goals and subgoals its *desires*. What I mean by saying this is a small step is that the notion of possession of information or misinformation is just as Intentional [sic!] a notion as that of belief.
>
> (Dennett 1971: 90)

Ein System verfügt also nach Dennett über Intentionen, wie Wünsche und Überzeugungen, wenn man anhand des Verhaltens „verläßlich und umfassend“ (Beckermann 2001: 309) auf intentionale Einstellungen rückschließen kann. Demgemäß

könnte man auch bei technischen Systemen anhand des beobachtbaren Verhaltens auf bestimmte innere Zustände schließen und objektiv wäre kein Unterschied zwischen Intentionalität von Menschen und Maschinen auszumachen.

Argumente gegen die Dennett'sche Lösung des Intersubjektivitätsproblems

Die Dennett'sche Lösung des Intersubjektivitätsproblems kann als Möglichkeit gesehen werden, doch von einer Gleichwertigkeit zwischen Intentionen bei Menschen und Maschinen auszugehen, also die Identitätsthese anzunehmen. Es gibt dabei jedoch ein paar Haken. Zum einen kann eine Person Wünsche oder Überzeugungen äußern, obwohl sie diese nicht vertritt. Sie kann beispielsweise den Wunsch äußern, etwas zu trinken, und möchte dabei einen Scherz mit ihrem Gegenüber machen. Demnach sind das Ziel, etwas zu trinken, und der geäußerte Wunsch, etwas trinken zu wollen, in diesem Fall nicht ernstgemeint. Das Ziel, das dem Gegenüber nicht klarwerden muss (je nachdem, wie gut die erste Person ihren Scherz zu verbergen versteht), ist vielmehr, einen Scherz mit der anderen Person zu treiben. In einem anderen Fall kann eine Person einer Selbsttäuschung unterliegen oder beispielsweise im Fall von Sucht sich etwas wünschen oder Ziele verfolgen, bei denen fraglich ist, ob sie diese wirklich sich wünschen und verfolgen möchte – im Hinblick auf das höhere Ziel des Überlebens beispielsweise. Diese Beispiele, die in der philosophischen Diskussion um Selbsttäuschung zentral sind (vgl. z. B. Beier 2010), könnten natürlich noch vielfach ergänzt und weiter ausgeführt werden.

Mit Bezug auf Matthias Raths Zitat oben kann zudem angeführt werden: Es stimmt, dass Menschen von sich selbst und eigenen Erfahrungen auf andere schließen – aber das geschieht ja keinesfalls immer fehlerfrei. Kommunikation birgt in vielen Fällen Missverständnisse, auch wenn die Informationen aus subjektiver Sicht noch so eindeutig geäußert wurden. Dies liegt auch an der Vielschichtigkeit des psychologischen Sender-Empfänger-Systems, bei dem die Inhalte bekanntermaßen nur eine geringe Rolle spielen neben Gestik, Mimik, Geruch und anderen Faktoren.

Äußerungen und Verhaltensweisen sind also zumindest bei Menschen fehleranfällig und können zu Missverständnissen führen. Bei technischen Systemen kann man zwar in einem ersten Schritt davon ausgehen, dass diese ja eindeutig zu einem bestimmten Ziel von Menschen konstruiert wurden, jedoch sind auch hier Missverständnisse in der *Mensch-Maschine-Kommunikation* nicht auszuschließen (s. hierzu auch S. 111). Wenn Technik nicht so funktioniert, wie Menschen es intendiert haben, ist das jedoch, wie oben gezeigt, in der Regel auf ein Nichtwissen von Menschen zurückzuführen und kein Zeichen von maschineller Irrationalität im Dennett'schen Verständnis – außer, einem Programm liegt ein entsprechendes Fehlverhalten ebenfalls absichtlicherweise (und dann wiederum von Menschen intendiert) zugrunde.

Des Weiteren ist fraglich, inwiefern sich der Dennett'sche Behaviorismus auf die anderen zentralen Begriffe übertragen lässt. Ist ein Mensch, geschweige denn eine Maschine, frei nur, weil sie sich so verhält, als wäre sie frei? Die Beantwortung der Frage müsste natürlich mit einer Definition von Freiheit in einem jeweils spezifischen Kontext einhergehen und kann nur daran gemessen werden. Daraus ergibt sich die Frage, in welchem Maß Menschen technischen Systemen Werte und Zustände in Form von sozialen Projektionen zuschreiben, die objektiv betrachtet gar nicht vorhanden sind. Studien mit älteren Menschen im Umgang mit automatisierten Kuscheltieren zeigen zum Beispiel, dass die Menschen soziale Haltungen gegenüber den Robotertieren entwickeln können (vgl. z. B. McGlynn et al. 2017), die mit Anthropomorphisierungen verbunden sein können.

> We humans are prone to anthropomorphize nonhuman, and even inanimate, elements of our environment where we can attribute agency. This can easily lead to assuming that robots and other AIs are more human-like and more capable than they actually are.
>
> (Kuipers 2020: 430 f.; vgl. darin auch Eply, Waytz und Cacioppo 2007)

Die Tatsache, dass „technische Gegenstände als soziale Phänomene und technische Entwicklungen als sozialer[. . . r] Prozeß“ verstanden werden können (vgl. Burkart 1987: 44 in Werle 2021: 128) ist jedoch nicht neu und sagt mehr über Menschen aus als über Technik. Aussagen über das Vorliegen und die Qualität von inneren Prozessen, wie Intentionalität, und Werte, wie Freiheit, können auf dieser Basis nicht getroffen werden. Dennett betont, dass er keine völlige Gleichwertigkeit von menschlichen mentalen Zuständen und intentionalen Einstellungen in technischen Systemen postuliert.

> All that has been claimed is that on occasion a purely physical system can be so complex, and yet so organized, that we find it convenient, explanatory, pragmatically necessary for prediction, to treat it as if it had beliefs and desires and was rational.
>
> (Dennett 1971: 91 f.)

Dieser pragmatische Ansatz kann neben der analogen Rede von Freiheit und Freiheit$_T$, wie sie oben vorgeschlagen wurde, eine weitere Herangehensweise sein, um mit der Frage umzugehen, ob sich die oben analysierten Begriffe auf ethische Maschinen übertragen lassen – oder nicht. Demnach muss nicht abschließend geklärt werden, ob Programme wirklich über intentionale Zustände *verfügen* oder nicht; zugrundegelegt wird bei der analogen Schreibweise aber ein prinzipieller Unterschied zwischen Begriffen bei technischen Systemen und Menschen.

Auf der Basis der Dennett'schen Theorie kann man konstatieren, dass es technische Systeme gibt, die so komplex sind, dass Menschen ihnen zumindest menschenähnliche Prozesse zuschreiben. Dies sagt mehr über die Menschen aus als über die

technischen Systeme selbst. Über das eigentliche Vorliegen jener inneren Prozesse wird damit keine sichere Aussage getroffen. Somit beinhaltet die These Dennetts eine größere Erkenntnis auf anthropologischer denn als auf technikphilosophischer Ebene und ist zwar möglicherweise nicht falsch, aber eben auch nicht relevant für ontologische Fragen bezüglich der Fähigkeiten und Eigenschaften technischer Systeme.

Das Intersubjektivitätsproblem kann wiederum als Bekräftigung für die analoge Redeweise von Freiheit, Autonomie, Rationalität und Intentionalität bei technischen Systemen herangezogen werden, um zu zeigen: Wenn schon die Zuschreibung interner Zustände bei Menschen aus epistemologischer Sicht schwierig ist, wieso sollte man dann diesen Nachweis bei Maschinen verlangen?

5.4 Ethische Theorien auf Maschinen angewandt

Neben der Strategie, einzelne Werte, wie Freiheit und Autonomie, oder interne Zustände, wie Bewusstsein und Intentionalität, für ethische Maschinen anzunehmen, gibt es auch Überlegungen dazu, ganze Theorien in technischen Systemen zu verorten. Auf Wallach und Allen (2009) geht die Unterscheidung zurück, den ersten Ansatz als *bottom-up*-Strategie zu bezeichnen und den zweiten als *top-down*. Weiter oben habe ich bereits zu zeigen versucht, dass diese Unterscheidung nur bedingt unterschiedliche Ansätze darstellt. Denn wie die Analyse wesentlicher Begriffe für ethische Reflexionsfähigkeit im ersten Teil dieses Kapitels gezeigt hat, ist selbst, wenn man Freiheit, Autonomie und Bewusstsein für ethische Techniksysteme als Voraussetzung akzeptierte, noch nicht klar, wie ein technisches System in einem konkreten Fall auf die Frage *Was soll ich tun?* reagieren, wie es sich *verhalten* sollte. Folgt man dennoch der Unterscheidung, so kann man auch sagen: *bottom-up* ist nicht ohne *top-down* möglich, denn durch den Versuch, ganze Theorie-Ansätze in Maschinen zu verorten, kann man der Frage, was in einer bestimmten Situation zu tun ist, insofern näherkommen, als sich auf der Basis der Theorien Präzedenzfälle generieren lassen, die einem System eine Regel von der Form *Immer wenn x, dann y* zugrundelegen können. Ob dies auch bei *top-down*-Ansätzen für technische Systeme so zuverlässig funktionieren kann, beziehungsweise welche Schwierigkeiten diese mit sich bringen können, wird in den folgenden Abschnitten deutlich werden.

5.4.1 Utilitaristische Ansätze

Bei utilitaristischen Ansätzen werden vor allen Dingen Effekte von Aktionen oder Handlungen bewertet. Dies bedeutet nicht, dass interne Zustände ausgeklammert werden könnten, allerdings sind sie nur dann relevant, wenn sie mit den Folgen in einem kausalen Zusammenhang stehen. Catrin Misselhorn spricht von einem *moralischen Kalkül*, das der „Moralimplementation“ entgegenkomme,

> denn ein solcher Kalkül scheint grundsätzlich algorithmisierbar zu sein. Der entsprechende Algorithmus im Fall des klassischen Utilitarismus soll diejenige Handlungsalternative berechnen, die die beste Bilanz aus Lust und Leid im Hinblick auf alle Betroffenen gewährleistet. Der dazu benötigte Input besteht in der Anzahl der betroffenen Personen, der Intensität und Dauer der jeweils für sie zu erwartenden Lust bzw. des Leids und der Wahrscheinlichkeit ihres Eintretens für jede Handlungsoption. Ein Computerprogramm verspricht diese Form der Nutzenaggregation deutlich schneller und akkurater durchzuführen als ein Mensch und entspricht dabei womöglich auch dem utilitaristischen Ideal der Unparteilichkeit besser.
> (Misselhorn 2019: 97)

Die Idee, die dem zugrundeliegt, scheint die zu sein, dass beispielsweise anhand großer Datenmengen mit höherer Wahrscheinlichkeit in bestimmten Situationen das Vorliegen von Kriterien für *Lust* und *Leid*[119] definiert und demnach eine *moralische* Vorgehensweise regelrecht berechnet werden könnte. Die damit verbundene Nützlichkeit (*utility*) bezieht sich dabei auf den kollektiven Nutzen und nicht, wie man meinen könnte, auf den individuellen (vgl. Birnbacher 2021: 160). Catrin Misselhorn beschreibt genauer, wie die Umsetzung eines utilitaristischen Ansatzes in einem technischen System aussehen könnte:

> Versucht man, einen utilitaristischen Ansatz zu implementieren, so müssen vier Schritte umgesetzt werden: 1. eine Beschreibung der Ausgangssituation, 2. eine Möglichkeit, um Handlungen hervorzubringen, 3. einen Weg, um vorherzusagen, was geschieht, wenn eine Handlung in der Ausgangssituation durchgeführt wird und [sic!] 4. eine Methode zur Bewertung der daraus entstehenden Folgen.
> (Misselhorn 2019: 97)

Zunächst müsse entschieden werden, „welches die relevanten Aspekte einer Situation“ (ebd.: 97 f.) für eine größtmögliche Zahl an Menschen innerhalb eines bestimmten Moralkreises sind (vgl. Birnbacher 2021: 161). Dazu gehört die Frage, „wer in den Kalkül eingeschlossen wird“ (Misselhorn 2019: 97 f. s. hierzu auch S. 68). Zusätzlich zu der Frage, ob ein moralisches Problem für alle Menschen, eine Teilgruppe, Tiere oder auch Ökosysteme gelöst werden soll, spielt beim Utilitarismus auch die Gewichtung eine Rolle – dies wird auch als *Prioritarismus* beziehungsweise *Präferenzutilitarismus* bezeichnet:[120]

> Sind alle Betroffenen gleich zu bewerten oder sind Lust bzw. Leid oder Präferenzen von Menschen und Tieren unterschiedlich zu gewichten? Selbst wenn man sich nur

[119] Es kann diskutiert werden, ob diese Ausdrücke treffend und sinnvoll sind. Andere Ausdrücke werden ebenfalls verwendet (s. u.), dies soll jedoch für die knappe Darstellung des Ansatzes hier keine Rolle spielen.

[120] Birnbacher (2021: 161) sieht darin zwei Spielarten des Utilitarismus, neben dem grundlegenden Zwei-Ebenen-Modell der größtmöglichen Lust für die größtmögliche Zahl.

> auf Menschen konzentriert, müsste der Input im Extremfall in einer Darstellung der aktuellen Lust bzw. Schmerzbilanz aller Menschen bzw. einer Aufzählung aller ihrer Präferenzen bestehen, um dann entscheiden zu können, auf wen sich eine Handlung auswirkt.
>
> (Misselhorn 2019: 97 f.)

Je mehr für eine Handlung relevante Faktoren es gibt, „desto detaillierter müssen auch die Handlungsoptionen bestimmt werden", so Misselhorn (ebd.: 98). Ebenso müsse man direkte und indirekte Folgen berücksichtigen, wobei sich die Frage stelle, wie weit in die Zukunft hinein sich Folgen abschätzen lassen (ebd.).

Neben den vielen Fragen, die sich innerhalb einer utilitaristischen Theorie stellen – Welche Kriterien werden für *Lust* und *Leid* angesetzt? Bedürfnisse welcher Teilgruppen oder Spezies werden einkalkuliert? Wie werden die einzelnen Präferenzen gewichtet? usw. –, stellt sich bei der Anwendung auf technische Systeme außerdem die Frage, welche Werte einem System von Menschen vorgegeben werden sollten – und welche das System selbst generieren sollte. Die Idee der ethischen Maschine ist ja, dass sie selbstständig reflektieren, Begründungen generieren, bestimmte Normen generieren und Argumente gegeneinander abwägen können sollte. Selbst wenn man einem Computerprogramm heute bereits mit der Klassifizierung großer Datenmengen in Kategorien (beispielsweise Präferenzen) beauftragen könnte und je nach Vorliegen einer Präferenz in einer bestimmten Situation eine spezifische Vorgehensweise an die Hand geben könnte, so ist fraglich, ob Menschen es riskieren wollten, die Abwägung von Präferenzen komplett Maschinen zu überlassen. Jessica Heesen et al. (2020a: 3) stellen die Anforderungen an aktuelle Entwicklungen im Bereich *künstlicher Intelligenz* in ihrem Whitepaper aus der *Plattform Lernende Systeme* so dar:

> Die Entwicklung und der Einsatz von KI-Systemen sollten immer das Ziel verfolgen, zur Förderung des Gemeinwohls beizutragen. Daraus ergibt sich, dass KI-Systeme so eingesetzt werden sollen, dass Schaden für Einzelpersonen, die Gemeinschaft und die Umwelt vermieden wird, die Rechtskonformität von KI-Systemen in der Praxis von Entwickelnden, Anbietenden und Nutzenden gewährleistet ist und das eingesetzte KI-System die notwendige technische Robustheit erfüllt, dass es zu keinem Zeitpunkt ein unannehmbares Sicherheitsrisiko darstellt.

Die Anwendung utilitaristischer Systeme ist nicht denkbar, ohne dass Menschen, Philosoph∗innen, Ethiker∗innen, Unternehmer∗innen und / oder Politiker∗innen, bestimmte Kriterien festlegen, anhand derer bestimmte Entscheidungen getroffen werden sollen. So attraktiv utilitaristische Ansätze auf den ersten Blick aufgrund ihrer Kalkulierbarkeit und dem Fokus auf den Effekt von Aktionen[121] also

[121] Birnbacher (2021: 160) spricht auch von einer *Ethik der Effizienz*.

auch scheinen, so schwierig ist die konkrete Ausgestaltung utilitaristischer Prinzipien. Aus metaethischer Sicht wäre es angemessen, weitere Ansätze innerhalb konsequentialistischer Theorien zu differenzieren und diese zu spezifizieren. Neben monistischen Ansätzen bezüglich dessen, was moralisch richtig ist, wie dem Utilitarismus, bei dem je nach Formulierung *Lust*, *Glück* oder *Bedürfnisbefriedigung* u. a. als alleinige Ziele im Zentrum moralischen Handelns stehen, gibt es auch pluralistische Ansätze, die mehr Faktoren als *moralisch gut* zulassen (vgl. u. a. Larry und Moore 2021). Auch Pflichten oder Rechte können als utilitaristische Ziele (in pluralistischen Ansätzen) geltend gemacht werden (vgl. Nozick 1974 in Larry und Moore 2021). Die vorangegangenen Überlegungen stellen daher nur einen kleinen Auszug aus möglichen denkbaren konsequentialistischen Debatten dar.

5.4.2 Deontologische Ansätze

Deontologische Ansätze werden immer wieder als Gegensatz zu utilitaristischen oder konsequentialistischen Ansätzen dargestellt, da bei ihnen bei der ethischen Bewertung nicht die Folgen von Handlungen oder Aktionen im Fokus stehen, sondern die Pflichten (gr. *déon*) beziehungsweise das *das Gesollte* im Hinblick auf das moralisch *Richtige*:

> Consequentialists hold that choices—acts and/or intentions—are to be morally assessed solely by the states of affairs they bring about. Consequentialists thus must specify initially the states of affairs that are intrinsically valuable—often called, collectively, "the Good." They then are in a position to assert that whatever choices increase the Good, that is, bring about more of it, are the choices that it is morally right to make and to execute. [...] In contrast to consequentialist theories, deontological theories judge the morality of choices by criteria different from the states of affairs those choices bring about. The most familiar forms of deontology, and also the forms presenting the greatest contrast to consequentialism, hold that some choices cannot be justified by their effects—that no matter how morally good their consequences, some choices are morally forbidden.
>
> (ebd.)

Mit deontologischen Ansätzen wird, insbesondere in maschinenethischen Diskussionen, in erster Linie der kategorische Imperativ Immanuel Kants in Verbindung gebracht; aber auch die Asimov'schen Robotergesetze (s. S. 57) werden von manchen Autor∗innen dieser Kategorie zugeordnet. Wallach und Allen (2009: 79) erwähnen ferner die Goldene Regel, die zehn Gebote, Yama und Niyama aus dem Hinduismus sowie eine Reihe von Tugenden.[122] Regeln oder Prinzipien werden gemäß Werner

[122] Zur Definitionsvielfalt vgl. auch Abschnitt 33.1 in Werner und Düwell 2021: 171.

und Düwell (2021: 173) jedoch allenfalls umgangssprachlich als *deontologisch* bezeichnet, da die Voraussetzung deontologischer Ethiken ein größerer theoretischer Rahmen sei. Sie verweisen ferner auch auf den Unterschied zwischen *pragmatisch deontologischen* und *kategorisch deontologischen* Begründungen, die jedoch beide innerhalb eines größeren Theorierahmens zum Tragen kommen müssten (vgl. Werner und Düwell 2021: 173). Die meisten Ethiker∗innen sind sich einig, dass die vielzitierten Asimov'schen Regeln zwar verlockend einfach und zugleich erfolgversprechend zu sein scheinen, jedoch aufgrund der Gefahr der Konflikte zwischen einzelnen Regeln und des hohen Abstraktionsgrads wenig brauchbar für konkrete technische Umsetzungen sind. Dennoch werden sie insbesondere im Kontext von Wissenschaftskommunikation (Vorträge u. a.) maschinenethischer Inhalte häufig zitiert.

Die Tatsache, dass deontologische Ansätze entgegen teleologischer Ethiken den Fokus auf die Einhaltung moralischer Grundsätze legen,[123] könnte ein Vorteil für die Implementierung von ethischen Theorien in technischen Systemen sein.

> A rule-based ethical theory is a good candidate for the practical reasoning of machine ethics because it generates duties or rules for action, and rules are (for the most part) computationally tractable.
> (Powers 2011: 46 f.)

Der *kategorische Imperativ* stellt gemäß Immanuel Kant eine verlässliche Prüfung für moralische Handlungen dar: „Handle nur nach derjenigen Maxime, durch die du zugleich wollen kannst, dass sie ein allgemeines Gesetz werde" (GMS: AA IV: 421). Legt man einem Programm allein die Ausrichtung am moralisch Richtigen gemäß dieser Maxime zugrunde, dürfte es ohne menschliches Zutun überfordert sein. Interpretiert man deontologische Maximen wiederum strikter, anhand bestimmter (absolutistischer; vgl. Werner und Düwell 2021: 172) Regeln – was, wie oben gezeigt, nicht selbstverständlich und ohne Theorierahmen nicht wissenschaftlich wäre –, dann ist fraglich, ob damit der kategorische Imperativ noch eingehalten werden könnte. Wenn Menschen ein Set universeller, moralischer Regeln für ein Programm festlegen würden, so Thomas Powers (2011: 47), könnte das Programm daraus eine ethische Theorie entwickeln „by applying the universalization step to individual maxims and then mapping them onto traditional *deontic categories* – namely, forbidden, permissible, obligatory actions – according to the results" (Herv. i. Orig.). Ein anderer Weg wäre, vorab bestimmte Maximen den entsprechenden Kategorien zuzuordnen – was aber in jedem Fall eine begrenzte Zahl von Regeln voraussetzen

[123] Die Unterscheidung wird auch durch das Anstreben des moralisch Guten (Konsequentialismus) im Unterschied zum moralisch Richtigen (deontologische Ethiken) definiert (vgl. Werner und Düwell 2021: 172; darin Frankena 1963 und Rawls 1971).

würde. Powers räumt ein, dass seine Überlegungen die Theorie Kants nur streifen und er keine strenge Interpretation behaupte. Legt man tatsächlich alle Faktoren des sehr voraussetzungsreichen Ansatzes des kategorischen Imperativs zugrunde, sehen einige Autor∗innen keine Möglichkeit der Integration dieses Ansatzes in Maschinen (vgl. z. B. Brieger 2019).

Wallach und Allen (2009: 95) setzen das Vorliegen von Zielen, eine klare Vorgehensweise sowie die Wahrnehmung der Umstände für solche Programme voraus, die technische Varianten des kantischen kategorischen Imperativs umsetzen können sollen. Die Kommunikation zwischen verschiedenen Systemen sowie Erfahrung und Kontextwissen könnten dabei durch große Datenmengen und technische Wahrnehmungssysteme, die sich mittlerweile auch die Digitalisierung zunutze machen können (Stichwort *Internet of Things*, vgl. Rieder 2021), realisiert werden. Angesichts von *big data* scheint dies kein allzu großes Problem mehr zu sein, da zum Beispiel Algorithmen, die auf Mustererkennung trainiert werden (somit auch Sprach-, Gesichtserkennung usw.), zunehmend besser funktionieren. Allerdings bleibt die Frage der Einordnung (insbesondere bei utilitaristischen Ansätzen) sowie der Definition der ethischen Normen (bei deontologischen Ansätzen) bestehen.

Ebenso wie in Bezug auf utilitaristische Systeme die Frage gestellt wurde, ob es technischen Systemen überlassen werden kann, über die Präferenzen bestimmter Gruppen zu entscheiden, stellt sich bei deontologischen Ansätzen die Frage, ob technische Systeme die Bewertung von Handlungsmaximen und gegebenenfalls eine Hierarchisierung vornehmen können. Denn selbst wenn ihnen bestimmte moralische Regeln (wie beispielsweise in den Geschichten von Isaac Asimov) von Menschen mitgegeben werden, so ist die Voraussetzung für ethische Techniksysteme ja insbesondere die selbstständige Entscheidungsfindung über Handlungsoptionen – unabhängig von Normen vorgebenden Menschen. Das Beispiel des Chatbots Tay (s. S. 80 f.) hat die Manipulierbarkeit solcher lernender Systeme verdeutlicht. Und auch andere Beispiele, unter anderem von Aktionen in Social-Media-Kanälen, zeigen: Große Datenmengen bilden nicht unbedingt eine große Bandbreite unterschiedlicher Haltungen ab, sondern bestimmte Narrative können – völlig unabhängig von ihrem moralischen Gehalt – sogar verstärkt und somit überrepräsentiert werden; Gleiches gilt, wie gezeigt, für soziale Ungerechtigkeiten, die sich in Datenmengen widerspiegeln können (vgl. zu der Thematik u. a. Thomaß et al. 2022; Richterich 2018; Richter 2016; Schmidt 2016).

Wallach und Allen (2009: 96) ergänzen, dass bei aller Priorität von ethischen Normen auch bei deontologischen Ansätzen in technischen Systemen die Folgen beachtet und hier an verschiedenen Stellen Bewertungen (*assessment*) vorgenommen werden müssten, bei denen fraglich ist, anhand welcher Kriterien technische Systeme dies sollten leisten können und wie flexibel die Kriterien angewandt werden könnten:

> An AMA that followed the Golden Rule would need to be able to (1) notice the effect of other's actions on itself, assess the effect (also in hypothetical situations), and choose its preferences; (2) assess the consequences of its own actions on the affective states of others, and decide whether they match its own preferences; and (3) take into account differences in individual psychology while working on (1) and (2), as people affected by the action might respond differently to the same treatment. The latter point would presume that the AMA has a capacity to discern and anticipate changes in the affective reactions of people to its decisions. Predicting the actual consequences of actions is difficult to impossible.
>
> (Wallach und Allen 2009: 96)

Spezifische Regeln sind zwar aus technischer Sicht recht einfach zu implementieren, allerdings stellen sich für komplexe Situationen Folgefragen, die entweder doch das konkrete Eingreifen von Menschen erforderlich machen oder technische Implementierung wesentlich komplexer gestalten würden (vgl. ebd.: 94). Werner und Düwell (2021: 171) beziehen sich auf den Definitionsvorschlag von William Frankena (1963), wonach das moralisch Richtige oder Gute in deontologischen Ansätzen kein *vormoralisches Gut* ist (wie etwa das Glück aller); vielmehr gebe es weitere Kriterien, die moralische Handlungen zu richtigen machen (wie etwa bestimmte moralische Pflichten, Versprechen usw.).

5.4.3 Tugenden für technische Systeme

Gemäß Platon, der in seinem Dialog *Menon* Sokrates mit ebendiesem sprechen lässt, kann man nicht anders als richtig zu handeln, wenn man tugendhaft ist. Tugend (*areté*) kann dabei sowohl als „Fähigkeit, Eignung oder Eigenschaft" (Frede 2008: 33) verstanden werden, moralisch gut, das heißt im Hinblick auf ein gelingendes Leben (*eudaimonía*) zu handeln, aber auch genereller als körperliche Tüchtigkeit im Sinn einer zweckdienlichen Tauglichkeit, die nicht nur Menschen, sondern auch Gegenständen zugeschrieben werden kann (ebd.: 34).

Gemäß der aristotelischen Tugendethik bedarf tugendhaftes Handeln einer charakterlichen Bildung, die aus einer Mischung von Gewöhnung und Erfahrung gewonnen wird und sich nur über einen längeren Zeitraum entwickeln kann.

> Die Tugend ist also eine Disposition (*hexis*), die sich in Vorsätzen äußert (*prohairetiké*), wobei sie in einer Mitte liegt, und zwar der Mitte in Bezug auf uns, die bestimmt wird durch die Überlegung (*logos*), das heißt so, wie der Kluge (*phronimos*) sie bestimmen würde.
>
> (NE: II5, 1106b35. Übers. v. Ursula Wolf)

Tugendhaftes handeln ist auch an der *Verlässlichkeit* erkennbar, mit der moralische Handlungen ausgeübt werden (vgl. hierzu z. B. Annas 2011: 9). Die Klugheit des Denkens (*phrónēsis*), die einen Tugendhaften Menschen die *richtige Mitte* in einer

bestimmten Situation erkennen lässt, sorgt wiederum für die Flexibilität, auch in vielen verschiedenen Situationen die richtige Wahl im Hinblick auf moralische Prämissen zu treffen – Faktoren, die große Vorteile für die Implementierung in Computerprogrammen mit sich bringen würden.

> Der Vorteil des aristotelischen Ansatzes mit Blick auf Roboter, die auf künstlichen neuronalen Netzwerken basieren, liegt darin, dass die Theorie nicht explizit in die Programmierung des Systems einfließen muss, sondern vielmehr die Rahmenbedingungen für das Training des künstlichen Akteurs liefert. Auch muss ein Programmierer nicht konkret angeben können, worin die moralische Signifikanz einer Situation besteht. Die Maschine wird die Bedingungen moralischen Handelns anhand von Beispielen oder durch Versuch und Irrtum lernen und selbstständig auf neue Herausforderungen anwenden können.
> (Brand 2018: 108)

Das Ziel tugendhaften Handelns ist *eudaimonía*, was auch mit Glück, Glückseligkeit oder dem guten Leben übersetzt wird. Tugendhaftigkeit ist dabei nur ein Baustein auf dem Weg zu einem glücklichen Leben, das in der aristotelischen Philosophie recht konkret ausdifferenziert wird, „wohlgeratene Kinder," und „Schönheit" gehören zum Beispiel dazu,

> denn wer sehr hässlich aussieht oder von niedriger Herkunft oder einsam und kinderlos ist, den kann man wohl nicht glücklich nennen, und noch weniger vielleicht den, der gänzlich schlechte Kinder oder Freunde hat oder gute, die gestorben sind.
> (NE: I 10, 1099b. Übers. v. Ursula Wolf)

Letztgenannte Bedingungen ebenso wie Freude an „werthaften Handlungen" (NE: I 10, 1099a15. Übers. v. Ursula Wolf) zählen aber nur zu den „günstigen Umständen" (NE: I 10, 1099b5), die *eudaimonía* bedingen können. Gemäß Aristoteles haben Dinge ebenso wie Lebewesen eine bestimmte Funktion (*érgon*) und wer gemäß dieser Funktion lebt und sich gut entwickelt, hat Voraussetzungen, tugendhaft zu werden.

> Wo Dinge einer Art spezifische Funktion (*ergon*) haben, reden wir davon, dass sie diese Funktion besser oder schlechter erfüllen, und wo ein Ding die Funktion gut erfüllt, nennen wir es ein gutes Ding dieser Art, ein Ding, das die *areté* dieser Art besitzt. (Messer haben die Funktion zu schneiden, und ein Messer, das gut schneidet, ist ein gutes Messer, es ist gut als Messer.) Aristoteles geht davon aus, dass auch jede Spezies von Lebewesen eine spezifische Funktion (*ergon*) hat. Diese ist nicht vorgegeben, sondern besteht – im Sinn einer internen Teleologie – im Vollzug der arttypischen Lebensweise. Was den Menschen von anderen Lebewesen unterscheidet ist seine Vernunftfähigkeit; dann besteht die typisch menschliche Lebensweise im Vollzug eines Lebens in der Betätigung dieser Fähigkeit und die *areté*

> des Menschen darin, dies auf gute Weise zu tun. Gut als Mensch ist also, wer sein Leben in der Weise der vernunftgemäßen Betätigung vollzieht. In diesem Sinn gut als Mensch zu sein, die menschliche *areté* auszuüben, soll nun aber gemäß der Argumentation in I 6 gerade den Inhalt der *eudaimonia* ausmachen.
> (Wolf 2013: 13)

Tugenden lassen sich also in Charaktertugenden einerseits (*areté éthikê*) und Tugenden des Denkens (*areté dianoéthiké*) andererseits einteilen (NE: II 1, 1103a14). Die entscheidende Frage im Hinblick auf die Möglichkeit tugendhafter Maschinen vor dem Hintergrund einer aristotelischen Tugendethik ist die, welches *érgon* einem technischen System jeweils zugrundeliegt, zu welchem Zweck es entwickelt und gebaut wurde sowie welche Funktion es erfüllen soll. Unabhängig davon, dass die primäre Funktion technischer Systeme freilich auch missbraucht werden kann, liegt eine solche immer zugrunde und in Anlehnung an die aristotelische Philosophie könnte man die These aufstellen, dass technische Systeme genau dann tugendhaft sind, wenn sie auch gemäß dem Zweck, zu dem sie erfunden, entwickelt und gebaut wurden, eingesetzt werden. Dies ist jedoch zu simpel gedacht, wenn man zum einen bedenkt, dass technische Entwicklungen von der Idee bis zur Ausführung oft auf langjährigen Prozessen beruhen, in deren verschiedenen Stadien auch unterschiedliche Funktionen als Ziele definiert werden können. Außerdem liegt die Tugendhaftigkeit bei Aristoteles ja gerade in der *eigenen* und selbstständigen Entwicklung, Gewöhnung und Ausübung der tugendhaften Tätigkeit, weshalb sich die Frage stellt, ob moralische Maschinen überhaupt als im weitesten Sinn tugendhaft gelten können, wenn sie doch stets als Werkzeuge von Menschen (vgl. Heil 2021: 426, 81.3) fungieren und damit nicht über den Objektstatus hinauskommen. Drittens wurde oben im Zusammenhang des Begriffs der Intentionalität bereits das Argument angeführt, dass es auch technische Systeme geben kann, die mit unmoralischen Zwecke entwickelt wurden. Lukas Brand (2018: 131) sieht eine Möglichkeit, wie menschliche Erfahrung mittels großer Datenmengen mit Methoden des maschinellen Lernens in tugendethischen Techniksystemen imitiert werden könnte:

> Die eigentliche Hürde, die eine künstliche Intelligenz in Bezug auf die nötige Veranlagung nehmen muss, ist, sich das Erfahrungswissen anzueignen, das für die Schulung der Charakter- und Verstandestugenden unerlässlich ist. [...] Die Zeit, die ein künstliches neuronales Netzwerk braucht, um sich die nötige Erfahrung anzueignen, ist zwar im Vergleich zum Menschen eventuell enorm reduziert, das Lernmaterial und die Lernbedingungen müssen aber in einer solchen Form gegeben sein, dass sie sich zum maschinellen Lernen eignen.

Brand führt anhand eines Beispiels aus, wie das maschinelle Lernen anhand einer Mischung aus überwachtem, unüberwachtem und verstärkendem Lernen auf der Basis eines großen Datensatzes aussehen könnte.

> Die Beispiele müssten hinsichtlich ihrer moralischen Qualität ausgezeichnet (*labeled*) sein, das heißt, es muss klar sein, ob sie eher ein gutes oder schlechtes Beispiel für eine Handlung geben, die das gute Leben verwirklichen. In Anlehnung an die Lebenserfahrung, die Aristoteles in Bezug auf den klugen Menschen vorschwebt, kommen hier vor allem eine Vielzahl von Einzelfallbeispielen infrage, von denen das künstliche neuronale Netzwerk lernen und abstrahieren kann. Das Lernmaterial könnte auf ähnliche Weise zustande kommen wie die Datensätze von Cambridge Analytica. [...] Schließlich könnte das Lernmaterial sowohl Szenen aus Literatur und Film, als [sic!] auch Alltagssituationen enthalten, in denen ein Verhalten oder eine Situation möglichst eindeutig als gut oder schlecht qualifiziert ist.
>
> (Brand 2018: 131 f.)

Eine große Frage ist dabei natürlich, wie die einzelnen Bewertungen (*labels*) vorgenommen werden und wer hier über moralisch *gut* und *nicht gut* entscheidet beziehungsweise nach welchen Kriterien. Entweder, so kann man meinen, werden Werte technischen Systemen bereits mitgegeben, dann haben wir es mit moralischen technischen Akteuren zu tun – oder Programme müssten selbst, beispielsweise mittels unüberwachten Lernens, die gelernten Werte auch auf neuen Situationen übertragen. Es ist aber auch denkbar, dass die beim maschinellen Lernen sonst bevorzugte Mustererkennung, insbesondere beim unüberwachten Lernen, also bei der statistischen Anwendung bisheriger Muster auf neue Situationen, gar nicht erwünscht ist. Zwei Situationen können sich in allen möglichen Faktoren sehr ähnlich sein und trotzdem mag es in dem einen Fall moralisch sinnvoll sein auf die eine Weise und im anderen Fall auf eine andere Weise zu entscheiden. Das müssen auch Menschen lernen und es bleibt abschließend fraglich, auf welcher Grundlage – wenn nicht durch menschliches Einschreiten – Programme lernen sollten, in welcher Situation sie gewohnte Heuristiken anwenden dürften und in welchen sie davon, aus ethischen Gründen welcher Art auch immer, abweichen sollten. Diese Lücke lässt den tugendethischen Ansatz als wenig praktikabel für die konkrete Umsetzung in ethischen Maschinen erscheinen, auch wenn Teile davon, wie zum Beispiel die analoge Erfahrungsgewinnung, vielversprechend sein mögen.

In ähnlicher Weise stellt sich das Problem dar, wenn man von konkreten Tugenden ausgeht, beispielsweise Gerechtigkeit, Tapferkeit usw. Wer entscheidet, was in einer bestimmten Situation gerecht ist? Auch hierfür bräuchte es zum einen eine große Datenbasis mit gelabelten Fällen, die als *gerecht* oder *ungerecht* bezeichnet wurden, wobei die Deutung dessen, was gerecht ist und was nicht, auf politischer Ebene im internationalen Raum bekanntermaßen recht unterschiedlich ist. Für die Ausrichtung einer ethischen Maschine am Konzept der Gerechtigkeit fehlt ein solches Konzept, das flexibel genug ist, um es auf Situationen unterschiedlicher Art anwenden zu können (vgl. hierzu auch Le Bui und Noble 2020).

Die Analogie menschlicher Erfahrung und maschinellen Lernens stellt also allenfalls eine Annäherung an die maschinelle Abbildung charakterlicher Tugenden dar – die Abbildung dianoethischer Tugenden bedürfte darüber hinaus explizi-

ter Regeln oder Prinzipien, die gemäß Wallach und Allen (2009: 119) *top-down* vorgegeben werden müssten (vgl. hierzu auch Hillerbrand und Poznic 2021: 168, Abschnitt 32.6).

5.4.4 Kulturelle Aspekte

Kulturen unterscheiden sich in Bezug auf Traditionen, soziale Praktiken – und auch im Hinblick auf moralische Vorstellungen. Demnach gibt es zwischen verschiedenen Kulturen unterschiedliche moralische Prinzipien und Vorstellungen davon, was moralisch richtig und falsch ist. Dieser kulturelle Relativismus kann eine Herausforderung für die Entwicklung ethischer Techniksysteme darstellen (vgl. Rath 2020). Die prinzipielle Idee ist, dass verschiedene Situationen in unterschiedlichen Gesellschaften und / oder Ländern unterschiedliche moralische Prinzipien erfordern. Jayanta Banerjee (2020) hat beispielsweise untersucht, welche Rolle kulturelle Unterschiede in der Ausbildung von Ingenieurswissenschaftler∗innen spielen können. Jansen et al. (2016: 10) stellen in ihrem Bericht die ethischen Prinzipien aus verschiedenen Ländern (primär der Europäischen Union) im Hinblick auf „research and innovation (R&I)“ dar. In diesem Zusammenhang machen sie acht Grundwerte für R&I aus, die sich aus Interviews mit 153 Interessenvertreter∗innen ergeben haben:

> research integrity, social responsibility, avoidance of and openness about potential conflicts of interest, protection of and respect for human research participants, protection of and respect for animals used in research, protection and management of data, protection of researchers and the research environment, dissemination of research results.
>
> (Ebd.: 184)

Welche konkreteren Anforderungen dabei in Systemen künstlicher Intelligenz enthalten sein sollten, wird beispielsweise im Whitepaper der *Plattform Lernende Systeme*, *Zertifizierung von KI-Systemen. Kompass für die Entwicklung und Anwendung vertrauenswürdiger KI-Systeme*, deutlich.

> 1.) eine Zertifizierung von KI-Systemen ermöglicht die Einhaltung wichtiger gesellschaftlicher und ökonomischer Prinzipien; 2.) eine Zertifizierung von KI-Systemen kann bei Bürgerinnen und Bürgern Vertrauen schaffen und eine entlastende Entscheidungshilfe in Bezug auf Nutzungsoptionen geben; 3.) eine Zertifizierung kann zu besseren Produkten im Sinne europäischer Werte führen sowie 4.) eine Zertifizierung von KI-Systemen kann die nationale und internationale Marktdynamik beeinflussen.
>
> (Heesen et al. 2020b)

Doch aus diesen und ähnlichen Grundlagen für Kontroll- und Regulierungsmethoden geht natürlich noch nicht hervor, wie sich eine verlässliche, moralische gute Aktionsweise von technischen Systemen systematisch integrieren lassen könnte.

> Ethical complexity comes from at least a couple of sources. On the one hand, there are the nuanced discussions within ethical theory about the fundamental concepts of the discipline. On the other hand, there are the difficult issues that arise from attempting to make normative judgments about real-world situations. Morality in humans is a complex activity and involves skills that many either fail to learn adequately or perform with limited mastery. Although there are shared values that transcend cultural differences, cultures and individuals differ in the details of their ethical systems and mores.
>
> (Wallach und Allen 2009: 76)

Das Problem an dieser Stelle ist, noch vor dem technologischen, ein prinzipiell ethisches. Denn es ist fraglich, ob man kulturelle Unterschiede auch in Bezug auf mögliches moralisches Verhalten bei technischen Systemen aus unterschiedlichen Kulturkreisen einfach hinnehmen muss – ebenso wie ja auch der Umgang und die Akzeptanz von technischen Entwicklungen interkulturell differieren[124] – oder ob es in Bezug auf manche Technologien einen gerechtfertigten Anspruch einer einheitlichen Lösung gibt.
Bis heute gilt, dass in Einzelfällen entschieden werden muss, welche moralischen Werte in technischen Systemen zum Einsatz kommen und wie diese genutzt werden. Ein Verständnis von Technik, welches Technik als „Kulturleistung handelnder Menschen“ (Janich 2021: 104) versteht, kann dabei hilfreich sein. Wendet man diese Idee auf das Konzept ethischer Techniksysteme an, könnten technische Systeme, welche die ethischen Reflexionsleistungen von Menschen adaptieren und simulieren und somit eigenständig ethische Theorien entwickeln beziehungsweise moralische Thesen begründen können, selbst als kulturschaffende Subjekte verstanden werden. Dies führt konsequenterweise zu einem Verständnis der Möglichkeit der *Selbstreproduktion* technischer Systeme auf der Basis eines kulturellen Schaffungsprozesses, in dem technische Systeme sowohl Subjekte als auch Objekte sein können. Dies hätte nicht nur Folgen für den Kulturbegriff, sondern auch für das Selbstverständnis, weil auf dieser Basis die Unterscheidung zwischen sich selbst in dieser Form reproduzierenden Technik und Technik schaffender Menschen ins Wanken gerät (s. hierzu auch Kapitel 7).

Diese These bringt auch im Licht des radikalen Konstruktivismus interessante Folgen mit sich, die hier kurz inform eines Gedankenexperiments dargelegt werden

[124] Vgl. hierzu auch die *Unified Theory of Acceptance and Use of Technology* von Venkatesh et al. 2003.

sollen.[125] Gemäß dieser Ansicht erzeugen Menschen die Realität durch ihre Kommunikation, die Voraussetzungen dafür, also ihre Wahrnehmung, Entscheidungen und Handlungen. Wie bereits an einigen Stellen in der vorliegenden Arbeit gezeigt, haben Programme, die auf große Datenmengen zugreifen können, die Möglichkeit, einen Großteil dessen, was Menschen kommunizieren, wahrnehmen und via soziale Medien oder andere Kanäle verbreiten, auszuwerten. Auf der Basis dieser Annahme ist fraglich, was Realität gestaltende Menschen dann von Realität gestaltenden technischen Systemen unterscheidet – zumal der Inhalt in diesen Fällen, wie bei Menschen, auch von Menschen kommt. Eine Umdrehung weiter gedacht, und das Beispiel von *Beta Writer* oben einbezogen, kann man natürlich die Frage anschließen, ob die konstruktivistische Einschätzung maschineller Bewertungen auf der Basis menschlichen Inputs anders ausfallen würde als auf der Basis, *eigenen* technischen Inputs.[126] Was wäre also, wenn gemäß dem radikalen Konstruktivismus technische Systeme auf der Basis selbst produzierter Daten Realität erschüfen, hätten wir es dann mit einer genuin technischen Realität zu tun? Inwiefern unterscheide sich diese von menschlich konstruierter Realität und welche Konsequenzen hätte dies?[127]

5.5 Zusammenfassung: Theoretische Bedingungen ethischer Techniksysteme

Ethische Techniksysteme sind nicht nur Maschinen, Computerprogramme oder andere technische Systeme, die einen gewissen Output produzieren, welcher im Hinblick auf die zugrundeliegende Moralität ethisch reflektiert werden kann, sondern es sind Programme, welche die ethische Reflexion, die bislang von Menschen kommt, selbst übernehmen können. Durch die Begriffsanalyse und Darstellung der möglichen Umsetzung ethischer Theorien in technischen Systemen habe ich versucht, einer Antwort auf die Frage, inwiefern solche technischen Systeme auf der Grundlage menschlicher ethischer Reflexionen möglich sein könnten, etwas näher zu kommen. Aufgrund der Breite des untersuchten Begriffsspektrums wurden dabei viele Aspekte nur exemplarisch beleuchtet.

Das Zwischenergebnis in Abschnitt 5.3 hat gezeigt, dass für ethische Maschinen, ebenso wie für ethische Reflexions- und Argumentationsfähigkeiten ein gewisses

[125] Für diesen Hinweis danke ich Jessica Heesen.

[126] Diese Frage ist umso brisanter, da wenige große Firmen weltweit derzeit das Machtmonopol über die Priorisierung bestimmter Inhalte und Daten im Internet innehaben.

[127] Da eine ausführliche Beantwortung der Frage den Rahmen sprengen würde, sei an dieser Stelle nur auf die Anmerkung verwiesen, dass für diese Sicht der Subjektstatus von Technik akzeptiert werden müsste.

Maß an Freiheit$_T$, Autonomie$_T$, Rationalität$_T$ und Intentionalität$_T$ vorausgesetzt werden kann. Die Identitätsthese, also die Annahme, dass es keinen Unterschied macht, ob man bei Menschen davon spricht, dass sie frei handeln und entscheiden können, oder bei Maschinen, wurde abgewogen, aber für die Maschinenethik für nicht sinnvoll befunden. Die Begründung fand zum einen vor dem Hintergrund des pragmatisch-sprachphilosophischen Arguments statt, dass etwas Anderes *gemeint* ist, wenn man beispielsweise von *Freiheit von Menschen* oder *Freiheit von Maschinen* spricht. Trotz einiger Gemeinsamkeiten scheint es auch ontologische Unterschiede zwischen Freiheit und Freiheit$_T$ zu geben, die ebenfalls dargelegt wurden. Eine qualitative Bewertung liegt den unterschiedlichen Ausdrücken dabei nicht zugrunde. In der sowohl umgangssprachlichen als auch wissenschaftlichen Praxis werden Ausdrücke, wie *Freiheit*, *Autonomie*, *Rationalität* und *Intentionalität*, vor allen Dingen aber auch *Entscheidungen* und *Handlungen* auf technische Systeme angewandt. Dies ist aber in erster Linie ein Hinweis auf die menschliche Wahrnehmung und Einschätzung von Technik, also ein anthropologischer Aspekt.

Neben einzelnen Begriffen, Werten und kognitiven Qualitäten wurden ebenfalls einige Ansätze für die Umsetzung ethischer Theorien in technischen Systemen vorgestellt. Im Hinblick auf die Systematisierbarkeit moralisch richtiger Handlungen als verlässliche Technikfolge bringen Konzepte, die versuchen, ethische Theorien in technischen Systemen umzusetzen, Vorteile mit sich. Gleichwohl gibt es sowohl auf konzeptioneller als auch auf praktischer Ebene Schwierigkeiten in der Umsetzung. Alle Theorien setzen die Definition ethisch erstrebenswerter Normen und Vorgehensweisen in konkreten Situationen voraus, die im Fall einer Maschinenethik *von* und *für* Menschen (s. S. 30) vorab definiert werden müssen. Unterschiede zwischen der technischen Umsetzung von beispielsweise deontologischen und utilitaristischen Ansätzen sind technisch realistisch, allerdings nicht ohne vorherige, menschliche Normsetzung, die technischen Systemen den großen Handlungsrahmen vorgibt.

Im folgenden Kapitel werde ich, meine Untersuchung abschließend, noch genauer auf mögliche Herausforderungen bei der Umsetzung genuin ethischer Techniksysteme eingehen und dabei auch mögliche Beispiele ethischer Techniksysteme konkretisieren. Die Prüfung des Konzepts anhand des *Weißbuchs zur Künstlichen Intelligenz* der Europäischen Kommission kann bei der Einschätzung helfen, wie realistisch ethische Techniksysteme – technische Details einmal außen vorgelassen – Stand heute sind.

6 Herausforderungen bei der Umsetzung genuin ethischer Techniksysteme

> It has been suggested that machine learning approaches to moral decisions will be limited because they will at best result in human-level moral decision making [sic!]; they will never exceed the morality of humans.
> (Conitzer et al. 2017: 4)

Bezüglich der Implementierung ethischer Theorien in Maschinen wurden bereits einige theoretische Herausforderungen aufgezeigt, die unter anderem Fragen betreffen wie: In welchem Sinn sind technische Systeme moralische Akteure? Kann ein technisches System handeln wie ein Mensch? Welche ethische Theorie sollte in Programmen umgesetzt werden, sofern dies technisch möglich ist? Die Auseinandersetzung mit einigen zentralen Begriffen hat gezeigt, dass eine analoge Rede von Entscheidungen$_T$ usw. sinnvoll sein kann, um Missverständnisse zu vermeiden. Diese Schreibweise soll jedoch nicht als Bewertung der Qualität verstanden werden. Es ist durchaus möglich, die Freiheit$_T$ eines technischen Systems in manchen Hinsichten oder Kontext als höher zu bewerten als menschliche – zum Beispiel weil im Moment davon auszugehen ist, dass technische Systeme nicht durch (ggf. irreführende) Gefühle beeinflusst werden. Dies *kann* eine Erleichterung bei der Reflexion wichtiger ethischer Entscheidungen darstellen.

Außerdem stellte sich bereits die Frage nach möglichen technischen Umsetzungen einzelner Theorien, wofür zum Beispiel für Methoden des maschinellen Lernens bereits Ansätze dargelegt wurden, wie die Kategorisierung einzelner Daten auf der Basis vorab definierter Werte. Doch hier stellt sich rasch die Frage, inwiefern Programme in der Lage sein können, bislang unbekannte Daten in die vorhandenen Kategorien einzuordnen. Hierzu ist es notwendig, dass Ethiker∗innen und Ingenieur∗innen eng zusammenarbeiten, damit ethische Reflexionen in den technischen Prozessen verankert werden können:

> Ethicists must partly be capable of grasping technical details with their intellectual framework. That means reflecting on the ways data are generated, recorded, curated, processed, disseminated, shared, and used [...], on the ways of designing algorithms and code, respectively [...], or on the ways training data sets are selected [...].
> (Hagendorff 2020: 111)[128]

[128] Vgl. hierzu auch de Bruin und Floridi 2017; Gebru et al. 2020; Kitchin 2017; Kitchin und Dodge 2017 in Hagendorff 2020.

Eine möglichst konkrete Diskussion von Anwendungsfällen sei insbesondere deswegen wichtig, da Ethik „no enforcement mechanisms reaching beyond a voluntary and non-binding cooperation between ethicists and individuals working in research and industry“ habe, so Thilo Hagendorff (2020: 108).[129]

Vincent Conitzer et al. (2017) schlagen vor, dass man im Rahmen von Verfahren maschinellen Lernens Daten in einem Datenset (zum Beispiel mit möglichen Verfahrensweisen in Dilemma-Situationen) vorab als *moralisch* oder *unmoralisch* kategorisieren und mithilfe einer Kombination aus überwachtem verstärkendem Lernen einem Programm so das jeweils richtige Verhalten beibringen könnte. Dies ist jedoch nur für eine begrenzte Anzahl von Daten und somit möglicher Verhaltensweisen sinnvoll denkbar. Die Herausbildung von Mustern bestimmter *moralisch richtiger* Vorgehensweisen ist nur für sehr ähnliche Situationen denkbar, erscheint aber nicht praktikabel für eine Verallgemeinerung, auf deren Basis ein Programm wirklich *eigenständig* in einer völlig unbekannten Situation moralisch agieren können soll. Die Autor∗innen kommen zu dem Schluss:

> It has been suggested that machine learning approaches to moral decisions will be limited because they will at best result in human-level moral decision making [sic!]; they will never exceed the morality of humans.
> (Ebd.: 4)

Bei solchen Aussagen stellt sich dabei die Frage, was genau mit *human-level moral decision-making* gemeint ist. Selbst auf der Basis von Reflexionen und Argumenten kommen Menschen – bezüglich Corona oder der im März 2022 aktuellen Kriegssituation – zu den unterschiedlichsten moralischen Überzeugungen. Sowohl viele Impfgegner∗innen als auch -befürworter∗innen brachten in der öffentlichen Debatte Argumente für Ihre Positionen an. Aus ethischer Sicht ist dabei abzuwägen, welche die überzeugenderen sind in Bezug auf das (gute, s. Abschnitt 4.3) gemeinschaftliche Zusammenleben.

Auch Wallach und Allen (2009: 215) halten die Umsetzung von ethischen Theorien derart, dass sie Programme zu komplett eigenständigem Denken und Handeln führt, für eher unrealistisch:

> We started with the deliberately naive idea that ethical theories might be turned into decision procedures, even algorithms. But we found that top-down ethical theorizing is computationally unworkable for real-time decisions.

Nach der Untersuchung der vorliegenden Arbeit scheinen die Umsetzungsschwierigkeiten dabei eher auf konzeptioneller Ebene zu liegen als auf technischer. Die

[129] S. hierzu auch Abschnitt 4.7.

Analyse der Ansätze zur Umsetzung ethischer Theorien hat gezeigt, dass die Auswahl einer Theorie nicht entscheidend ist für die ethischen Normen und Prinzipien, die davon unabhängig und vorab festgelegt werden müssen. Ein beispielsweise utilitaristischer Ansatz kann also nur bedingt als Methode zur Erlangung einer normativen Vorgehensweise führen, da die Bewertung von *Lust* und *Leid* oder bestimmter Präferenzen beteiligter Subjekte a priori abgewogen und vorgenommen werden muss. In dieser Hinsicht sind ethische Techniksysteme als völlig unabhängige Instanzen von Menschen Stand heute nicht realistisch.

6.1 Ethische Techniksysteme: Ein paar Beispiele

Statistische Modelle, die auf der Basis großer Datenmengen versuchen, Muster zu erkennen und die gemäß meiner Definition in den Bereich moralischer Maschinen fallen, könnten auch eine Grundlage für Ethische Techniksysteme sein. Ethisch-technische Systeme könnten demnach auf die Erkennung bisheriger Verhaltensweisen, die als *moralisch richtig* kategorisiert wurden, mittels verschiedener Formen maschinellen Lernens trainiert werden und auf dieser Basis neue Verhaltensweisen oder Handlungsvorschläge generieren. Die Abwägung wirtschaftlicher und ethischer Faktoren in bestimmten Kontexten, das Anpassen von Programmen im Hinblick auf deren Kompatibilität bezüglich (beispielsweise deutscher) demokratischer Grundwerte, die Entwicklung von Vorgehensweisen bei bestimmten Fragen, die in Ethikräten diskutiert werden – dies alles könnten ethische Techniksysteme und Programme leisten; und zwar über eine Befolgung von Regeln hinaus. Solche Programme sind dabei entweder als fortgeschrittene Assistenzprogramme denkbar, die Menschen bei ihren Entscheidungen unterstützen könnten – oder ethische Algorithmen könnten in Programmen und Maschinen zum Einsatz kommen, um deren unabhängiges Verhalten$_T$ gemäß menschlicher, ethischer Normen in verschiedenen Situationen gewährleisten zu können.

Ethische Assistenzprogramme könnten ähnlich, wie bereits heute eingesetzte Algorithmen, in medizinischen Diagnoseverfahren zum Tragen kommen, also als Unterstützung primär menschlicher Entscheidungsprozesse. Im Gegensatz zu moralischen Akteuren wäre dabei das Ziel des ethischen Abwägungsprozesses nicht menschlich vorgegeben. 2020 wurde die GPT3-basierte Software *philosopher AI* bekannt, ein von Murat Ayfer entwickeltes Programm, das komplexe Fragen beantworten können sollte.[130] Erinnern wir uns an das oben zitierte Beispiel von *Siri*

[130]Im März 2022 sind mindestens drei verschiedene Internetseiten unter dem Namen zu finden, https://philosopherai.com/, https://philosopherai.xyz/ und https://www.philosopherai.net/, deren Urheberschaft zumindest bei den letzten beiden nicht eindeutig zu erkennen ist und deren Fragefunktion bei einem Testversuch am 04.03.2022 nicht funktionierte.

(S. 61) – andere Hersteller haben andere, vergleichbare Systeme. Auch diesen Programmen kann man ähnliche Fragen stellen, wie: Was ist der Sinn des Lebens? Was ist Moral? Was ist Liebe? Warum führen Menschen Krieg? Die Antworten basieren zum einen auf ausgewählten, verfügbaren Daten und zum anderen auf der statistischen Berechnung der Wahrscheinlichkeit, wie angemessen eine Antwort ist. In diesem Sinn unterscheiden sich Programme wie *philospherai*, *Siri*, *Alexa* und Co. nicht wesentlich voneinander. Bei allen handelt es sich um Beispiele, die gemäß der hier vorgenommenen Unterscheidung in den Bereich *moralischer Techniksysteme* fallen würden, da sie keine *eigenen* Regeln entwerfen oder Bewertungen anhand eigenständig konstruierter Bewertungssysteme vornehmen, sondern auf der Basis bereits vorhandener, menschengemachter Daten ihren Output erzeugen.

Ethisch-technische Systeme könnten darüber hinaus auf der Basis eigenständig generierter Wertsysteme in menschliche Entscheidungsfindungen involviert sein, wie beispielsweise bei der Suche nach moralisch angemessenen Lösungen für die Endlagerung von Atommüll. In der leichten Form der Assistenzfunktion hätten dabei stets Menschen das letzte Wort – ebenso wie bei derzeit sich im Einsatz befindlichen Assistenzsystemen. Genuin ethische Systeme könnten darüber hinaus ohne menschliches Zutun selbst verschiedene moralische Entscheidungen$_T$ treffen auf der Basis guter Gründe gemäß einer jeweiligen, zugrundeliegenden Theorie auch danach handeln$_T$.[131]

Zum einen bestünde hier aber dasselbe Problem der Güte von in der Vergangenheit erhobenen Daten wie bei anderen Programmen, die mittels Methoden des maschinellen Lernens trainiert werden (Stichwort: *bias*). Zum anderen gibt es einen entscheidenden Unterschied zwischen beispielsweise Programmen und Texten, die Sprache verarbeiten, und solchen, die moralische Regeln erkennen und verarbeiten sollen: Den meisten Zielen von Programmen zur Sprach- und Texterkennung liegt wesentlich eine Abbildung der Realität zugrunde. Man möchte mithilfe der Programme die Sprache, die von Menschen verwendet wird, rekonstruieren und Programme dazu bringen, ebenso gut wie Menschen zu kommunizieren. Angestrebt wird hier eine Kommunikationssimulation – der Weg geht also andersherum als Günter Ropohl (1991) es prophezeite (s. S. 52). Bei der Entwicklung von Waschmaschinen und Einparkhilfen war ebenfalls von vornherein klar, zu welchem Zweck sie entwickelt wurden: zum Wäschewaschen und um Menschen das Einparken zu erleichtern. Anders bei der Entwicklung ethischer Techniksysteme: Hier ist oft unklar, an welchem Referenzobjekt oder Referenzverhalten man sich orientieren soll. *Richtiges Verhalten* oder *richtiges Handeln* ist nichts, was ein für alle Mal definiert werden kann; vielmehr entscheidet sich, was richtig ist, auf der Basis

[131] In diesem Fall müsste genauer untersucht werden, was die Entscheidungen$_T$ von menschlichen Entscheidungen unterscheidet.

unterschiedlicher Faktoren: abhängig von dem betroffenen Sozialkreis, abhängig von einer jeweiligen Kultur und dem spezifischen Kontext sowie abhängig von der die Handlung oder das Verhalten beurteilenden Person oder dem Personenkreis. Unabhängig von der Gefahr möglicher Wert-Verzerrungen, ist die Vorlage für die Handlungen$_T$ und Entscheidungen$_T$ ethischer Maschinen unklar, da die möglichen Situationen der Welt zu komplex sind, um sie verallgemeinern zu können.

Allerdings gibt es durchaus moralische Werte, die in demokratischen Gesellschaften beispielsweise durch das Einfordern mittels Gesetzen festgeschrieben werden, was als Referenzrahmen dienen könnte. Orwat und Bless (2016: 29 f.) plädieren beispielsweise dafür, dass Werte wie Menschenrechte institutionell verankert sein und durch Methoden der Rechtsstaatlichkeit gewährleistet werden müssen:

> Therefore, *institutional frameworks*, in the form of market law, consumer protection, privacy regulation, competition policy, and so on are needed in order to establish or stimulate markets, mitigate market failures or constrain abusive market behavior. Here, the institutional frameworks have functions to enhance the efficiency of markets for products and services that contribute to the realization of human rights or which do not violate human rights during production and consumption. (Herv. i. Orig.)

Ethische Aussagen haben keine Rechtsverbindlichkeit, sofern sie oder moralische Werte nicht Teil von Gesetzestexten und somit Gegenstand juristischen Rechts sind. Bislang gibt es zahlreiche Weißbücher, Stellungnahmen und Positionspapiere zur Regulierung von künstlich intelligenten Algorithmen, die häufig unter dem Stichwort *governance* vereint werden (s. Abschnitt 6.3). Stand heute sind es aber noch immer Menschen, die die grundlegenden Normen und Prinzipien für ein gutes Zusammenleben einer mehr oder weniger großen Gesellschaft festlegen müssten und könnten (vgl. hierzu auch Kuipers 2020: 440).

6.2 Ergänzung: Informationsverarbeitung und technische Realisierungen

Mit David Marr (2010: 25) können drei Ebenen unterschieden werden, „at which any machine carrying out an information-processing task must be understood“: die Ebene der *computational theory*, die Ebene von *representation and algorithm* und die Ebene der *hardware implementation*. Damit sind ihm zufolge folgende Fragen verbunden:

> What is the goal of the computation, and what is the logic of the strategy by which it can be carried out?
> How can this computational theory be implemented? In particular, what is the

> representation for the input and output, ans what is the algorithm for the transformation?
> How can the representation and algorithm be realized physically?
> (Marr 2010: 25)

Die Abwägung der Notwendigkeit kognitiver Prozesse kann auf der computationalen Ebene angesiedelt werden; die Überlegungen zur Umsetzung möglicher Theorien auf der algorithmischen Ebene. Auf technischer Ebene stellen sich dann je nach System andere Fragen, wie sie beispielsweise Soares und Fallenstein (2017: 4) aufwerfen:

> What, formally, is the induction problem faced by an intelligent agent embedded within and computed by its environment? What is the set of environments which embed the agent? What constitutes a simplicity prior over that set? How is the agent scored?

Konkret in Bezug auf automatisierte Fahrzeuge hält Benjamin Kuipers (2020: 424. Herv. i. Orig.) eine Differenzierung des ersten Robotergesetzes Isaac Asimovs für plausibel.

> *(SN-1) A robot will never* deliberately *harm a human being.*
> *(SN-2) In a given situation, a robot will be no more likely than a skilled and alert human to* accidentally *harm a human being.*[132]

Diese Gesetze bedürfen natürlich technischer Umsetzungen und auch Möglichkeiten, insbesondere bei automatisierten Fahrzeugen, z. B. die Fähigkeit des Fahrzeugs, anzuhalten, wenn die Regeln nicht eingehalten werden können (vgl. ebd.: 430). Die grundlegende Frage bei der Modellierung von Szenarien, bei der Formalisierung von Problemen und Verhaltensweisen, also der Umsetzung ethischer Denkstrukturen in technischen Systemen, scheint dabei, wie anhand der Diskussion der Theorien deutlich wurde, nicht die technische Umsetzung an sich zu sein, sondern die Tatsache, dass ethische Prinzipien sich nicht für sämtliche Fälle international verallgemeinern lassen – beziehungsweise: Ethische Theorien können bei der Suche möglicher technischer Lösungswege für bestimmte Probleme zwar konzeptionell hilfreich sein, in Bezug auf normatives Handeln stellen sie jedoch eher Unterstützungen für die Suche nach angemessenen Verfahrenswegen dar und beinhalten selbst noch keine normativen Richtlinien. Letzteres wird aber suggeriert, indem die Rede davon ist, dass es *eine ethische Theorie* geben könne oder ein *Bündel ethischer Prinzipien*, die in technischen Systemen verlässlicherweise zu ethisch gewünschten Zielen führen könnten.

[132] „SN“ steht für *social norm*.

6.3 Ethische Techniksysteme gemessen am *Weißbuch zur Künstlichen Intelligenz* der Europäischen Kommission

In ihrem *Weißbuch zur Künstlichen Intelligenz* spricht sich die Europäische Kommission einerseits für klare wirtschaftliche Ziele und das Bestreben, die „Technologieführerschaft der EU zu wahren und sicherzustellen", aus, andererseits für „Regulierung" (Europäische Kommission 2020: 1). Dass dies kein Widerspruch sein muss, zeigen Heesen et al. (2020a: 24 f.) in ihrem *Ethik-Briefing* anhand einiger Praxisbeispiele:

> Die Einhaltung von ethischen Werten und Prinzipien, die nicht vorrangig der Logik des Marktes folgen, kann dazu führen, im freien Wettbewerb Nachteile zu erleiden und damit wirtschaftlich abgehängt zu werden, so lautet eine gängige Befürchtung. Diese Sorge teilen die befragten Unternehmen jedoch nicht. [...] Viele Unternehmen sind sich der Bedeutung ethischer Werte zur Realisierung eines verantwortungsvollen Technologieentwicklungsprozesses nicht nur bewusst, sie haben auch bereits Maßnahmen ergriffen, wie die Einhaltung dieser Werte unter anderem gewährleistet werden kann.

Im Folgenden werde ich mich auf die Aspekte *Risiko* und *Vertrauenswürdigkeit* konzentrieren und ihre Definition im Weißbuch vor dem Hintergrund der Möglichkeit ethischer Maschinen diskutieren. Menschen müssten Technologien vertrauen können, schreibt die Europäische Kommission (2020: 1): „Vertrauenswürdigkeit ist eine Voraussetzung für ihre Akzeptanz" (ebd.: 11). Vollständig automatisierte, ethische Techniksysteme, wie ich sie in der vorliegenden Arbeit hypothetisch modelliert habe, sind dabei als Entwicklungsmöglichkeit bislang gar nicht vorgesehen.

> Die menschliche Aufsicht hilft, dafür zu sorgen, dass ein KI-System die menschliche Autonomie nicht untergräbt oder sich sonst nachteilig auswirkt. Das Ziel einer vertrauenswürdigen, ethischen und auf den Menschen ausgerichteten KI kann nur erreicht werden, wenn dafür gesorgt wird, dass Menschen bei KI-Anwendungen mit hohem Risiko gebührend mitwirken.
>
> (Ebd.: 25)

Die Achtung und Berücksichtigung europäischer Werte und Rechtsstaatlichkeit spielten für die Vertrauensbildung eine ebenso große Rolle wie für die Risikoregulierung. Von *Grundwerten* ist ferner die Rede (ebd.: 3). Die zugrundeliegenden Ziele – wirtschaftlicher Erfolg, Einhaltung (und gegebenenfalls Erweiterung) des rechtlichen Rahmens sowie die Einhaltung europäischer Werte – scheinen dabei gleich stark im Fokus sein. Ferner beruft sich die Kommission auf die bereits 2019 veröffentlichten *Ethik-Leitlinien für eine vertrauenswürdige KI* (Europäische Kommission 2019), in denen „sieben Kernanforderungen" (Europäische Kommission 2020: 11) festgehalten wurden:

- Vorrang menschlichen Handelns und menschlicher Aufsicht
- Technische Robustheit und Sicherheit
- Privatsphäre und Datenqualitätsmanagement
- Transparenz
- Vielfalt, Nichtdiskriminierung und Fairness
- Gesellschaftliches und ökologisches Wohlergehen und
- Rechenschaftspflicht

Um mögliche Risiken zu vermeiden, schlägt die Kommission vor, folgende Anforderungen zu berücksichtigen:

> - Anforderungen, die gewährleisten, dass die KI-Systeme in allen Phasen ihres Lebenszyklus robust und genau sind oder zumindest ihren Genauigkeitsgrad korrekt wiedergeben;
> - Anforderungen, die gewährleisten, dass die Ergebnisse reproduzierbar sind;
> - Anforderungen, die gewährleisten, dass KI-Systeme in allen Phasen ihres Lebenszyklus Fehler und Unstimmigkeiten angemessen bewältigen können;
> - Anforderungen, die gewährleisten, dass KI-Systeme sowohl gegen offene Angriffe als auch gegen subtilere Versuche, Daten oder die Algorithmen selbst zu manipulieren, widerstandsfähig sind und dass in solchen Fällen Abhilfemaßnahmen ergriffen werden.
> (Europäische Kommission 2020: 24 f.)

Beim Versuch der Anwendung dieser Vorschläge auf genuin ethische Techniksysteme fällt auf, dass die Anforderungen zwar einen Rahmen für mögliche Entwicklungen und Regulierungen darstellen können, darüber hinaus aber in konkreten Anwendungsfällen einer erheblichen Konkretisierung bedürften. Dies entspricht auch der oben erwähnten Einschätzung, dass die Diskussion der Anwendung ethischer Theorien in technischen Systemen nicht zielführend ist, sofern nicht über konkrete Werte und Normen gesprochen wird, die es umzusetzen gilt.

Lässt man die Möglichkeit zur menschlichen Aufsicht einmal außen vor und überlegt mögliche Handlungsspielräume für selbstständig denkende, argumentierende und agierende Programme und Maschinen, so hilft der beispielsweise angegebene Bezug auf *europäische Grundwerte* bei der Konkretisierung nicht weiter. Diese Angabe ist ähnlich konkret wie die Robotergesetze von Isaac Asimov (s. S. 57), demgemäß Menschen *keinen Schaden* im Umgang mit selbstständiger Technik nehmen sollen. Selbst wenn man europäische Werte gemäß Artikel 2 des *Vertrags über die Europäische Union* konkreter benennt – „Achtung der Menschenwürde“, „Freiheit“, „Demokratie“, „Gleichheit“, „Rechtsstaatlichkeit“, „Achtung der Menschenrechte, einschließlich der Rechte von Minderheiten“ (Europäische Union 2016) –,

müsste für konkrete Situationen oder Anwendungen definiert werden, was dies im Einzelnen bedeutet. Was bedeutet die Achtung der Menschenwürde im Kontext von Abtreibungsdebatten, zu denen ethische Maschinen sich äußern$_T$ könnten? Wie kann die Achtung der Menschenrechte im Kontext militärischer Szenarien gewahrt werden? Wie können automatisierte Fahrzeuge so gestaltet werden, dass sie gleichzeitig den Ansprüchen der Freiheit, aber auch denen der Demokratie und Rechtsstaatlichkeit gerecht werden? Auch ethische Techniksysteme, die hier für ethische Debatten Argumentationen liefern könnten, bräuchten vorab eine Konkretisierung der Definitionen der einzelnen Werte für bestimmte Situationen. In ihrer Stellungnahme zum Weißbuch betonen Orwat et al. (2020: 3),

> dass die zur Risikoregulierung notwendige Operationalisierung bzw. Konkretisierung ein aufwendiges Vorgehen von Wissenschaft und Politik und letztlich politisch-normative Abwägungen und Entscheidungen erfordern.

Sie empfehlen daher die „Stärkungen der Antidiskriminierungsstellen mit Kompetenzen und Befugnissen“, die „Verbesserung der Nachweismöglichkeiten und das Verbandsklagerecht“ sowie „Dokumentationen“ von Zugangsmöglichkeiten, „die nach der DSGVO erstellt werden müssen“ (ebd.). Diese Forderungen bewegen sich auf juristischer Ebene. Ethik ist, wie bereits betont, nicht in ähnlicher Weise verbindlich, wie juristische Gesetze es sind. Allerdings können und sollten ethische und philosophische Überlegungen die Grundlage für die Konkretisierung zentraler Begriffe, wie Diskriminierung und Transparenz, sein. Zusammengenommen mit Grundlagen aus den Ingenieurswissenschaften oder der Informatik, müsste die Umsetzung ethischer Techniksysteme vor dem Hintergrund technischer Machbarkeit, juristischer Kongruenz und ethischer Vertretbarkeit durchgeführt werden.

7 Fazit und Konsequenzen

> To make practical decisions, individual humans need concise and understandable ethical principles. For these principles to be useful for the long-term survival of the society, they must be explainable and teachable to individuals entering the society, such as children and immigrants. If intelligent nonhuman agents such as robots and corporations are to apply ethical principles to their own behavior, these principles must be capable of being learned or programmed.
> (Kuipers 2020: 424)

Die Arbeitshypothese für die vorliegende Arbeit war, dass gemäß der möglichen Unterscheidung von *Moral* und *Ethik* möglicherweise auch ein Unterschied zwischen moralischen und ethischen Techniksystemen bestehen könnte – dahingehend, dass moralisch-technische Akteure zu moralischem Verhalten in der Lage sind, wohingegen ethische Akteure auch zu ethischen Reflexionen, dem Aufstellen ethischer Theorien, dem Abwägen ethischer Argumente imstande wären und dementsprechenden Output in Form von Handlungsvorschlägen oder (je nach Art des technischen Systems) eigenen Handlungen liefern könnten. Diese Hypothese wurde unter der Voraussetzung aufgestellt, dass die Rede von *ethischen Menschen*, wenngleich unüblich, insofern sinnvoll ist, wenn man damit die Fähigkeit zu ebenjener Theoriearbeit auf einer reflektierenden Metaebene unterstellt.

Die Analyse hat zunächst ergeben, dass moralische Techniksysteme oder technische Systeme als moralische Akteure – entgegen durch Studien belegte, subjektiv empfundene Ängste oder medial vermittelte Vorbehalte gegen *moral machines* – insofern unproblematisch sind, da *Moral* über regelgeleitetes Handeln definiert werden kann und algorithmengestützte Computerprogramme ebenfalls einer logischen Abfolge von Regeln folgen. Technische Systeme wurden also als *moralische Akteure* definiert, wenn sie zuvor festgelegten, ethischen Prinzipien folgen und diese umsetzen können. Die moralische Güte bemisst sich dabei nicht an den Intentionen der Entwickler∗innen oder den Intentionen$_T$ eines technischen Systems, sondern an den Umsetzungsergebnissen, was als *technischer Pragmatismus* bezeichnet wurde. Denn auch technische Systeme, die gemäß den intendierten Regeln *funktionieren*, können zu unmoralischen Ergebnissen führen, wie anhand einiger Beispiele verdeutlicht wurde. Ob es sich bei moralischen Techniksystemen also zugleich um moralisch *gute* Techniksysteme (im Sinn der Regeleinhaltung am Maßstab menschlicher Normen) handelt, muss im Einzelfall geprüft werden.

Weiterhin wurden die theoretischen Bedingungen der Umsetzung einzelner Werte und kognitiver Fähigkeiten in technischen Systemen untersucht, wie Freiheit, Autonomie und Bewusstsein u. a., und es wurde die Notwendigkeit deren Vorliegen für ethische Reflexionen geprüft. Diese Analyse wurde als synoptische Hinführung an die Thematik verstanden, die der Erörterung des Konzepts hypothetischer

ethischer Techniksysteme dienlich sein kann, gleichwohl aber aufgrund der thematischen Breite nicht in der notwendigen Ausführlichkeit geleistet werden konnte. Die Analyse ergab, dass insbesondere ein gewisses Maß an Freiheit, Autonomie, Rationalität und Intentionalität auf der Basis eines breiten Verständnisses der Begriffe für ethische Techniksysteme als notwendig erachtet werden können, wobei es sinnvoll ist, bei technischen Systemen eine prinzipielle Analogie zu Menschen (im Gegensatz zur Annahme der Identität der Prozesse) auszugehen und daher von Freiheit$_T$ usw. zu sprechen; so können mögliche Missverständnisse im direkten Vergleich mit menschlichen Fähigkeiten und Möglichkeiten vermieden werden. Anschließend wurden einige Standardtheorien der Ethik, Utilitarismus, deontologische Ansätze und Tugenden, im Hinblick auf deren Umsetzbarkeit in technischen Systemen geprüft, um der eingangs aufgeworfenen Frage, welche Form von *Maschinenethik* am plausibelsten ist, näher zu kommen.

Zu Beginn der vorliegenden Arbeit wurden drei mögliche Definitionsvarianten von *Maschinenethik* eröffnet: 1. eine genuine Menschenethik im Umgang mit Maschinen und technischen Systemen, 2. eine Maschinenethik von Menschen für Maschinen und 3. eine genuine Maschinenethik, die von Maschinen selbst generiert wird (S. 30). Für die anschließende Analyse legte ich die Arbeitshypothesen zugrunde, dass sich moralische Techniksysteme grundlegend von ethischen Techniksystemen unterscheiden und somit die dritte Form von Maschinenethik plausibel sein könnte. Die Analyse hat jedoch gezeigt, dass die Umsetzung ethischer Theorien in technischen Systemen zwar sowohl aus konzeptioneller als auch technischer Sicht nicht unrealistisch ist, dass jedoch auch technische Systeme für konkrete Anwendungsbereiche und -situationen stets der vorherigen Festlegung menschlicher Normen und Werte – und somit immer einer menschlichen Ethik – bedürfen. Genuin ethische Techniksysteme sind unter dieser Voraussetzung nicht möglich.

Die Analyse fußte also auf der Abwägung zweier *extremer Positionen* auf die Frage, ob Maschinen eine *neue* Ethik brauchten (nach Rath, Karmasin und Krotz 2019: 3 f.), die zur Erinnerung hier noch einmal zitiert werden:

> Hier lassen sich zwei extreme Positionen ausmachen:
> eine, die annimmt, dass Maschinen stets die Folge menschlicher Handlungen sind, dass es also im Kern nur um technische Hilfsmittel menschlichen Wollens ginge, und dass deswegen weder in begründungstheoretischer noch in praktischer Hinsicht neue ethische Konzeptionen von Nöten seien,
> die andere Position, die annimmt, dass selbstlernende Systeme und künstliche Intelligenz auch einen autonomen Willen von Maschinen begründen, der, wenn schon nicht dem menschlichen gleichzusetzen, diesem zumindest vergleichbar wäre – es wäre also sehr wohl von Nöten, den anthropozentrischen Fokus moderner Ethik zu Gunsten einer breiteren Konzeption von Ethik aufzugeben.

Die Autoren verstehen

> Maschinen*ethik* im eigentlichen Sinne, also nicht nur [als; L. S.] eine maschinisierte Moralanwendung, [. . . sondern; L. S.] als eine Spezifizierung der Medienethik [. . .], sofern digitale Maschinen, denen Künstliche Intelligenz zuzuschreiben und ggf. sogar moralische Intelligenz zu unterstellen wäre, nicht nur Objekte, sondern Akteure medial vermittelter Normansprüche sind.
>
> (Rath, Karmasin und Krotz 2019: 8, Herv. i. Orig.)

Zugleich bedarf es ihnen zufolge keiner grundlegend breiteren Konzeption von Ethik; gleichwohl würden sich neue ethische Fragen in Bezug auf aktuelle und zukünftige Technikentwicklungen stellen (s. S. 32). Die vorliegende, mitunter metaethische, Analyse kann als Untermauerung der These von Rath, Karmasin und Krotz dienen, insofern sie zeigt, dass eine genuin von Maschinen erzeugte Ethik nicht realistisch ist, da selbst ethische Theorien auf der Basis begründeter Argumente mit konkreten Normen und Werten gefüllt werden müssen, die prinzipiell von Menschen generiert werden (müssen). Insofern könnte es zwar plausible sein, um die eingangs aufgeworfene Frage noch einmal aufzugreifen, ob ein technisches System eine Dissertation schreiben kann, dass ein Programm auf der Basis bestimmter Texte eine systematische Auswertung liefert, doch die Kriterien hierfür müsste, um dem Unterfangen Sinnhaftigkeit zu verleihen, von Menschen vorgegeben werden. Unter diesen Umständen muss die Frage verneint werden, zumal der Sinn einer technisch generierten Dissertation, der hier nur als Gedankenexperiment aufgeführt wurde, natürlich prinzipiell infrage gestellt werden kann.

Dieses Ergebnis bedeutet aber weder, dass technische Systeme immer vollumfänglich kontrollierbar sind, noch, dass sie nicht auch zu unmoralischen Ergebnissen führen können – sei es aus Versehen oder auf menschlicher Seite beabsichtigt. Die Bewertung von Techniksystemen innerhalb der Maschinenethik muss also gemäß Thilo Hagendorff (2020: 114) von einer bloßen Auflistung allgemein gültiger moralischer Prinzipien wegkommen und sich eher auf spezielle Anwendungsfälle fokussieren:

> Checkbox guidelines must not be the only ‚instruments' of AI ethics. A transition is required from a more deontologically oriented, action-restricting ethic based on universal abidance of principles and rules, to a situation-sensitive ethical approach based on virtues and personality dispositions, knowledge expansions, responsible autonomy and freedom of action. Such an AI ethics does not seek to subsume as many cases as possible under individual principles in an overgeneralizing way, but behaves sensitively towards individual situations and specific technical assemblages. Further, AI ethics should not try to discipline moral actors to adhere to normative principles, but emancipate them from potential inabilities to act self-responsibly on the basis of comprehensive knowledge, as well as empathy in situations where morally relevant decisions have to be made.

Die Konsequenzen, die aus den vorliegenden Ergebnissen gezogen werden können, sind, dass im Hinblick auf die nicht ausgeschlossene Möglichkeit ethischer Techniksysteme und der bereits vorhandenen Schwierigkeiten mit moralischen Techniksystemen die ethische Zielsetzung und Bewertung noch konsequenter in Produktionsprozesse verankert werden sollte. Bislang haben wir es noch nicht mit Formen von starker künstlicher Intelligenz zu tun, doch die Möglichkeit eigenständig denkender und entscheidender Maschinen erfordert umso mehr auf menschlicher Seite ein Klarheit darüber, welche Technikentwicklungen legitim und erwünscht sind und welche nicht. Neben den bereits genannten Vertreter∗innen starker KI, gibt es Autor∗innen, die eine Überholung von Menschen durch Maschinen, auch in moralischer und ethischer Hinsicht, durchaus für möglich halten: „If progress in artificial intelligence continues unabated, AI systems will eventually exceed humans in general reasoning ability“ (Soares und Fallenstein 2017: 1). Vielmehr ist auch denkbar, dass technisch generierte, ethische Werte nicht zwangsläufig zum Nachteil menschlichen Lebens sein müssten, sondern vielmehr als Revolution menschlicher Ethik verstanden werden könnten – ebenso wie die Erfindung der Dampfmaschine als solche gesehen wurde (alle möglichen Vor- *und* Nachteile inbegriffen). Gleichwohl wäre damit die *Gefahr* verbunden, dass Techniksysteme nicht im Sinne von Menschen entscheiden und Prioritäten setzen, die nicht im Sinne menschlicher Interessen wären. Doch auch unter menschlichen Machtverhältnissen können natürlich unmoralische Entscheidungen getroffen werden, wie die oben dargelegten Beispiele gezeigt haben:

> Unfortunately, there will always be individuals and corporations who develop systems for their own ends. That is, the goals and values they program into (ro)bots may not serve the good of humanity.[133]
>
> (Wallach und Allen 2009: 194 f.)

Eine Voraussetzung für eine gelingende Technikgestaltung ist, dass die Vorstellungen menschlich *guten* Zusammenlebens, die allen ethischen Überlegungen zugrundeliegen, immer wieder neu und für bestimmte Personen- und Gesellschaftskreise definiert werden. Um weitreichende Folgen zu generieren, kann dies nicht nur auf individueller Ebene passieren; technikethische Diskussionen müssen auch von Entscheidungsträger∗innen wie Politiker∗innen bei der Gestaltung der Gesetzgebung und möglicher wirtschaftlicher Regularien aufgegriffen und berücksichtigt werden, um die technische Zukunft bestmöglich gestalten zu können.

[133] Dabei ist *the good of humanity* wieder ein sehr großer Begriff, der im Detail spezifiziert werden müsste.

Siglenverzeichnis

An	Aristoteles (2011): *Über die Seele. Griechisch / Deutsch*. Hrsg. und übers. von Gernot Krapinger. Stuttgart: Reclam.
Disc. Meth. AT	Descartes (2011): *Discours de la méthode. Französisch / Deutsch*. Hrsg. von Christian Wohlers. Philosophische Bibliothek 624. Hamburg: Meiner.
GMS	Immanuel Kant (2016): *Grundlegung zur Metaphysik der Sitten*. Hrsg. und mit einer Einl. vers. von Bernd Kraft und Dieter Schönecker. 2., durchgesehene Aufl. Erstausgabe 1785, Johann Friedrich Hartknoch: Riga. Mit aktualisierter Einleitung und Bibliographie. Hamburg: Meiner.
KpV	Immanuel Kant (2003): *Kritik der praktischen Vernunft*. Hrsg. von Heiner Klemme und Horst Brandt. Erstausgabe 1788, Johann Friedrich Hartknoch: Riga. Mit einer Einl., Sachanmerkungen und Bibliographie von Heiner Klemme. Hamburg: Meiner.
KrV	Immanuel Kant (1998): *Kritik der reinen Vernunft*. Hrsg. von Jens Timmermann. Erstausgabe 1781, Johann Friedrich Hartknoch: Riga. Mit einer Bibliographie von Heiner Klemme. Hamburg: Meiner.
MS	Immanuel Kant (2009): *Metaphysik der Sitten*. Hrsg. von Wilhelm Weischedel. Werkausgabe Band VIII. Erstausgabe 1792, Königsberg: Nicolovius. Frankfurt a. M.: Suhrkamp.
NE	Aristoteles (2013): *Nikomachische Ethik*. Hrsg. und übers. von Ursula Wolf. 4. Aufl. Rowohlts Enzyklopädie. Reinbek bei Hamburg: Rowohlt.
Tractatus	Ludwig Wittgenstein (2006): *Tractatus logico-philosophicus*. 1. Aufl. 1984. Werkausgabe Band I. Für die vorliegende Ausgabe wurde der Text neu durchgesehen von Joachim Schulte. Frankfurt a. M.: Suhrkamp.

Literaturverzeichnis

Hinweise zur Darstellung: Die folgenden Literaturangaben sind alphabetisch geordnet. Bei mehreren Einträgen eine∗r Autor∗in ist die Reihenfolge chronologisch aufsteigend. Erscheinungen eine∗r Autor∗in im selben Jahr werden mit den kleinen Anfangsbuchstaben des Alphabets versehen. Herausgeberschaften mit anderen Autor∗innen werden separat aufgelistet.

Ach, Johann und Borchers, Dagmar, Hrsg. (2018): *Handbuch Tierethik. Grundlagen – Kontexte – Perspektiven*. Stuttgart: J. B. Metzler.

Allen, Colin, Smit, Iva und Wallach, Wendell (2015): „Artificial morality: Top-down, bottom-up, and hybrid approaches“. In: *Ethics and Information Technology* 7, S. 149–155.

Allen, Colin, Varner, Gary und Zinser, Jason (2000): „Prolegomena to any future artificial moral agent“. In: *Journal of Experimental & Theoretical Artificial Intelligence*, S. 251–261.

Allen, Nicholas D. et al. (2010): „StatsMonkey: A Data-Driven Sports Narrative Writer“. In: *AAAI Association for the Advancement of Artificial Intelligence*. url: https://www.aaai.org/ocs/index.php/FSS/FSS10/paper/viewPaper/2305. Aufgerufen am 22.01.2022.

Alvarez, Maria (2018): „Reasons for action, acting for reasons, and rationality“. In: *Synthese* 195, S. 3293–3310.

Anderson, Michael und Leigh Anderson, Susan (2011): *Machine Ethics*. Cambridge: Cambridge University Press.

Anderson, Susan Leigh (2008): „Asimov's „three laws of robotics“ and machine metaethics“. In: *AI & Society* 22.4, S. 477–493.

Angel, Leonard (2018): *How to Build a Conscious Machine*. 1. Aufl. 1989: Westview Press. London; New York: Routledge.

Annas, Julia (2011): *Intelligent Virtue*. Oxford: Oxford University Press.

Aristoteles (2011): *Über die Seele. Griechisch / Deutsch*. Hrsg. und übers. von Gernot Krapinger. Stuttgart: Reclam.

— (2013): *Nikomachische Ethik*. Hrsg. und übers. von Ursula Wolf. 4. Aufl. Rowohlts Enzyklopädie. Reinbek bei Hamburg: Rowohlt.

Asimov, Isaac (1942): „Runaround“. In: *I, robot*. New York: Astounding Science Fiction.

— (2016): *Ich, der Roboter*. Übers. von Otto Schrag. 1. Aufl. 1950, Nightfall Inc. Titel der amerikanischen Originalausgabe: *I, ROBOT*. München: Heyne.

Austin, John (1962): *How to Do Things with Words*. Cambridge, MA: Harvard University Press.

Awad, Edmond (2017): *Moral Machine: Perception of Moral Judgment Made by Machines*. URL: https://www.media.mit.edu/publications/moral-machine-perception-of-moral-judgment-made-by-machines/. Aufgerufen am 22.01.2022.

AWO Bundesverband e. V., Hrsg. (2019): *Armut im Lebensverlauf. Kindheit, Jugend und junges Erwachsenenalter*. URL: https://jugendsozialarbeit.news/wp-content/uploads/2019/11/191104_Br_Armut_im_CV_bf.pdf. Aufgerufen am 22.01.2022.

Banerjee, Jayanta (2020): „Cultural Relativism and Technology Transfer in Engineering Education". In: *ASEE Virtual Annual Conference*. URL: https://peer.asee.org/cultural-relativism-and-technology-transfer-in-engineering-education.pdf. Aufgerufen am 22.02.2022.

Baranzke, Heike (2018): „Verrohungsargument". In: *Handbuch Tierethik. Grundlagen – Kontexte – Perspektiven*. Hrsg. von Johann Ach und Dagmar Borchers. Stuttgart: J. B. Metzler, S. 219–224.

Bareis, Jascha und Katzenbach, Christian (2021): „Talking AI into Being: The Narratives and Imaginaries of National AI Strategies and Their Performative Politics". In: *Science, Technology, & Human Values*. URL: https://journals.sagepub.com/doi/full/10.1177/01622439211030007. Aufgerufen am 22.01.2022.

Barooah, Rituparna (2019): „Physiology of Emotion". In: *Application of Biomedical Engineering in Neuroscience*. Hrsg. von Paul Studip. Singapur: Springer, Singapore, S. 415–435.

Barredo Arrieta, Alejandro et al. (2020): „Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information Fusion* 58, S. 82–115.

Beckermann, Ansgar (2001): *Analytische Einführung in die Philosophie des Geistes*. 2., überarbeitete Aufl. Berlin; New York: de Gruyter.

— (2003): *Einführung in die Logik*. Berlin: de Gruyter.

Beier, Kathi (2010): *Selbsttäuschung (Grundthemen Philosophie)*. Berlin; New York: de Gruyter.

Bendel, Oliver (2017): „Die Maschine in der Moral". In: *Cyber Security Report*, S. 4–6.

— Hrsg. (2018a): *Handbuch Maschinenethik*. Wiesbaden: VS Verlag für Sozialwissenschaften.

— (2018b): „Wozu brauchen wir die Maschinenethik?" In: *Handbuch Maschinenethik*. Hrsg. von Oliver Bendel. Wiesbaden: VS Verlag für Sozialwissenschaften, S. 1–20.

Beta Writer (2019): *Lithium-Ion Batteries – A Machine-Generated Summary of Current Research*. London: Springer.

Beyerer, Jürgen und Martini, Peter, Hrsg. (2020): *Rise of Artificial Intelligence in Military Weapons Systems. Position Paper*. Fraunhofer Group for Defense

and Security VVS. URL: https://www.fraunhofer.de/content/dam/zv/de/forschungsthemen/schutz-sicherheit/rise-of-intelligent-systems-in-military-weapon-systems-position-paper-fraunhofer-vvs.pdf. Aufgerufen am 22.01.2022.

Bigman, Yochanan und Gray, Kurt (2018): „People are averse to machines making moral decisions". In: *Cognition* 181, S. 21–34.

Birnbacher, Dieter (2021): „Utilitarismus". In: *Handbuch Technikethik*. Hrsg. von Armin Grunwald und Rafaela Hillerbrand. 2., aktualisierte und erweiterte Aufl. Stuttgart: J. B. Metzler, S. 160–164.

Blackford, Russell, Hrsg. (2017): *Science Fiction and the Moral Imagination*. Cham: Springer International.

Bostrom, Nick (2014): *Superintelligence. Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Braga, Adriana und Logan, Robert (2017): „The Emperor of Strong AI Has No Cloths. Limits to Artificial Intelligence". In: *Information* 156.8, S. 1–21.

Brand, Lukas (2018): *Künstliche Tugend. Roboter als moralische Akteure*. Regensburg: Friedrich Pustet.

Bratton, William, Morgan, John und Malinowski, Sean (2009): „Fighting crime in the information age: The promise of predictive policing". In: *Annual meeting of the American Society of Criminology*. Philadelphia.

Braun, Hans-Joachim (2020): „Intuition". In: *Technikanthropologie. Handbuch für Wissenschaft und Studium*. Hrsg. von Martina Hessler und Kevin Liggieri. Baden-Baden: Nomos, S. 573–578.

Braun, Robert (2019): „Artificial Intelligence: Socio-Political Challenges of Delegating Human Decision-Making to Machines". In: *IHS Working Paper 6*. Wien: Institut für Höhere Studien (IHS). URL: https://nbn-resolving.org/urn:nbn:de:0168-ssoar-62489-4. Aufgerufen am 22.01.2022.

Brentano, Franz (2008): *Psychologie vom empirischen Standpunkt. Von der Klassifikation psychischer Phänomene*. 2. Aufl. Herausgegeben, mit einem Vorwort und einem Index versehen von Thomas Binder und Arkadiusz Chrudzimski. Eingeleitet von Mauro Antonelli. 1. Ausgabe 1874. Wahlstedt: Ontos.

Brieger, Julchen (2019): „Über die Unmöglichkeit einer kantisch handelnden Maschine". In: *Maschinenethik. Normative Grenzen autonomer Systeme*. Hrsg. von Matthias Rath, Friedrich Krotz und Matthias Karmasin. Wiesbaden: VS Verlag für Sozialwissenschaften, S. 107–120.

Broadie, Sarah (2015): „Soul and Body in Plato and Descartes". In: *Proceedings of the Aristotelian Society* 101.1, S. 295–308.

Bruckenberger, Ulrike et al. (2013): „The Good, The Bad, The Weird: Audience Evaluation of a „Real" Robot in Relation to Science Fiction and Mass Media". In: *International Conference on Social Robotics* 8239, S. 301–310. URL: https://doi.org/10.1007/978-3-319-02675-6_30. Aufgerufen am 22.01.2022.

Brédart, Serge (2021): „The Influence of Anthropomorphism on Giving Personal Names to Objects“. In: *Advances in Cognitive Psychology* 17.1, S. 33–37.
Burckhardt, Martin (2018): *Philosophie der Maschine*. Berlin: Matthes & Seitz.
Burkart, Lutz (1987): „Das Ende des Technikdeterminismus und die Folgen – soziologische Technikforschung vor neuen Aufgaben und neuen Problemen.“ In: *Technik und sozialer Wandel. Verhandlungen des 23. Deutschen Soziologentages in Hamburg 1986*. S. 34–52. URL: https://nbn-resolving.org/urn:nbn:de:0168-ssoar-100737. Aufgerufen am 22.01.2022.
Bölker, Michael, Gutmann, Mathias und Hesse, Wolfgang, Hrsg. (2010): *Information und Menschenbild*. Ethics of Science and Technology Assessment (Schriftenreihe der Europäischen Akademie zur Erforschung von Folgen wissenschaftlich-technischer Entwicklungen Bad Neuenahr-Ahrweiler GmbH), Band 37. Berlin, Heidelberg: Springer.
Cabanac, Guillaume, Labé, Cyril und Magazinov, Alexander (2022): „‚Bosom peril'is not ‚breast cancer': How weird computer-generated phrases help researchers find scientific publishing fraud“. In: *Bulletin of the Atomic Scientists*. URL: https://thebulletin.org/2022/01/bosom-peril-is-not-breast-cancer-how-weird-computer-generated-phrases-help-researchers-find-scientific-publishing-fraud/. Aufgerufen am 22.01.2022.
Callaghan, Victor et al. Hrsg. (2017): *The Technological Singularity: Managing the Journey*. Berlin, Heidelberg: Springer.
Čapek, Karel (2005): *R.U.R. – Rossums Universal Robots*. Übers. von Otto Pick. Berlin. Bearbeitung (Wort): Dieter Hasselblatt. Technische Realisierung: Hans Greb und Regine Müller. Regieassistenz: Alexander Malachovsky. Eine Produktion des Bayerischen Rundfunks/ RIAS Berlin, 1978.
— (2017): *W.U.R. Werstand's Universal Robots*. Übers. von Otto Pick. Berlin: Holzinger. Originaltitel: „R.U.R. Rossum's Universal Robots“. Erstdruck 1920, Übersetzung 1922.
Caporael, Linnda (1986): „Anthropomorphism and mechanomorphism: Two faces of the human machine“. In: *Computers in Human Behavior* 2.3, S. 215–234.
Capurro, Rafel (2016): „Informationsethik und kulturelle Vielfalt“. In: *Handbuch Medien- und Informationsethik*. Hrsg. von Jessica Heesen. Stuttgart: J. B. Metzler, S. 331–336.
Carter, Matt (2007): *Minds and Computers. An Introduction to the Philosophy of Artificial Intelligence*. Edinburgh: Edinburgh University Press.
Celikates, Robin und Gosepath, Stefan (2017a): „Einleitung: Grundbegriffe, Grundprobleme und Grundmodelle der Moralphilosophie“. In: *Philosophie der Moral. Texte von der Antike bis zur Gegenwart*. Hrsg. von Robin Celikates und Stefan Gosepath. 5. Aufl. Frankfurt a. M.: Suhrkamp, S. 7–27.

Celikates, Robin und Gosepath, Stefan, Hrsg. (2017b): *Philosophie der Moral. Texte von der Antike bis zur Gegenwart*. 5. Aufl. Frankfurt a. M.: Suhrkamp.
— (2017c): „*Immanuel Kant*. Grundlegung zur Metaphysik der Sitten“. In: *Philosophie der Moral. Texte von der Antike bis zur Gegenwart*. Hrsg. von Robin Celikates und Stefan Gosepath. 5. Aufl. Frankfurt a. M.: Suhrkamp, S. 203–222.
Centrone, Stefania (2021): „Leibniz und die künstliche Intelligenz. Lingua characteristica und Calculus ratiocinator“. In: *Philosophisches Handbuch Künstliche Intelligenz*. Hrsg. von Klaus Mainzer. Wiesbaden: Springer VS, S. 1–27.
Chrisley, Ron (2008): „Philosophical foundations of artificial consciousness“. In: *Artificial Intelligence in Medicine* 44.2, S. 119–137. url: https://doi.org/10.1016/j.artmed.2008.07.011. Aufgerufen am 22.01.2022.
— (2020): „A Human-Centered Approach to AI Ethics“. In: *The Oxford Handbook of Ethics of AI*. Hrsg. von Markus Dubber, Frank Pasquale und Sunit Das. Oxford: Oxford University Press, S. 463–474.
Christian, Brian (2021): *The Alignment Problem. How Can Artificial Intelligence Learn Human Values?* London: Atlantic Books. First published in the United States in 2020 by W. W. Norton & Company, Inc., New York.
Chu, Hang, Urtasun, Raquel und Fidler, Sanja (2017): „Song from PI: A musically plausible network for pop music generation“. In: *International Conference on Learning Representations Workshops (ICLRW)*, S. 1–9. url: https://arxiv.org/pdf/1611.03477. Aufgerufen am 22.01.2022.
Collins, Harry und Kusch, Martin (1999): *The Shape of Actions. What Humans and Machines Can Do*. Cambridge, MA: MIT Press.
Conitzer, Vincent et al. (2017): „Moral Decision Making Frameworks for Artificial Intelligence“. In: *Association for the Advancement of Artificial Intelligence*. url: https://users.cs.duke.edu/~conitzer/moralAAAI17.pdf. Aufgerufen am 22.01.2022.
Dale, Robert (2020): „GPT-3: What's it good for?“ In: *Natural Language Processing* 27.1, S. 113–118.
Danaher, John (2020): „Welcoming Robots into the Moral Circle: A Defence of Ethical Behaviorism“. In: *Science and Engineering Ethics* 26, S. 2023–2049.
Darling, Kate (2021): *The New Breed. How To Think About Robots*. 1. Ausgabe: Henry Holt and Company, 2021. Dublin: Allen Lane – Penguin.
Dastin, Jeffrey (2018): *Amazon scraps secret AI recruiting tool that showed bias against women*. url: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G. Aufgerufen am 22.01.2022.

Davidson, Daniel (1970): „Mental Events“. In: *Experience and Theory*. Hrsg. von Lawrence Foster und Joe Swanson. London: Duckworth.

de Bruin, Boudewijn und Floridi, Luciano (2017): „The ethics of cloud computing“. In: *Science and Engineering Ethics* 23.1, S. 21–39. url: https://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID3835151_code2644503.pdf?abstractid=3835151&mirid=1. Aufgerufen am 22.01.2022.

de Sousa, Ronald, Hrsg. (2019): *Die Rationalität des Gefühls*. Frankfurt a. M.: Suhrkamp.

Decker, Michael (2010): „Ein Abbild des Menschen: Humanoide Roboter“. In: *Information und Menschenbild*. Hrsg. von Michael Bölker, Mathias Gutmann und Wolfgang Hesse. Ethics of Science and Technology Assessment (Schriftenreihe der Europäischen Akademie zur Erforschung von Folgen wissenschaftlich-technischer Entwicklungen Bad Neuenahr-Ahrweiler GmbH). Bd. 37. Berlin; Heidelberg: Springer, S. 41–62.

— (2016): „Roboterethik“. In: *Handbuch Medien- und Informationsethik*. Hrsg. von Jessica Heesen. Stuttgart: J. B. Metzler, S. 351–357.

Dennett, Daniel (1971): „Intentional Systems“. In: *Journal of Philosophy* 68, S. 87–106.

Der Spiegel (2000): *‚Würgende Lügen‘ im Wortlaut*. 02.07.2000. url: https://www.spiegel.de/kultur/literatur/erstes-maschinengedicht-wuergende-luegen-im-wortlaut-a-83591.html. Aufgerufen am 22.01.2022.

Descartes (2011): *Discours de la méthode. Französisch / Deutsch*. Hrsg. von Christian Wohlers. Philosophische Bibliothek 624. Hamburg: Meiner.

Diamant, Emanuel (2015): *Advances in artificial intelligence: Are you sure, we are on the right track?* url: https://arxiv.org/abs/1502.04791. Aufgerufen am 22.01.2022.

Diestel, Lukas (2021): *falschegefühle.de*. url: http://falschegefuehle.de/info/. Aufgerufen am 31.10.2021.

Dolezal, Eugen und Windegger, Moritz (2020): „KI – Künstler oder Werkzeug?“ In: *LIMINA – Grazer Theologische Perspektiven* 3.2, S. 217–235. url: https://limina-graz.eu/index.php/limina/article/view/57. Aufgerufen am 22.01.2022.

Donner, Susanne (2016): „René Descartes – Vater der Leib-Seele-Theorie“. In: *dasgehirn.info*. url: https://www.dasgehirn.info/entdecken/meilensteine/rene-descartes-vater-der-leib-seele-theorie. Aufgerufen am 22.01.2022.

Drux, Rudolf, Hrsg. (1994a): *Die Geschöpfe des Prometheus – Der künstliche Mensch von der Antike bis zur Gegenwart*. Bielefeld: Kerber.

— (1994b): „Die Geschöpfe des Prometheus. Zur künstlerischen Gestaltung und technischen Verwirklichung eines Mythems“. In: *Die Geschöpfe des Prometheus*

– *Der künstliche Mensch von der Antike bis zur Gegenwart*. Hrsg. von Rudolf Drux. Bielefeld: Kerber, S. 15–26.

Dubber, Markus, Pasquale, Frank und Das, Sunit, Hrsg. (2020): *The Oxford Handbook of Ethics of AI*. Oxford: Oxford University Press.

Duda, Justine (2016): *Ein mittelalterlicher Computer Ein Rechenautomat aus Papier und Pergament. Raimundus Lullus und seine Ars generalis ultima*. Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften, Universität Trier, 15.05.2016. url: https://dhmuseum.uni-trier.de/node/354. Aufgerufen am 22.01.2022.

Dupuy, Jean-Pierre, Hrsg. (2001): *The Mechanization of the Mind*. Princeton: Princeton University Press.

Döring, Sabine, Hrsg. (2009): *Philosophie der Gefühle*. Frankfurt a. M.: Suhrkamp.

— (2019): „Gefühlte Bewertungen. Emotionen in Moral- und politischer Philosophie“. In: *Emotionen. Ein interdisziplinäres Handbuch*. Hrsg. von Hermann Kappelhoff et al. Stuttgart: J. B. Metzler, S. 375–384.

Eisele, Jörg (2016): „Cyberkriminalität“. In: *Handbuch Medien- und Informationsethik*. Hrsg. von Jessica Heesen. Stuttgart: J. B. Metzler, S. 255–260.

Elgammal, Ahmed und Saleh, Babak (2015): „Quantifying Creativity in Art Networks“. In: *Proceedings of the Sixth International Conference on Computational Creativity*, 39–46. url: https://arxiv.org/abs/1506.00711. Aufgerufen am 22.01.2022.

Elkins, Katherine und Chun, Jon (2020): „Can GPT-3 Pass a Writer's Turing Test?“ In: *Journal of Cultural Analytics*, S. 1–16. url: https://culturalanalytics.org/article/17212.pdf. Aufgerufen am 22.01.2022.

Engel, Gisela und Karafyllis, Nicole, Hrsg. (2004): *Technik in der Frühen Neuzeit, Schrittmacher der europäischen Moderne*. Frankfurt a. M.: Klostermann.

Enzensberger, Hans-Magnus, Hrsg. (2000): *Mausoleum. Siebenunddreißig Balladen aus der Geschichte des Fortschritts*. Bibliothek Suhrkamp. Frankfurt a. M.: Suhrkamp.

— Hrsg. (2015): *Einladung zu einem Poesie-Automaten*. Edition Suhrkamp. Frankfurt a. M.: Suhrkamp.

Epley, Nicholas, Waytz, Adam und Cacioppo, John (2007): „On Seeing Human: A Three-Factor Theory of Anthropomorphism“. In: *Psychological Review* 114.4, S. 864–886.

Ernst, Gerhard (2009): *Die Objektivität der Moral*. 2., unveränderte Aufl. Paderborn: mentis.

Europäische Kommission (2019): *Ethik-Leitlinien für eine vertrauenswürdige KI*. Brüssel. url: https://op.europa.eu/o/opportal-service/download-handler?

identifier=d3988569-0434-11ea-8c1f-01aa75ed71a1&format=pdf&language=de&productionSystem=cellar&part=. Aufgerufen am 06.11.2021.

Europäische Kommission (2020): *Weißbuch zur Künstlichen Intelligenz – ein europäisches Konzept für Exzellenz und Vertrauen*. Brüssel. url: https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_de.pdf. Aufgerufen am 07.12.2021.

Europäische Union (2016): *Vertrag über die Europäische Union (Konsolidierte Fassung)*. Brüssel. url: https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX%3A12016M%2FTXT. Aufgerufen am 22.01.2022.

Fischer, Bob, Hrsg. (2021): *Animal Ethics. A Contemporary Introduction*. London: Routledge.

Fischer, Sarah und Petersen, Thomas (2018): *Was Deutschland über Algorithmen weiß und denkt. Ergebnisse einer repräsentativen Bevölkerungsumfrage*. Gütersloh. url: https://www.bertelsmann-stiftung.de/de/publikationen/publikation/did/was-deutschland-ueber-algorithmen-weiss-und-denkt/. Aufgerufen am 22.01.2022.

Floridi, Luciano und Chiriatti, Massimo (2020): „GPT-3: It's Nature, Scope, Limits, and Consequences". In: *Minds and Machines* 30, S. 681–694.

Floridi, Luciano und Sanders, John (2004): „On the Morality of Artificial Agents". In: *Minds and Machine* 14, 349–379.

Ford, Kenneth M., Glymour, Clark und Hayes, Patrick, Hrsg. (1995): *Android Epistemology*. Cambridge, MA: MIT Press.

Frankena, William (1963): *Ethics*. Englewood Cliffs: Prentice-Hall.

Franssen, Maarten, Lokhorst, Gert-Jan und van de Poel, Ibo (2018): „Philosophy of Technology". In: *The Stanford Encyclopedia of Philosophy*. Hrsg. von Edward N. Zalta. url: https://plato.stanford.edu/archives/fall2018/entries/technology/. Aufgerufen am 22.01.2022.

Frede, Dorothea (2008): „Platon (428/27-348/47 v. Chr.)" In: *Klassiker der Philosophie. 1 – Von den Vorsokratikern bis David Hume*. Hrsg. von Otfried Höffe. München: C. H. Beck, S. 26–49.

Gaber, Mohamed Medhat, Hrsg. (2009): *Scientific Data Mining and Knowledge Discovery. Principles and Foundations*. Berlin; Heidelberg: Springer.

Gabriel, Gottfried et al. (2017): „Was leistet die Begriffsanalyse?" In: *Information Philosophie* 2, S. 36–44. url: https://www.information-philosophie.de/?a=1&t=8560&n=2&y=1&c=60. Aufgerufen am 22.02.2022.

Gamez, David (2008): „Progress in machine consciousness". In: *Consciousness and Cognition* 17.3, S. 887–910.

Gandorfer, Sarah (2015): „Big Data: ‚Shit in, Shit out'". In: *IT-Business*. url: https://www.it-business.de/big-data-shit-in-shit-out-a-485770/. Aufgerufen am 22.01.2022.

Gatys, Leon A., Ecker, Alexander S. und Bethge, Matthias (2015): „A Neural Algorithm of Artistic Style“. In: *Computer Vision and Pattern Recognition*. url: https://arxiv.org/pdf/1508.06576. Aufgerufen am 13.01.2021.

Gebru, Timnit et al. (2020): „Datasheets for datasets“. In: *arXiv*, S. 1–17.

Gettier, Edmund L. (1963): „Is Justified True Belief Knowledge?“ In: *Analysis* 23.6, S. 121–123. url: https://www.jstor.org/stable/3326922. Aufgerufen am 22.01.2022.

Gfrereis, Heike und Raulff, Ulrich, Hrsg. (2015): *Die Seele. Die Dauerausstellung im Literaturmuseum der Moderne*. 68. Marbach a. N.: Deutsche Schillergesellschaft.

Gibson, Rebecca (2020): *Desire in the Age of Robots and AI. An Investigation in Science Fiction and Fact*. Cham: Palgrave Pivot.

Gips, James (1995): „Toward the Ethical Robot“. In: *Android Epistemology*. Hrsg. von Kenneth M. Ford, Clark Glymour und Patrick Hayes. Cambridge, MA: MIT Press, S. 243–252.

Goldie, Peter (2009): „Emotionen und Gefühle“. In: *Philosophie der Gefühle*. Hrsg. von Sabine Döring. Frankfurt a. M.: Suhrkamp, S. 369–397.

Gosepath, Stefan, Hrsg. (2018): *Aufgeklärtes Eigeninteresse. Eine Theorie theoretischer und praktischer Raktionalität*. Frankfurt a. M.: Suhrkamp.

Graff, Bernd (2016): *Rassistischer Chat-Roboter: Mit falschen Werten bombardiert*. 3. April. url: http://www.sueddeutsche.de/digital/microsoft-program m-tay-rassistischer-chat-roboter-mit-falschen-werten-bombardiert-1.2928421. Aufgerufen am 22.01.2022.

Grimm, Petra, Keber, Tobias und Zöllner, Oliver, Hrsg. (2019): *Digitale Ethik. Leben in vernetzten Welten*. Stuttgart: Reclam.

Grunwald, Armin (2020): „Ethik und Technik“. In: *Technikanthropologie. Handbuch für Wissenschaft und Studium*. Hrsg. von Martina Hessler und Kevin Liggieri. Baden-Baden: Nomos, S. 69–82.

Grunwald, Armin und Hillerbrand, Rafaela, Hrsg. (2021): *Handbuch Technikethik*. 2., aktualisierte und erweiterte Aufl. Stuttgart: J. B. Metzler.

Gunkel, David, Hrsg. (2012): *The Machine Question. Critical Perspectives on AI, Robots, and Ethics*. Cambridge: Cambridge University Press.

— Hrsg. (2018): *Robot Rights*. Cambridge, MA: MIT Press.

Göcke, Benedikt Paul (2020): „Could Artificial General Intelligence be an End-In-Itself?“ In: *Artificial Intelligence. Reflections in Philosophy, Theology, and the Social Sciences*. Hrsg. von Benedikt Paul Göcke und Astrid Rosenthal-von der Pütten. Leiden; Boston: Brill; Mentis, S. 221–240.

Göcke, Benedikt Paul und Rosenthal-von der Pütten, Astrid, Hrsg. (2020): *Artificial Intelligence. Reflections in Philosophy, Theology, and the Social Sciences*. Leiden; Boston: Brill; Mentis.

Görz, Günther, Schmid, Ute und Braun, Tanya, Hrsg. (2021): *Handbuch der Künstlichen Intelligenz*. 6. Aufl. Berlin; Boston: de Gruyter.

Habermas, Jürgen (1991): *Erläuterungen zur Diskursethik*. Frankfurt a. M.: Suhrkamp.

— (2020): „Kommunikation und Handeln“. In: *Handlungstheorie. Eine Einführung*. Hrsg. von Wolfgang Bonss et al. 2. Aufl. Bielefeld: transcript Verlag, S. 255–270.

Hagendorff, Thilo (2020): „The Ethics of AI Ethics: An Evaluation of Guidelines“. In: *Minds and Machines* (30), S. 99–120.

Hancock, Peter, Hrsg. (2017): *Mind, Machine and Morality. Toward a Philosophy of Human-Technology Symbiosis*. Boca Raton: CRC Press; Taylor & Francis.

Heesen, Jessica, Hrsg. (2016): *Handbuch Medien- und Informationsethik*. Stuttgart: J. B. Metzler.

— (2021a): „Informationsethik“. In: *Handbuch Technikethik*. Hrsg. von Armin Grunwald und Rafaela Hillerbrand. 2., aktualisierte und erweiterte Aufl. Stuttgart: J. B. Metzler, S. 219–223.

— (2021b): „Wie kommt Ethik in die künstliche Intelligenz?“ In: *Digitale Welt* 5, S. 48–50.

Heesen, Jessica et al. (2020a): *Ethik-Briefing. Leitfaden für eine verantwortungsvolle Entwicklung und Anwendung von KI-Systemen – Whitepaper aus der Plattform Lernende Systeme*. Lernende Systeme – Die Plattform für Künstliche Intelligenz. München. URL: https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG3_Whitepaper_EB_200831.pdf. Aufgerufen am 22.01.2022.

Heesen, Jessica et al. (2020b): *Zertifizierung von KI-Systemen. Kompass für die Entwicklung und Anwendung vertrauenswürdiger KI-Systeme*. Lernende Systeme – Die Plattform für Künstliche Intelligenz. München. URL: https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG1_3_Whitepaper_Zertifizierung_KI_Systemen.pdf. Aufgerufen am 22.01.2022.

Hehl, Walter (2016): *Wechselwirkung. Wie Prinzipien der Software die Philosophie verändern*. Heidelberg: Springer.

Heil, Reinhard (2021): „Künstliche Intelligenz/ Maschinelles Lernen“. In: *Handbuch Technikethik*. Hrsg. von Armin Grunwald und Rafaela Hillerbrand. 2., aktualisierte und erweiterte Aufl. Stuttgart: J. B. Metzler, S. 424–428.

Heinaman, Robert (1990): „Aristotle and the Mind-Body Problem“. In: *Phronesis* 35.1, S. 83–102.

Hennen, Leonhard und Nierling, Linda (2019): „The politics of technology assessment. Introduction to the special issue“. In: *Technological Forecasting and Social Change* 139, S. 17–22.

Hessler, Martina (2020a): „Ersetzung“. In: *Technikanthropologie. Handbuch für Wissenschaft und Studium*. Hrsg. von Martina Hessler und Kevin Liggieri. Baden-Baden: Nomos, S. 263–269.
— (2020b): „Fehlerhafte Menschen“. In: *Technikanthropologie. Handbuch für Wissenschaft und Studium*. Hrsg. von Martina Hessler und Kevin Liggieri. Baden-Baden: Nomos, S. 303–307.
— (2020c): „Maschine“. In: *Technikanthropologie. Handbuch für Wissenschaft und Studium*. Hrsg. von Martina Hessler und Kevin Liggieri. Baden-Baden: Nomos, S. 256–262.
Hessler, Martina und Liggieri, Kevin, Hrsg. (2020): *Technikanthropologie. Handbuch für Wissenschaft und Studium*. Baden-Baden: Nomos.
Hildt, Elisabeth (2019): „Artificial Intelligence: Does Consciousness Matter?“ In: *Frontiers in Psychology*. url: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.01535/full. Aufgerufen am 22.01.2022.
Hillerbrand, Rafaela und Poznic, Michael (2021): „Tugendethik“. In: *Handbuch Technikethik*. Hrsg. von Armin Grunwald und Rafaela Hillerbrand. 2., aktualisierte und erweiterte Aufl. Stuttgart: J. B. Metzler, S. 165–170.
Himma, Kenneth Einar (2009): „Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent?“ In: *Ethics and Information Technology* 11, S. 19–29. url: https://link.springer.com/article/10.1007/s10676-008-9167-5. Aufgerufen am 22.01.2022.
Hubig, Christoph (2021): „Technik und Medium“. In: *Handbuch Technikethik*. Hrsg. von Armin Grunwald und Rafaela Hillerbrand. 2., aktualisierte und erweiterte Aufl. Stuttgart: J. B. Metzler, S. 123–127.
Höltgen, Stefan (2021): „Von der Sprachphilosophie zu ELIZA“. In: *Philosophisches Handbuch Künstliche Intelligenz*. Hrsg. von Klaus Mainzer. Wiesbaden: Springer VS, S. 1–22.
Hübner, Dietmar, Hrsg. (2021): *Einführung in die philosophische Ethik*. 3., erneut durchgesehene und korrigierte Aufl. Göttingen: Vandenhoeck & Ruprecht.
IBM (2021): *Explainable AI*. url: https://www.ibm.com/watson/explainable-ai. Aufgerufen am 22.01.2022.
Irrgang, Bernhard (1990): *Künstliche Intelligenz*. Stuttgart: Hirzel.
— (2011): „Technikphilosophie“. In: *Philosophie*. Hrsg. von Peggy Breitenstein und Johannes Rohbeck. Stuttgart: J. B. Metzler, S. 335–344.
— (2020): *Roboterbewusstsein, automatisiertes Entscheiden und Transhumanismus*. Würzburg: Königshausen & Neumann.
iStock (2018): *Nahaufnahme menschlicher Finger Touch Roboter Finger wie der Erschaffung Adams für nächsten Jahrhundert neue Ära Zukunft der Roboter-Age-Konzept – Stockfoto*. 10.08.2018. url: https://www.istockphoto.com/de/foto/nahaufnahme-menschlicher-finger-touch-roboter-finger-wie-der-

erschaffung-adams-f\%C3\%BCr-gm1015825948-273340442. Aufgerufen am 22.01.2022.

Jacob, Pierre (2019): „Intentionality“. In: *The Stanford Encyclopedia of Philosophy*. Hrsg. von Edward Zalta. url: https://plato.stanford.edu/archives/win2019/entries/intentionality/. Aufgerufen am 22.02.2022.

Janich, Peter (2021): „Kulturalistische Technik“. In: *Handbuch Technikethik*. Hrsg. von Armin Grunwald und Rafaela Hillerbrand. 2., aktualisierte und erweiterte Aufl. Stuttgart: J. B. Metzler, S. 104–108.

Jansen, Philip et al. (2016): *A reasoned proposal for shared approaches to ethics assessment in the European context*. Deliverable D4.1. url: https://research.utwente.nl/files/52452462/D4.1_Proposal_Ethics_Assessment_Framework.pdf. Aufgerufen am 22.01.2022.

Johnson, Deborah und Snapper, John (1985): „Ethical issues in the use of Computers“. In: *Metaphilosophy* 16.4, S. 322–326.

Kainz, Lisa (2021): *3 Gründe, warum Bio-Fleisch aus „guter Haltung“ nicht reicht*. 08.11.2021. url: https://www.peta.de/themen/bio-fleisch-gute-haltung/. Aufgerufen am 22.01.2022.

Kant, Immanuel (1998): *Kritik der reinen Vernunft*. Hrsg. von Jens Timmermann. Erstausgabe 1781, Johann Friedrich Hartknoch: Riga. Mit einer Bibliographie von Heiner Klemme. Hamburg: Meiner.

— (2003): *Kritik der praktischen Vernunft*. Hrsg. von Heiner Klemme und Horst Brandt. Erstausgabe 1788, Johann Friedrich Hartknoch: Riga. Mit einer Einl., Sachanmerkungen und Bibliographie von Heiner Klemme. Hamburg: Meiner.

— (2009): *Metaphysik der Sitten*. Hrsg. von Wilhelm Weischedel. Werkausgabe Band VIII. Erstausgabe 1792, Königsberg: Nicolovius. Frankfurt a. M.: Suhrkamp.

— (2016): *Grundlegung zur Metaphysik der Sitten*. Hrsg. und mit einer Einl. vers. von Bernd Kraft und Dieter Schönecker. 2., durchgesehene Aufl. Erstausgabe 1785, Johann Friedrich Hartknoch: Riga. Mit aktualisierter Einleitung und Bibliographie. Hamburg: Meiner.

Kappelhoff, Hermann et al. Hrsg. (2019): *Emotionen. Ein interdisziplinäres Handbuch*. Stuttgart: J. B. Metzler.

Keil, Geert (2009): *Willensfreiheit und Determinismus*. Stuttgart: Reclam.

Kerschreiter, Adrian (2017): „Nach dem Nokia 3310: Nächster Retro-Klassiker kehrt zurück“. In: *Chip*. 12.04.2017. url: https://www.chip.de/news/Original-Tamagotchi-guenstig-kaufen-Bei-diesen-Haendlern-koennen-Sie-bestellen_112756753.html. Aufgerufen am 22.01.2022.

Kirchin, Simon (2013): *Thick concepts*. Oxford: Oxford University Press.

Kitamura, Tadashi, Tahara, Tomoko und Asami, Ken-Ichi (2000): „How can a robot have consciousness?“ In: *Advanced Robotics* 14.4, S. 263–275.

Kitchin, Rob (2017): „Thinking critically about and researching algorithms“. In: *Information, Communication & Society* 1.20, S. 1–31. URL: http://mural.maynoothuniversity.ie/11591/1/Kitchin_Thinking_2017.pdf. Aufgerufen am 22.01.2022.

Kitchin, Rob und Dodge, Martin: *Code/Space: Software and Everyday Life*. Cambridge, MA: MIT Press.

Knobel, Cory und Bowker, Geoffrey (2011): „Values in Design“. In: *Communication of the ACM* 54.7, S. 26–28.

Korb, Kevin (2004): „Introduction: Machine Learning as Philosophy of Science“. In: *Minds and Machines* 14, S. 433–440.

Kornwachs, Klaus (2021): „Positionen der Technikphilosophie“. In: *Philosophisches Handbuch Künstliche Intelligenz*. Hrsg. von Klaus Mainzer. Wiesbaden: Springer VS, S. 1–44.

Korsgaard, Christine (2017): „Mit Tieren interagieren: Ein kantianischer Ansatz“. In: *Tierethik. Grundlagentexte*. Hrsg. von Friederike Schmitz. Frankfurt a. M.: Suhrkamp, S. 243–286.

Krzanowski, Roman et al. (2016): „Is Machine Ethics computable, non-computable or nonsensical?“ In: *Machine Ethics and Machine Law*, S. 8–10. URL: https://www.researchgate.net/publication/320324648_Is_Machine_Ethics_computable_non-computable_or_nonsensical. Aufgerufen am 22.01.2022.

Krüger, Oliver (2020): „Technologische Singularität“. In: *Technikanthropologie. Handbuch für Wissenschaft und Studium*. Hrsg. von Martina Hessler und Kevin Liggieri. Baden-Baden: Nomos, S. 270–276.

Kuipers, Benjamin (2020): „Perspectives on Ethics of AI: Computer Science“. In: *The Oxford Handbook of Ethics of AI*. Hrsg. von Markus Dubber, Frank Pasquale und Sunit Das. Oxford: Oxford University Press, S. 421–441.

Kurzweil, Raymond (2005): *The Singularity Is Near: When Humans Transcend Biology*. London: Viking – Penguin.

Larry, Alexander und Moore, Michael (2021): „Deontological Ethics“. In: *The Stanford Encyclopedia of Philosophy*. Hrsg. von Edward Zalta. URL: https://plato.stanford.edu/entries/ethics-deontological/. Aufgerufen am 22.01.2022.

Latour, Renate Hackel-de (2015): „Automaten, Algorithmen, Drohnen. Die Hilfstruppen des Journalismus – Potentiale, Grenzen, Gefahren“. In: *Communicatio Socialis* 48.1, S. 4–5.

Laugée, Françoise (2014): „Robots et journalistes, l'info data-driven“. In: *La revue européenne des médias et du numérique* 32. URL: https://la-rem.eu/2014/12/robots-et-journalistes-linfo-data-driven/. Aufgerufen am 22.01.2022.

Le Bui, Matthew und Noble, Safiya Umoja (2020): „We're missing a moral framework of justice in artificial intelligence. On the Limits, Failings, and Ethics of Fairness“. In: *The Oxford Handbook of Ethics of AI*. Hrsg. von Markus

Dubber, Frank Pasquale und Sunit Das. Oxford: Oxford University Press, S. 163–179.
Leben, Derek (2018): *Ethics for Robots. How to Design a Moral Algorithm*. London: Routledge.
Lee-Peuker, Mi-Yong, Scholtes, Fabian und Schumann, Olaf, Hrsg. (2007): *Kultur – Ökonomie – Ethik*. München: Hampp.
Lenk, Hans und Ropohl, Günter (1993a): „Einführung. Technik zwischen Können und Sollen". In: *Technik und Ethik*. Hrsg. von Hans Lenk und Günter Ropohl. 2. Aufl. Stuttgart: Reclam, S. 5–21.
— Hrsg. (1993b): *Technik und Ethik*. 2. Aufl. Stuttgart: Reclam.
Lerner, Jennifer, Valdesolo, Piercarlo und Kassam, Karim (2014): „Emotion and Decision Making". In: *Annual Review of Psychology* 66, S. 799–823.
Levy, David (2009): „The Ethical Treatment of Artificial Conscious Robots". In: *International Journal of Social Robotics* 1, S. 209–216.
Life Institute, The Future of (2021): *Benefits & Risks of Artificial Intelligence*. url: https://futureoflife.org/background/benefits-risks-of-artificial-intelligence/?cn-reloaded=1. Aufgerufen am 01.03.2022.
Loh, Janina (2017): „Roboterethik. Janina Loh über eine noch junge Bereichsethik". In: *Information Philosophie* 1. Hrsg. von Peter Moser, S. 20–33. Aufgerufen am 29.12.2018.
— (2019a): *Roboterethik. Eine Einführung*. Berlin: Suhrkamp.
— (2019b): „Verantwortung und Roboterethik. Ein Überblick und kritische Reflexionen". In: *Maschinenethik. Normative Grenzen autonomer Systeme*. Hrsg. von Matthias Rath, Friedrich Krotz und Matthias Karmasin. Wiesbaden: VS Verlag für Sozialwissenschaften, S. 91–106.
Lopez-Jimenez, Francisco et al. (2020): „Artificial Intelligence in Cardiology: Present and Future". In: *Mayo Clinic Proceedings* 95.5, S. 1015–1039.
Loriot (1972a): *Loriots großer Ratgeber*. Berlin; Darmstadt; Wien: Bertelsmann. Lizenzausgabe mit Genehmigung des Diogenes Verlags AG, Zürich.
— (1972b): „Umgang mit Robotern". In: *Loriots großer Ratgeber*. Berlin; Darmstadt; Wien: Bertelsmann, S. 67–68. Lizenzausgabe mit Genehmigung des Diogenes Verlags AG, Zürich.
Lullus, Raimundus (1999): *Ars brevis*. Hrsg., übers. und mit einer Einl. vers. von Alexander Fidora. Hamburg: Meiner.
Lutz-Bachmann, Matthias (2013a): *Ethik*. Grundkurs Philosophie, Band 7. Stuttgart: Reclam.
— (2013b): „Philosophische Ethik". In: *Ethik*. Grundkurs Philosophie, Band 7. Stuttgart: Reclam, S. 13–26.
Maas, Stefan (2017): „Geschichte der Zukunft. Joachim Radkau vergleicht Erwartungen und Realität". In: *Deutschlandfunk*. url: https://www.deutschlandfu

nk.de/geschichte-der-zukunft-joachim-radkau-vergleicht.1310.de.html?dram:article_id=377706. Aufgerufen am 22.01.2022.

Mainzer, Klaus, Hrsg. (2021): *Philosophisches Handbuch Künstliche Intelligenz*. Wiesbaden: Springer VS.

Mamak, Kamil (2021): *Whether to Save a Robot or a Human: On the Ethical and Legal Limits of Protections for Robots*. 07.07.2021. URL: https://www.frontiersin.org/articles/10.3389/frobt.2021.712427/full. Aufgerufen am 22.01.2022.

Marchetti, Giorgio (2018): „Consciousness: a unique way of processing information". In: *Cognitive Processing* 19, S. 435–464. URL: https://link.springer.com/article/10.1007/s10339-018-0855-8. Aufgerufen am 22.01.2022.

Marr, David (2010): *Vision – A Computational Investigation into the Human Representation and Processing of Visual Information*. Cambridge, MA; London: MIT Press.

Marsiske, Hans-Arthur (2021): „Robotik und Genderforschung: ‚Roboter sind queer'". In: *heise online*. URL: https://www.heise.de/news/Robotik-und-Genderforschung-Roboter-sind-queer-6165082.html. Aufgerufen am 22.01.2022.

Mayor, Adrienne (2018): *Gods and Robots. Myths, Machines and Ancient Dreams of Technology*. Princeton: Princeton University Press.

McCorduck, Pamela (2004): *Machines Who Think. A Personal Inquiry into the History and Prospects of Artificial Intelligence*. 2. Aufl. Natick: A. K. Peters.

McEwan, Ian (2019): *Maschinen wie ich*. Zürich: Diogenes.

McGlynn, Sean et al. (2017): „Understanding the potential of PARO for healthy older adults". In: *International Journal of Human-Computer Studies* 100, S. 33–47.

Meinecke, Lisa und Voss, Laura (2018): „I Robot, You Unemployed: Robotics in Science Fiction and Media Discourse". In: *Schafft Wissen: Gemeinsames und geteiltes Wissen in Wissenschaft und Technik. Proceedings der 2. Tagung des Nachwuchsnetzwerks ‚INSIST', 07.–08. Oktober 2016, München*. Hrsg. von Julia Engelschalt et al. S. 203–221. URL: https://www.ssoar.info/ssoar/bitstream/handle/document/58220/ssoar-2018-engelschalt_et_al-Schafft_Wissen_Gemeinsames_und_geteiltes.pdf?sequence=1&isAllowed=y&lnkname=ssoar-2018-engelschalt_et_al-Schafft_Wissen_Gemeinsames_und_geteiltes.pdf. Aufgerufen am 01.03.2022.

Mey, Günter und Mruck, Katja, Hrsg. (2018a): *Handbuch Qualitative Forschung in der Psychologie*. Wiesbaden: Springer.

— (2018b): „Qualitative Interviews in der psychologischen Forschung". In: *Handbuch Qualitative Forschung in der Psychologie*. Hrsg. von Günter Mey und Katja Mruck. Wiesbaden: Springer, S. 1–21.

Michelfelder, Diane und Doorn, Neelke, Hrsg. (2021): *The Routledge Handbook of the Philosophy of Engineering*. New York: Taylor & Francis.

Millar, Jason (2020): „Social Failure Modes in Technology and the Ethics of AI: An Engineering Perspective“. In: *The Oxford Handbook of Ethics of AI*. Hrsg. von Markus Dubber, Frank Pasquale und Sunit Das. Oxford: Oxford University Press, S. 443–461.

Misselhorn, Catrin (2018): „Maschinenethik und Philosophie“. In: *Handbuch Maschinenethik*. Hrsg. von Oliver Bendel. Wiesbaden: VS Verlag für Sozialwissenschaften, S. 1–23.

— (2019): *Grundfragen der Maschinenethik*. 4., durchgesehene und überarbeitete Aufl. Stuttgart: Reclam.

Mitcham, Carl, Hrsg. (2005): *Encyclopedia of Science, Technology and Ethics*. New York: Macmillan.

Moor, James H. (2006): „The Nature, Importance, and Difficulty of Machine Ethics“. In: *IEEE Intelligent Systems* 21.4, S. 18–21.

Mosakas, Kestutis (2021): „On the moral status of social robots: considering the consciousness criterion“. In: *AI & Society* 36, S. 429–443. https://link.springer.com/article/10.1007/s00146-020-01002-1. Aufgerufen am 22.01.2022.

Mühlbauer, Peter (2013): „Drei Seiten geteilt durch sieben Autoren ist gleich ein Dr. med.“ In: *heise online*. 03.06.2013. url: https://www.heise.de/tp/features/Drei-Seiten-geteilt-durch-sieben-Autoren-ist-gleich-ein-Dr-med-3399548.html. Aufgerufen am 22.01.2022.

Müller-Salo, Johannes, Hrsg. (2020): *Analytische Philosophie. Eine Einführung in 16 Fragen und Antworten*. Wien: Böhlau.

Nagel, Thomas (1974): „What Is It Like to Be a Bat?“ In: *The Philosophical Review* 83.4, S. 435–450.

Nath, Rajakishore (2009): *Philosophy of Artificial Intelligence: A Critique of the Mechanistic Theory of Mind*. Irvine: Universal Publishers.

Nida-Rümelin, Julian, Hrsg. (2005): *Angewandte Ethik. Die Bereichsethiken und ihre theoretische Fundierung. Ein Handbuch*. 2. Aufl. Stuttgart: Alfred Kröner.

Nierling, Linda und Torgersen, Helge (2019): „Normativität in der Technikfolgenabschätzung. Einleitung in das TATuP-Thema“. In: *TATuP – Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis* 1.28, S. 11–14.

Nissenbaum, Helen (2005): „Values in Technical Design“. In: *Encyclopedia of Science, Technology and Ethics*. Hrsg. von Carl Mitcham. New York: Macmillan, S. xvi–xx.

Nozick, Robert (1974): *Anarchy, State and Utopia*. Berlin: Basic Books.

Offray de La Mettrie, Julien (1909): *Der Mensch – eine Maschine (L'homme machine, 1946)*. Übers. von Max Brahn. Leipzig: Verlag der Dürr'schen Buchhandlung.

Orwat, Carsten und Bless, Roland (2016): „Values and Networks – Steps Toward Exploring their Relationships“. In: *ACM SIGCOMM Computer Communication Review* 46.2.

Orwat, Carsten et al. (2020): *Risikoregulierung der KI: normative Herausforderungen und politische Entscheidungen. Stellungnahme zum Weißbuch der Europäischen Kommission „Zur Künstlichen Intelligenz – ein europäisches Konzept für Exzellenz und Vertrauen“*. Karlsruher Institut für Technologie (KIT). url: https://publikationen.bibliothek.kit.edu/1000121489/85506536. Aufgerufen am 22.01.2022.

Paltrinieri, Nicola, Comfort, Louise und Reniers, Genserik (2019): „Learning about risk: Machine learning for risk assessment“. In: *Safety Science* 118, S. 475–486.

Pape, Wilhelm (1954): *Griechisch-deutsches Handwörterbuch*. Nachdruck der 3. Aufl., bearbeitet von M. Sengbusch. Graz: Vieweg.

Parthemore, Joel und Blay, Whitby (2013): „What Makes Any Agent a Moral Agent? Reflections on Machine Consciousness And Moral Agency“. In: *International Journal of Machine Consciousness* 5.2, S. 105–129.

Pasquale, Frank (2015): *The Black Box Society. The Secret Algorithms That Control Money and Information*. Cambridge, MA: Harvard University Press.

Pastor, Oscar und Iborra, José (2004): *Automatic software production system*. url: https://patents.google.com/patent/US6681383B1/en. Aufgerufen am 22.01.2022.

Pavaloiu, Alica und Kose, Utku (2017): „Ethical Artificial Intelligence – An Open Question“. In: *Journal of Multidisciplinary Developments* 2.2, S. 15–27.

Pepperell, Robert (2018): „Consciousness as a Physical Process Caused by the Organization of Energy in the Brain“. In: *Frontiers in Psychology*. url: https://www.frontiersin.org/articles/10.3389/fpsyg.2018.02091/full. Aufgerufen am 22.01.2022.

Pieper, Annemarie, Hrsg. (2017): *Einführung in die Ethik*. 7., aktualisierte Aufl. Tübingen: A. Francke.

Pirner, Manfred und Rath, Matthias, Hrsg. (2003): *Homo Medialis. Perspektiven und Probleme einer Anthropologie der Medien*. München: Kopäd.

Pitt, Joseph C. (1999): *Thinking About Technology: Foundations of the Philosophy of Technology*. New York: Seven Bridges Press.

Powers, Thomas (2011): „Prospects for a Kantian Machine“. In: *Machine Ethics*. Hrsg. von Michael Anderson und Susan Leigh Anderson. Cambridge: Cambridge University Press, S. 464–475.

Prinzing, Marlies, Debatin, Bernhard und Köberer, Nina, Hrsg. (2020): *Kommunikations- und Medienethik reloaded? Wegmarken für eine Orientierungssuche im Digitalen*. Baden-Baden: Nomos.

Quante, Michael (2020): *Philosophische Handlungstheorie*. Wien: Böhlau.
Radder, Hans (2009): „Why Technologies are Inherently Normative“. In: *Philosophy of Technology and Engineering Sciences. Handbook of the Philosophy of Science*. Hrsg. von Anthonie Meijers. Burlington: Elsevier.
Radkau, Joachim (2017): *Geschichte der Zukunft: Prognosen, Visionen, Irrungen in Deutschland von 1945 bis heute*. München: Hanser.
Ramge, Thomas (2019): *Who's afraid of AI? Fear and Promise in the Age of Thinking Machines*. New York: The Experiment.
Rath, Matthias (2014): *Ethik der mediatisierten Welt. Grundlagen und Perspektiven*. Wiesbaden: Springer Fachmedien.
— (2019): „Zur Verantwortungsfähigkeit künstlicher ‚moralische Akteure'. Problemanzeige oder Ablenkungsmanöver?“ In: *Maschinenethik. Normative Grenzen autonomer Systeme*. Hrsg. von Matthias Rath, Friedrich Krotz und Matthias Karmasin. Wiesbaden: VS Verlag für Sozialwissenschaften, S. 223–242.
— (2020): „Sind „moral machines“ kulturrelativ? Maschinenethische Anmerkungen zu einem psychologisch-informatischen Experiment“. In: *Kommunikations- und Medienethik reloaded? Wegmarken für eine Orientierungssuche im Digitalen*. Hrsg. von Marlies Prinzing, Bernhard Debatin und Nina Köberer. Baden-Baden: Nomos, S. 105–114.
Rath, Matthias, Karmasin, Matthias und Krotz, Friedrich (2019): „Brauchen Maschinen Ethik? Begründungstheoretische und praktische Herausforderungen“. In: *Maschinenethik. Normative Grenzen autonomer Systeme*. Hrsg. von Matthias Rath, Friedrich Krotz und Matthias Karmasin. Wiesbaden: VS Verlag für Sozialwissenschaften, S. 1–10.
Rath, Matthias, Krotz, Friedrich und Karmasin, Matthias, Hrsg. (2019): *Maschinenethik. Normative Grenzen autonomer Systeme*. Wiesbaden: VS Verlag für Sozialwissenschaften.
Rawls, John (1971): *A Theory of Justice*. Cambridge, MA: Harvard University Press.
Reggia, James (2013): „The rise of machine consciousness: studying consciousness with computational models“. In: *Neural Networks. The Official Journal of the International Neural Network Society* 44, S. 112–131.
Richardson, Rashida, Schultz, Jason und Crawford, Kate (2019): „Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice“. In: *New York University Law Review* 192, S. 192–164. url: https://ssrn.com/abstract=3333423. Aufgerufen am 22.01.2022.
Richter, Philipp (2016): „Big Data“. In: *Handbuch Medien- und Informationsethik*. Hrsg. von Jessica Heesen. Stuttgart: J. B. Metzler, S. 210–216.

Richterich, Annika (2018): „The Big Data Agenda: Data Ethics and Critical Data Studies“. In: London: University of Westminster Press.
Rieder, Gernot (2021): „Big Data“. In: *Handbuch Technikethik*. Hrsg. von Armin Grunwald und Rafaela Hillerbrand. 2., aktualisierte und erweiterte Aufl. Stuttgart: J. B. Metzler, S. 310–314.
Robinson, Howard (2009): „Mind and Body in Aristotle“. In: *The Classical Quaterly* 28.1, S. 105–124.
Rollin, Bernard (2011): *The Unheeded Cry. Animal Consciousness, Animal Pain, and Science*. Oxford: Oxford University Press.
Ropohl, Günter (1985): *Die unvollkommene Technik*. Suhrkamp Taschenbuch. Frankfurt a. M.: Suhrkamp.
— (1991): *Technologische Aufklärung. Beiträge zur Technikphilosophie*. Frankfurt a. M.: Suhrkamp.
Ros, Arno (2005a): *Materie und Geist. Eine philosophische Untersuchung*. Paderborn: mentis.
— (2005b): „Sind psychische Phänomene intentionale Phänomene?“ In: *Materie und Geist. Eine philosophische Untersuchung*. Paderborn: mentis, S. 166–179.
Rothaas, Julia (2016): „Die Roboterfrau hat Angst, fett auszusehen“. In: *Süddeutsche Zeitung*. 03.12.2016. URL: https://www.sueddeutsche.de/digital/kuenstliche-intelligenz-die-roboterfrau-hat-angst-fett-auszusehen-1.3276843. Aufgerufen am 22.01.2022.
Russell, Daniel, Hrsg. (2013): *The Cambridge Companion to Virtue Ethics*. Cambridge: Cambridge University Press.
Russell, Stuart und Norvig, Peter, Hrsg. (2010): *Artificial Intelligence. A Modern Approach*. 3. Aufl. Upper Saddle River: Pearson.
— Hrsg. (2012): *Künstliche Intelligenz. Ein moderner Ansatz*. 3. Aufl. München: Pearson.
Ryle, Gilbert (1945): „Knowing How and Knowing That. The Presidential Address“. In: *Proceedings of the Aristotelian Society* 46, S. 1–16. URL: https://www.informationphilosopher.com/solutions/philosophers/ryle/Ryle_KnowHow.pdf.
Schippl, Jens und Hillerbrand, Rafaela (2021): „Automatisiertes Fahren“. In: *Handbuch Technikethik*. Hrsg. von Armin Grunwald und Rafaela Hillerbrand. 2., aktualisierte und erweiterte Aufl. Stuttgart: J. B. Metzler, S. 378–382.
Schmidt, Jan-Hinrik (2016): „Ethik des Internets“. In: *Handbuch Medien- und Informationsethik*. Hrsg. von Jessica Heesen. Stuttgart: J. B. Metzler, S. 284–291.
Schmidt, Sebastian (2017): „Why we should promote irrationality“. In: *Grazer Philosophische Studien* 94.4, S. 605–615.

Schmidt, Sebastian (2020): „Rationality and Responsibility“. In: *Australasian Philosophical Review* 4.4, S. 379–385.

Schmitt, Arbogast (2011): *Denken und Sein bei Platon und Descartes: Kritische Anmerkungen zur ‚Überwindung‘ der antiken Seinsphilosophie durch die moderne Philosophie des Subjekts*. Heidelberg: Winter.

Schmitter, Amy (2021): „17th and 18th Century Theories of Emotions“. In: *The Stanford Encyclopedia of Philosophy*. Hrsg. von Edward Zalta. URL: https://plato.stanford.edu/entries/emotions-17th18th/. Aufgerufen am 22.01.2022.

Schmitz, Friederike (2017a): „Tierethik – eine Einführung“. In: *Tierethik. Grundlagentexte*. Hrsg. von Friederike Schmitz. Frankfurt a. M.: Suhrkamp, S. 13–73.

— Hrsg. (2017b): *Tierethik. Grundlagentexte*. 3. Aufl. Frankfurt a. M.: Suhrkamp.

Scholtes, Fabian (2007): „Zur Einleitung. Kultur als Herausforderung an Ökonomie und Wirtschaft“. In: *Kultur – Ökonomie – Ethik*. München: Hampp, S. 9–27.

Schulenburg, Mathias (2007): „Keine Spielereien. Vor 225 starb der französische Erfinder Jacques de Vaucanson“. In: *Deutschlandfunk*. 21.11.2007. URL: https://www.deutschlandfunk.de/keine-spielereien.871.de.html?dram:article_id=126066.

Schönecker, Dieter (2005): *Kants Begriff transzendentaler und praktischer Freiheit*. Kantstudien Ergänzungshefte 149. Berlin; New York: de Gruyter.

Searle, John (1969): *Speech Acts. An Essay in the Philosophy of Language*. Cambridge: Cambridge University Press.

— (1980): „Minds, Brains and Programs“. In: *Behavioral and Brain Sciences* 3, S. 417–457.

— (1994): „Geist, Gehirn, Programm“. In: *Künstliche Intelligenz. Philosophische Probleme*. Hrsg. von Walter Ch. Zimmerli und Stefan Wolf. Stuttgart: Reclam, S. 232–265.

Selinger, Evan (2021): „The line between human and machine begins to blur“. In: *The Boston Globe*. 06.10.2021. URL: https://www.bostonglobe.com/2021/10/06/opinion/line-between-human-machine-begins-blur/. Aufgerufen am 22.01.2022.

Seng, Leonie (2017): „Autonomes Fahren – Eine Frage der Ethik? Oder: Kant fährt Dein Auto gegen die Wand“. In: *Scilogs – Wissenschaftsblogs*. URL: https://scilogs.spektrum.de/feuerwerk-der-neuronen/autonomes-fahren-eine-frage-der-ethik-oder-kant-faehrt-dein-auto-gegen-die-wand/. Aufgerufen am 22.01.2022.

— (2019a): „Maschinenethik und Künstliche Intelligenz“. In: *Handbuch Maschinenethik*. Hrsg. von Oliver Bendel. Wiesbaden: VS Verlag für Sozialwissenschaf-

ten, S. 1–21. URL: https://link.springer.com/referenceworkentry/10.1007/978-3-658-17484-2_13-1. Aufgerufen am 22.01.2022.

Seng, Leonie (2019b): „Mein Haus, mein Auto, mein Roboter? Eine (medien-) ethische Beurteilung der Angst vor Robotern und *künstlicher Intelligenz*“. In: *Maschinenethik. Normative Grenzen autonomer Systeme*. Hrsg. von Matthias Rath, Friedrich Krotz und Matthias Karmasin. Wiesbaden: VS Verlag für Sozialwissenschaften, S. 57–72.

— (2020): „Current challenges in Ethics of AI or: old wine in new bottles“. In: *Artificial Intelligence. Reflections in Philosophy, Theology, and the Social Sciences*. Hrsg. von Benedikt Paul Göcke und Astrid Rosenthal-von der Pütten. Leiden; Boston: Brill/ Mentis, S. 159–172.

Simon, Judith (2016): „Values in Design“. In: *Handbuch Medien- und Informationsethik*. Hrsg. von Jessica Heesen. Stuttgart: J. B. Metzler, S. 357–364.

Simonite, Tom (2020): „Did a Person Write This Headline, or a Machine? GPT-3, a new text-generating program from OpenAI, shows how far the field has come – and how far it has to go“. In: *Wired*. 22.07.2020. URL: https://www.wired.com/story/ai-text-generator-gpt-3-learning-language-fitfully/. Aufgerufen am 22.01.2022.

Singer, Peter (2009): *Animal Liberation: The Definitive Classic of the Animal*. Modern Classics. New York: Harper Perennial.

Soares, Nate und Fallenstein, Benya (2017): „Agent Foundations for Aligning Machine Intelligence with Human Interests: A Technical Research Agenda“. In: *The Technological Singularity: Managing the Journey*. Hrsg. von Victor Callaghan et al. Berlin; Heidelberg: Springer, S. 1–13.

Sorabji, Richard (2009): „Body and Soul in Aristotle“. In: *Philosophy* 49.187, S. 63–89.

Spitzer, Manfred (2012): *Digitale Demenz: Wie wir uns und unsere Kinder um den Verstand bringen*. München: Droemer HC.

Springer Nature Group (2019): *Springer Nature publishes its first machine-generated book*. London; Heidelberg. URL: https://group.springernature.com/de/group/media/press-releases/springer-nature-machine-generated-book/16590134. Aufgerufen am 22.01.2022.

Steiner, Anna (2016): „Zum Nazi und Sexisten in 24 Stunden“. In: *Frankfurter Allgemeine Zeitung*. Aktualisiert am 24.03.2016. URL: https://www.faz.net/aktuell/wirtschaft/netzwirtschaft/microsofts-bot-tay-wird-durch-nutzer-zum-nazi-und-sexist-14144019.html. Aufgerufen am 22.01.2022.

Steinfath, Holmer (2021): „Gutes Leben“. In: *Handbuch Technikethik*. Hrsg. von Armin Grunwald und Rafaela Hillerbrand. 2., aktualisierte und erweiterte Aufl. Stuttgart: J. B. Metzler, S. 191–195.

Stiller, Sebastian (2015): *Planet der Algorithmen*. München: Knaus Verlag.

Tesla Inc., Hrsg. (2021): *Artificial Intelligence & Autopilot*. URL: https://www.tesla.com/en_EU/AI. Aufgerufen am 22.01.2022.

Tessloff Verlag (2012): *Die Frage der Woche: Was ist Moral?* In der Reihe *WAS IST WAS*, 20.01.2012. URL: https://www.wasistwas.de/archiv-geschichte-details/die-frage-der-woche-was-ist-moral.html. Aufgerufen am 01.11.2021.

Tetens, Holm (2006): *Philosophisches Argumentieren*. München: C. H. Beck.

Thimm, Caja (2019): „Die Maschine – Materialität, Metapher, Mythos: Ethische Perspektiven auf das Verhältnis zwischen Mensch und Maschine". In: *Die Maschine: Freund oder Feind? Mensch und Technologie im digitalen Zeitalter*. Hrsg. von Caja Thimm und Thomas Bächle. Wiesbaden: Springer VS, S. 17–39.

Thimm, Caja und Bächle, Thomas (2019): „Autonomie der Technologie und autonome Systeme als ethische Herausforderung". In: *Maschinenethik. Normative Grenzen autonomer Systeme*. Hrsg. von Matthias Rath, Friedrich Krotz und Matthias Karmasin. Berlin: VS Verlag für Sozialwissenschaften, S. 73–87.

Thomass, Barbara et al. Hrsg. (2022): *Ethik der öffentlichen Kommunikation. Eine kommunikationswissenschaftliche Einführung*. Heidelberg; New York: Springer.

Tonry, Michael, Hrsg. (2011): *The Oxford Handbook of Crime and Criminal Justice*. Oxford: Oxford University Press.

Toyka-Seid, Christiane (2021): *HanisauLand. Lexikon: Moral*. Hrsg. von der *Bundeszentrale für politische Bildung*. URL: https://www.hanisauland.de/wissen/lexikon/grosses-lexikon/m/moral.html. Aufgerufen am 22.01.2022.

Tugendhat, Ernst und Wolf, Ursula (1986): *Logisch-semantische Propädeutik*. Stuttgart: Reclam.

Turing, Alan (1950): „Computing Machinery and Intelligence". In: *Mind* 49, S. 433–460.

Twitter, @jackyalcine (2015): *Google Photos, y'all fucked up. My friend's not a gorilla.* URL: https://twitter.com/jackyalcine/status/615329515909156865. Aufgerufen am 05.11.2021.

Ullrich, Stefan (2014): „Informationelle Mü(n)digkeit. Über die unbequeme Selbstbestimmung." In: *Datenschutz und Datensicherheit* 10, S. 696–700.

— (2019a): „Algorithmen, Daten und Ethik. Ein Beitrag zur Papiermaschinenethik". In: *Handbuch Maschinenethik*. Hrsg. von Oliver Bendel. Wiesbaden: VS Verlag für Sozialwissenschaften, S. 1–26.

— (2019b): „Moralische Maschinen. Was die Maschine über die Moral ihrer Schöpferinnen und Schöpfer verrät". In: *Maschinenethik. Normative Grenzen autonomer Systeme*. Hrsg. von Matthias Rath, Friedrich Krotz und Matthias Karmasin. Wiesbaden: VS Verlag für Sozialwissenschaften, S. 243–261.

van de Poel, Ibo (2021a): „Values and Design“. In: *The Routledge Handbook of the Philosophy of Engineering*. Hrsg. von Diane Michelfelder und Neelke Doorn. New York: Taylor & Francis, S. 300–314.
— (2021b): „Werthaltigkeit der Technik“. In: *Handbuch Technikethik*. Hrsg. von Armin Grunwald und Rafaela Hillerbrand. 2., aktualisierte und erweiterte Aufl. Stuttgart: J. B. Metzler, S. 132–136.
Van den Hoven, Jeroen und Lockhorst, Gert-Jan (2002): „Deontic Logic and Computer-Supported Computer Ethics“. In: *Cyberphilosophy: The Intersection of Computing and Philosophy*. Hoboken: Blackwell, S. 280–289.
van Huijstee, Mariëtte et al. (2021): *Tackling deepfakes in European policy*. EPRS – European Parliament Research Service. url: https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf. Aufgerufen am 22.01.2022.
van Rysewyk, Simon und Pontier, Matthijs, Hrsg. (2015a): *Machine Medical Ethics*. Cham: Springer.
van Rysewyk, Simon Peter und Pontier, Matthijs (2015b): „A Hybrid Bottom-Up and Top-Down Approach to Machine Medical Ethics: Theory and Data“. In: *Machine Medical Ethics*. Hrsg. von Simon Peter van Rysewyk und Matthijs Pontier. Cham: Springer, S. 93–110.
Venkatesh, Viswanath et al. (2003): „User Acceptance of Information Technology: Toward a Unified View“. In: *MIS Quaterly* 27.3, S. 425–478. Management Information Systems Research Center, University of Minnesota.
Verbeek, Peter-Paul (2011): *Moralizing Technology. Understanding and Designing the Morality of Things*. Chigaco: The University of Chicago Press.
Versenyi, Laszlo (1974): „Can Robots be Moral?“ In: *Ethics* 84, S. 248–259.
Vincent, James (2018): „Google ‚fixed‘ its racist algorithm by removing gorillas from its image-labeling tech“. In: *The Verge*. 12.01.2018. url: https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai. Aufgerufen am 22.01.2022.
Voss, Laura (2021): *More Than Machines? The Attribution of (in)Animacy to Robot Technology*. Bielefeld: Transcript.
Voss, Laura et al. (2018): *Not my robots*. url: https://notmyrobot.home.blog. Aufgerufen am 22.01.2022.
Vígh, Tamás (2016): „People want their robots to be expressive and communicative“. In: *envienta. open source everything*. 06.09.2016. url: https://envienta.net/hardware-robotics-ai/people-want-their-robots-to-be-expressive-and-communicative/. Aufgerufen am 22.01.2022.
Wagner, Wolf-Rüdiger (2014): *Ein Blick auf Platons „Schriftkritik“ – eine Anleitung zum angemessenen Umgang mit schriftlichen Texten*. Hildesheim: Niedersächsiches Landesinsitut für schulische Qualitätsentwicklung. url: https:

/ / www . nibis . de / uploads / 2medfach / files / Ein _ Blick _ auf _ Platons _ Schriftkritik.pdf. Aufgerufen am 01.03.2022.

Wallach, Wendell und Allen, Colin (2009): *Moral Machines. Teaching Robots Right from Wrong*. Oxford: Oxford University Press.

Wallach, Wendell und Asaro, Peter (2016): *Machine Ethics and Robot Ethics*. London: Routledge.

Wallach, Wendell, Franklin, Stan und Allen, Colin (2010): „A conceptual and computational model of moral decision making in human and artificial agents". In: *TopiCS* 2.3, S. 454–485.

Waytz, Adam, Heafner, Joy und Epley, Nicholas (2014): „The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle". In: *Journal of Experimental Social Psychology* 52, S. 113–117.

Weber, Marc Andree und Yolcu, Nadja-Mira, Hrsg. (2019): *Edmund L. Gettier: Is Justified True Belief Knowledge? / Ist gerechtfertigte, wahre Überzeugung Wissen?* Englisch / Deutsch. Great Papers Philosophy. Stuttgart: Reclam.

Wendland, Karsten (2022): „Demystifying Artificial Consciousness – About Attributions, Black Swans, and Suffering Machines". In: *Journal of AI Humanities* 9, S. 137–166.

Werle, Raymund (2021): „Technik als soziale Konstruktion". In: *Handbuch Technikethik*. Hrsg. von Armin Grunwald und Rafaela Hillerbrand. 2., aktualisierte und erweiterte Aufl. Stuttgart: J. B. Metzler, S. 128–131.

Werner, Micha und Düwell, Marcus (2021): „Deontologische Ethik". In: *Handbuch Technikethik*. Hrsg. von Armin Grunwald und Rafaela Hillerbrand. 2., aktualisierte und erweiterte Aufl. Stuttgart: J. B. Metzler, S. 171–175.

Wikipedia (2021): *Moral*. 93 Wikipedia-Autor∗innen: https://xtools.wmflabs.org/articleinfo-authorship/de.wikipedia.org/Moral?uselang=de, 21.02.2022. url: https://de.wikipedia.org/wiki/Moral. Aufgerufen am 01.03.2022.

Williamson, Jon (2009): „The Philosophy of Science and its relation to Machine Learning". In: *Scientific Data Mining and Knowledge Discovery. Principles and Foundations*. Hrsg. von Mohamed Medhat Gaber. Berlin; Heidelberg: Springer, S. 77–89.

Witte, Hannah (2021): *Typohacks. Handbuch für gendersensible Sprache und Typografie*. Frankfurt a. M.: form.

Wittgenstein, Ludwig (2006): *Tractatus logico-philosophicus*. 1. Aufl. 1984. Werkausgabe Band I. Für die vorliegende Ausgabe wurde der Text neu durchgesehen von Joachim Schulte. Frankfurt a. M.: Suhrkamp.

Wolf, Ursula (2013): „Einleitung". In: *Aristoteles. Nikomachische Ethik*. 4. Aufl. Reinbek bei Hamburg: Rowohlts Enzyklopädie, S. 9–22.

Wölm, Erik (2019): „Warum mein Auto nie allein schuld sein wird. Über die Teilverantwortlichkeit autonomer Akteure". In: *Maschinenethik. Normative Gren-*

zen autonomer Systeme. Hrsg. von Matthias Rath, Friedrich Krotz und Matthias Karmasin. Wiesbaden: VS Verlag für Sozialwissenschaften, S. 173–191.

Xanke, Lisa und Bärenz, Elisabeth (2012): „Künstliche Intelligenz in Literatur und Film – Fiktion oder Realität?“ In: *Journal of New Frontiers in Spatial Concepts* 4, S. 36–43. url: https://publikationen.bibliothek.kit.edu/1000027215/2055128. Aufgerufen am 22.01.2022.

Zeng, Daniel (2015): „AI Ethics. Science Fiction Meets Technological Reality“. In: *IEEE Intelligent Systems* 30, S. 2–5.

Stichwort- und Personenverzeichnis

A Anhang

Abbildungen

Seite 31:

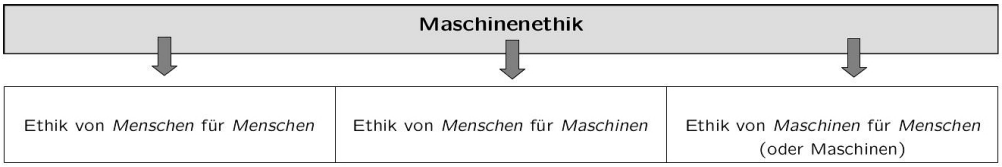

Abbildung 1: Drei mögliche Lesarten des Begriffs *Maschinenethik*.

Seite 60:

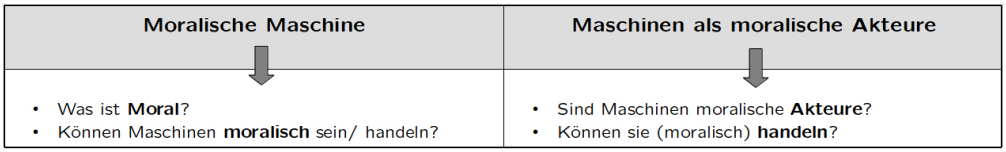

Abbildung 2: Moralische Maschinen und moralische Akteure.

Seite 65:

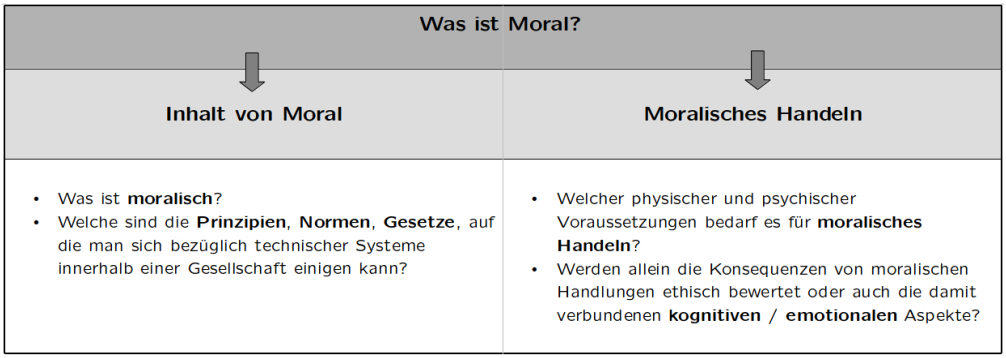

Abbildung 3: Inhalt von Moral und die Voraussetzungen moralischen Handelns.