

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

**„Die Utopie der Verantwortungszuschreibung in der Mensch-Maschine-
Interaktion“**

verfasst von / submitted by

Theresia Liebetreu, BEd

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of
Master of Education (MEd)

Wien, 2023 / Vienna, 2023

Studienkennzahl lt. Studienblatt /
degree programme code as it appears on
the student record sheet

UA 199 506 525 02

Studienrichtung lt. Studienblatt/
degree programme as it appears on
the student record sheet

Masterstudium Lehramt Sek (AB)
Unterrichtsfach Deutsch Unterrichtsfach
Psychologie und Philosophie

Betreut von / Supervisor:

Mag. Dr. Katharina Lacina

Abstract

In Anbetracht der rasanten Weiterentwicklung von künstlicher Intelligenz, artifiziellen Systemen und Robotern rückt die Frage der Verantwortungszuschreibung insbesondere im Kontext der Mensch-Maschine-Interaktion immer mehr in den Fokus gesellschaftlichen Interesses. Trotz des Bestehens diverser Konzepte, die die Zuschreibung von Verantwortung im zwischenmenschlichen Bereich regeln und strukturieren, ergibt sich in der Anwendung ebendieser Konzepte auf nicht-menschliche Instanzen eine Problematik, die dringend zu bearbeiten ist.

Dieser Arbeit liegt die These zugrunde, die Verantwortungszuschreibung in der Mensch-Maschine-Interaktion sei grundsätzlich utopisch. Um diese These zu evaluieren, werden unterschiedliche Verantwortungskonzepte einander gegenübergestellt und auf die Möglichkeit ihrer Anwendung im Bereich der Mensch-Maschine-Interaktion geprüft. Auch die Verortung artifizieller Systeme als moralische Handlungssubjekte oder moralische Handlungsobjekte wird in diesem Kontext berücksichtigt.

Die philosophische Auseinandersetzung zeigt mittels einer kritischen Reflexion theoretischer Konzepte sowie anhand der Betrachtung einer Reihe an Beispielen aus unterschiedlichen Bereichen der Roboterethik, dass in Hinblick auf Maschinen und Roboter eine Zuschreibung von Verantwortung tatsächlich utopisch ist. Dieses Erkenntnis kann einerseits im Philosophieunterricht der Sekundarstufe II genutzt werden, um Schüler*innen für die Probleme in der Mensch-Maschine-Interaktion empfänglich zu machen. Andererseits dient die im Rahmen dieser Arbeit detailliert dargelegte Utopie der Verantwortungszuschreibung als Basis für weitere wissenschaftliche Forschung und Entwicklung neuer Konzepte.

Inhaltsverzeichnis

Abkürzungsverzeichnis	1
1. Einleitung	2
2. Abgrenzungen	4
2.1. [Künstliche] Intelligenz	4
2.2. Ausprägungen künstlicher Intelligenz	6
2.2.1. AGI, ANI und ASI	7
2.2.2. Affective Computing	9
3. Maschinen- und Roboter[ethik]	12
3.1. Maschinen	13
3.2. Roboter	14
4. Verantwortung	17
4.1. Kontextualisierung	17
4.2. Terminologie und Abgrenzungen	18
4.3. Formen der Verantwortungszuschreibung	20
4.3.1. Verantwortung bei Otfried Höffe und Claudia Blöser	20
4.3.2. Verantwortung bei Claudia Blöser	23
4.3.3. Verantwortung in der Wissenschaft	24
4.4. Verantwortungsmodelle im wissenschaftlichen Kontext	25
4.4.1. Rollenverantwortung – Interne Verantwortung	25
4.4.2. Kausalhandlungsverantwortung – Externe Verantwortung	26
5. Problematisierung	28
5.1. Bedingungen der Verantwortungszuschreibung	28
5.2. Das Problem der Verantwortungszuschreibung in der Maschinen- und Technikethik	29
5.3. Herstellungsorientierte vs. gebrauchorientierte Verantwortung	32
6. Arbeitsfelder der Roboterethik	33
6.1. Roboter als moralische Handlungssubjekte	33
6.1.1. Definition moralischer Akteursschaft (moral agency)	33
6.1.2. Bedingungen für moralische Handlungssubjekte	37
6.2. Konzepte moralischer Akteursschaft	41
6.2.1. J.P. Sullins vs. L. Floridi und J.W. Sanders	41
6.2.2. Moral agency nach C. Misselhorn	45
6.2.3. Moral agency nach J.H. Moor	46
6.3. Roboter als moralische Handlungsobjekte	48

7. Die Dichotomie der Verantwortungszuschreibung.....	51
7.1. Utopische Zuschreibungen in Therapie- Medizin und Pflegerobotik	51
7.1.1. DaVinci.....	52
7.1.2. Paro.....	55
7.1.3. Care-O-Bot und TUG	56
7.2. Utopische Zuschreibungen in der Militärrobotik	60
8. Fachdidaktische Erläuterungen	63
8.1. Kontextualisierung	63
8.2. Unterrichtsplanung	67
8.2.1. Stundenbild 1	69
8.2.2. Themenspezifischer Bezug.....	69
8.2.3. Stundenbild 2.....	71
8.2.4. Themenspezifischer Bezug.....	72
9. Fazit.....	76
10. Literaturverzeichnis.....	82

Abkürzungsverzeichnis

AI	Artificial Intelligence
AGI	Artificial General Intelligence
ANI	Artificial Narrow Intelligence
ASI	Artificial Superintelligence
FACS	Facial Action Coding System
GA	Gruppenarbeit
KI	Künstliche Intelligenz
LSG	Lehrer*innen-Schüler*innen-Gespräch
ML	Machine Learning/Maschinelles Lernen
MMI	Mensch-Maschine-Interaktion
PA	Partner*innenarbeit
POL	Problemorientiertes Lernen
SuS	Schüler* und Schülerinnen*

1. Einleitung

Unter anderem durch die rasante Weiterentwicklung artifizierter Systeme und künstlicher Intelligenz ergeben sich philosophisch im wissenschaftlichen wie auch im gesellschaftlichen Diskurs unterschiedlich breit gefächerte Frage- und Problemstellungen, die sich nicht pauschal beantworten oder behandeln lassen. So ist beispielsweise die Frage nach der Verantwortungszuschreibung in der Mensch-Maschine-Interaktion höchst komplex, zeichnet sie sich schließlich nicht nur durch eine multidimensionale Komponente in Bezug auf die Art jener Instanzen, die in die MMI involviert sind, sondern auch durch weitere Aspekte aus. Beispielsweise spielt die Fähigkeit eines mit künstlicher Intelligenz ausgestatteten Systems, autonom zu lernen oder lediglich die ihm implementierten Algorithmen zu nutzen, um Handlungen zu vollziehen, hierbei eine entscheidende Rolle.

Mehr noch stehen allerdings der Bereich und das Arbeitsfeld des jeweiligen Systems im Vordergrund. Die Unterscheidung von Robotern als moralischen Handlungssubjekten¹ oder moralischen Handlungsobjekten², die unter anderem von Janina Loh getroffen wird, korreliert eng mit der Frage nach jenem Maß an Verantwortung, welches einem artifizierten System zugeschrieben werden kann oder muss, um einen kritischen Diskurs in Bezug auf zukünftige Zusammenarbeit oder Zusammenwirkung menschlicher und artifizierter Handlungsträger zu gewährleisten. Wie diese Verantwortung philosophisch einzuordnen, zu klassifizieren und zu verteilen ist, und inwiefern die Frage der Verantwortungszuschreibung im Rahmen einer Interaktion zwischen Menschen und Maschinen utopisch ist, soll in der zu verfassenden Arbeit beleuchtet werden.

Diese Arbeit verfolgt hierbei kein empirisches Interesse. Aufbau und Struktur sind auf eine philosophisch-theoretische Bearbeitung der vorliegenden Thematik ausgerichtet. Unter anderem geschieht dies durch zentrale Begriffsarbeit sowie den Vergleich bestehender philosophischer und ethischer Konzepte, die ihrerseits auf die gestellten Forschungsfragen übertragen werden. Hierfür wird Literatur aus unterschiedlichen Anwendungsbereichen herangezogen. Auch mittels einer fachdidaktischen Auseinandersetzung soll der breit gefächerte Themenkomplex aufbereitet werden. Hierdurch sollen Schüler*innen der Sekundarstufe II durch eine Kombination aus Kompetenzorientierung und problemorientiertem Lernen an die kritische Auseinandersetzung mit artifizierten Systemen herangeführt werden, um ein grundlegendes Verständnis für die Utopie der

¹ Loh, Janina: Roboterethik. Eine Einführung. Stuttgart: Suhrkamp 2019, S. 48.

² Ebd. S. 72.

Verantwortungszuschreibung in der MMI zu erlangen und jene Problemfelder zu begreifen, die sich aus einer philosophisch-ethischen Perspektive ergeben. Um zu verdeutlichen, wie komplex diese Perspektive ist und wie viele Subkategorien sie umfasst, werden unterschiedliche Aspekte der MMI, der Roboterphilosophie, Roboter- und Maschinenethik sowie der Robotik und Technikethik beleuchtet, um die bestehende Problematik sowohl wissenschaftlich als auch pädagogisch aufgreifen und bearbeiten zu können.

Der vorliegenden Masterarbeit liegt die These zugrunde, die Verantwortungszuschreibung in der Mensch-Maschine-Interaktion sei grundsätzlich utopisch. Da bislang die Frage nach einer scharf abgegrenzten Definition des Verantwortungsbegriffs in Bezug auf artifizielle Systeme nicht vollständig geklärt ist, wird ein Teil dieser Arbeit genau diese Problematik erläutern und insbesondere die Unterschiede zwischen moralischen Handlungssubjekten und moralischen Handlungsobjekten herausarbeiten. Auch alternative Ansätze wie jener von J.H. Moor, die einen Kompromiss zwischen *moral agency* und *moral patiency* zu finden suchen, werden in diesem Kontext beleuchtet. Der vorliegenden Arbeit liegen vier zentrale Forschungsfragen zugrunde, die durch eine stetige Erweiterung, Verknüpfung und Beantwortung eine detaillierte Reflexion der Master-These gewährleisten sollen:

- Wie kann Verantwortung definiert werden?
- Welche Kriterien für die Verantwortungszuschreibung gibt es?
- Inwiefern ist der Diskurs um die Verantwortungszuschreibung in der MMI sowie in Bezug auf autonome artifizielle Systeme utopisch?
- Wie können Schüler*innen der Sekundarstufe II philosophisch und problemorientiert an die vorliegende Thematik herangeführt werden?

Im Rahmen einer Auseinandersetzung mit den obig dargelegten Fragestellungen werden zunächst jene für die Arbeit zentralen Begriffe beleuchtet und definiert. Anschließend wird auf die Arbeitsfelder der Robotik und ausgewählte Bereiche eingegangen, um eine solide Basis für die Problematisierung und philosophisch-ethische Auseinandersetzung mit der postulierten Utopie der Verantwortungszuschreibung in der MMI zu bilden. Insbesondere die Frage nach der fachdidaktischen Umsetzung wird am Ende der Arbeit unter Berücksichtigung der bisher erarbeiteten Inhalte beantwortet. Die schüler*innengerechte Aufbereitung für den schulischen Kontext erhebt keinen Anspruch auf Vollständigkeit, sondern dient in erster Linie dazu, junge Menschen durch eine gezielte Heranführung an die Thematik empfänglich für zentrale philosophische und ethische Dilemmata zu machen, die sich durch die rapide Weiterentwicklung künstlicher Intelligenz und artifizieller Systeme

ergeben. Diese Masterarbeit soll hiermit letztlich nicht nur einen Überblick über den Themenkomplex „Verantwortung“ in Bezug auf artifizielle Systeme und künstliche Intelligenz geben und die abgebildeten Forschungsfragen kritisch beleuchten, sondern auch die bisher bestehende Lücke in der Frage nach der Verantwortungszuschreibung in der Mensch-Maschinen-Interaktion detailliert darlegen. Lösungsansätze auszuarbeiten ist hierbei genauso wenig Anspruch dieser Arbeit wie eine Implementierung in den wissenschaftlichen Kanon. Vielmehr wird das breite Spektrum an Themenfeldern kategorisch und fachdidaktisch aufbereitet, um die bestehende Problematik der Verantwortungsdiffusion in den Fokus zu rücken und herauszuarbeiten, inwiefern bei der Zuschreibung von Verantwortung in der MMI eine Utopie vorliegt, die sowohl wissenschaftlich als auch gesellschaftlich und fachdidaktisch einer genaueren Auseinandersetzung bedarf.

2. Abgrenzungen

2.1. [Künstliche] Intelligenz

Umgangssprachlich werden die Termini „Künstliche Intelligenz“ bzw. „Artifizielle Intelligenz“, häufig mit Begriffen wie „Roboter“, „Maschine“ oder „Artifizielle(s) System(e)“ gleichgesetzt. Wenngleich diese synonyme Verwendung zum Zwecke der Vermeidung von Wortwiederholungen in schriftlichen Texten taktisch nachvollziehbar ist, ist es philosophisch und ethisch relevant, eine Unterscheidung zu treffen, insbesondere, wenn über rechtliche Rahmenbedingungen und Verantwortungszuschreibung in Bezug auf eine oder mehrere dieser Kategorien reflektiert wird. Zunächst sei in diesem Kontext der Begriff der künstlichen Intelligenz näher beleuchtet. Da dieser ein Kompositum aus dem Adjektiv „künstlich“ sowie dem Nomen „Intelligenz“ ist, muss zunächst auf die Definition von „Intelligenz“ als eigenständiger Begrifflichkeit eingegangen werden.

“Intelligence”, die von der American Psychological Association als „[...] the ability to derive information, learn from experience, adapt to the environment, understand, and correctly utilize thought and reason [...]”³ definiert wird, weist darauf hin, dass die Möglichkeit der Nutzung der Ratio im Zuge eines Denkprozesses maßgeblich für die Zuschreibung von Intelligenz ist. Interessant ist, dass das *APA Dictionary of Psychology* den Terminus nicht lediglich auf lebendige Individuen – insbesondere Menschen – beschränkt,

³ American Psychological Association. (o. D.) Intelligence. In: APA Dictionary of Psychology. <https://dictionary.apa.org/intelligence> (24.2.2023)

sondern den Aspekt der Vernunft und des Denkens ungeachtet der äußeren Rahmenbedingungen in den Vordergrund stellt. Anders verhält es sich, wenn die Definition des *Duden Online* herangezogen wird. Dieser bezeichnet Intelligenz als die „Fähigkeit [des Menschen], abstrakt und vernünftig zu denken und daraus zweckvolles Handeln abzuleiten“⁴ Der eingeklammerte Hinweis auf die Kontextualisierung von Intelligenz als etwas dem Menschen Zugehöriges kann hierbei unterschiedlich interpretiert werden.

Einerseits besteht die Möglichkeit, den Terminus ohne die eckige Klammer zu lesen und somit die Anwendung des Begriffs in einem breiteren Kontext zu gewährleisten. Hierdurch wäre es beispielsweise möglich, das Kompositum „Künstliche Intelligenz“ nicht als semantischen Widerspruch zu verstehen. Andererseits wiederum impliziert der Hinweis des Duden, Intelligenz stehe mit abstraktem Denken in Verbindung, dass es sich bei einem*iner Intelligenzträger*in um einen Menschen handeln müsse. Gerechtfertigt werden kann diese Interpretation unter anderem dadurch, dass zum derzeitigen Stand der Entwicklung künstlicher Intelligenz zwar durchaus gewünschte Effekte und Antworten von einer KI generiert werden können, diese allerdings häufig fehlerhaft sind, weil gerade abstrakte Zusammenhänge nicht oder nicht ausreichend erfasst und umgesetzt werden können.⁵

Ein Blick in die Literatur, der über die eindimensionale Definition des Intelligenzbegriffs hinausgeht, legt die Komplexität des Begriffs nahe. Die psychologische Differenzierung von Intelligenz in spezifische Teilbereiche nach Gardner basiert beispielsweise auf einer Unterscheidung von visuell-räumlicher, körperlich-kinästhetischer, musikalischer, interpersonaler, intrapersonaler, sprachlicher sowie logisch-mathematischer Intelligenz⁶, während andere Wissenschaftler*innen den Aspekt der emotionalen Intelligenz hinzufügen und dessen Relevanz betonen. Neumann-Endres beispielsweise versteht emotionale Intelligenz als Fähigkeit, eigene Gefühle sowie die von anderen Individuen wahrnehmen, beeinflussen und begreifen zu können.⁷

⁴ Dudenredaktion. (o. D.). Intelligenz. In: Duden online. <https://www.duden.de/rechtschreibung/Intelligenz> (24.2.2023)

⁵ Anm.: Auf diese Problematik wird in späteren Kapiteln dieser Arbeit vertieft eingegangen. Die Randbemerkung im Rahmen dieses Kapitels dient lediglich der einleitenden Auseinandersetzung mit terminologischen Abgrenzungen und deren semantischen Besonderheiten.

⁶ Vgl. Gardner, Howard: *Frames of Mind – The Theory of Multiple Intelligences*. New York: Basic Books 1983 [Dt.: *Abschied vom IQ – Die Rahmentheorie der vielfachen Intelligenzen*. Stuttgart: Klett-Cotta 1991], zitiert nach Misselhorn, Catrin: *Grundfragen der Maschinenethik*. Ditzingen, Stuttgart: Reclam⁴ 2019, S. 17.

⁷ Neumann-Endres, Meike-Lesley: *Die Wirkung der Emotionalen Intelligenz auf die interkulturelle Sensibilität – ein wissenschaftliches Essay*. In: Valeva, Milena (Hrsg.): *Emotionale Intelligenz*. München: Rainer Hampp Verlag 2019, S. 16.

In Hinblick auf die Suche nach einer Definition künstlicher Intelligenz kann Folgendes vorweggenommen werden: Während Intelligenz ein multidimensional zu interpretierender Terminus ist, der in unterschiedlichen Disziplinen und Kontexten differenziert betrachtet wird und der nicht einheitlich definiert wird⁸, verhält sich dies bei künstlicher Intelligenz anders. Als Kompositum ist diese zwar vermeintlich selbsterklärend, semantisch allerdings widersprüchlich, da anhand der bereits erläuterten Aspekte fraglich ist, ob Intelligenz – sofern sie „künstlich“ generiert ist, überhaupt als solche im ursprünglichen Sinne bezeichnet werden kann.

2.2. Ausprägungen künstlicher Intelligenz

Problematisch bei dem Versuch, künstliche Intelligenz zu definieren, ist nicht nur, dass trotz mehreren bestehenden Definitionen offenbar kein wissenschaftlicher Konsens darüber besteht, welche Begriffserläuterung tatsächlich dasjenige beschreibt, was sie bezeichnen soll.⁹ Auch die Frage danach, was dieses vermeintlich zu Bezeichnende ist, scheint in Anbetracht der folgenden vielfältigen Auffassungen fragwürdig. Die *Hochrangige Expertengruppe für künstliche Intelligenz* definiert KI beispielsweise folgendermaßen:

Artificial intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals AI-based systems can be purely software-based, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g. advanced robots, autonomous cars, drones or Internet of Things applications).¹⁰

Hier wird interessanterweise keine Unterscheidung zwischen unterschiedlichen Formen oder Ausprägungen artifizieller Intelligenz getroffen. Vielmehr wird alles, was gemeinhin als artifizielle(s) System(e) bezeichnet werden könnte, unter dem Hyperonym „Künstliche Intelligenz“ subsummiert. Diese Definition erscheint einerseits geeignet, wenn der Terminus als Überbegriff für eine Vielzahl an komplexen Inhalten gelten soll. Genau dies verfügt im wissenschaftlichen Kontext allerdings auch über das Potenzial, höchst dringlich zu differenzierende Konzepte artifizieller Intelligenz zu vereinheitlichen, während es insbesondere im Kontext der Verantwortungszuschreibung unbedingt notwendig wäre, genau zwischen ebendiesen zu differenzieren. Betrachtet man weitere Definitionsvorschläge, ergibt sich rasch der Eindruck, künstliche Intelligenz sei eine Art

⁸ Vgl. Misselhorn (2019), S. 17.

⁹ Vgl. Dobler, Ralph-Niklas: Mensch und Maschine. In: Brandstetter, Nicole und Ralph-Niklas Dobler u.a. (Hrsg.): Künstliche Intelligenz Interdisziplinär. München: UVK Verlag 2020, S. 27.

¹⁰ Europäische Kommission: High-Level Expert Group on Artificial Intelligence. A definition of AI: Main capabilities and scientific disciplines. Brüssel 2018, S. 1.
https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_december_1.pdf (23.7.2023)

Stellvertreter*in für etwas, dessen tatsächlicher Bedeutungsinhalt nicht im Entferntesten greifbar ist. So schreibt Beranek:

Unter KI versteht man eine Softwareanwendung, die nicht rein deterministische bzw. statische Lösungswege beschreitet, sondern auf Grund von Erfahrungen ihre Ausgangskonfiguration weiterentwickelt. Im Kern handelt es sich hierbei um angewandte Mathematik, genauer: Stochastik, die in großen Datenmengen Muster erkennt und hieraus Rückschlüsse zieht, welche sich dann im maschinell erzeugten, intelligenten Verhalten wiederfinden.¹¹

Dieser Versuch, KI zu definieren, beruht auf einer ähnlichen, wenn doch nicht identen Annahme. Neben der Tatsache, dass künstliche Intelligenz mit einer nicht näher erläuterten Softwareanwendung gleichgesetzt wird, wird auch der Aspekt der Erfahrung im Kontext mathematischer Datenverarbeitung erwähnt. Die Fähigkeit, Inhalte klassifizieren, verstehen und deuten zu können, scheint also beiden bislang abgebildeten Begriffserklärungen gemein zu sein. Unterschiedlich dürfte allerdings die Spannbreite des Terminus sein. Während die *Hochrangige Expertengruppe für künstliche Intelligenz* verdeutlicht, dass es sich bei künstlicher Intelligenz um Systeme handelt, die auf einer Software basieren, begreift Beranek KI selbst als Software. Dobler wiederum wendet sich gegen den gebräuchlichen Begriff „Künstliche Intelligenz“ und plädiert dafür, diesen durch die Termini „Kognitive Systeme“ bzw. „maschinelles Lernen“, im Englischen auch als „Machine Learning“ – kurz als ML bezeichnet - zu ersetzen.¹² Problematisch erscheint hier in erster Linie die Bezeichnung künstlicher Systeme als „kognitiv“, da hiermit einerseits die Existenz eines ausgeprägten [menschlichen oder dem Menschen ähnlichen] Denkvermögens impliziert wird und andererseits die damit verbundene uneingeschränkte Fähigkeit, komplexe Inhalte kognitiv zu verarbeiten. Aus diesem Grund wird in der vorliegenden Arbeit von dieser Bezeichnung abgesehen, während andere Definitionen und Subkategorien zu Anwendung kommen.

2.2.1. AGI, ANI und ASI

Manche Wissenschaftler*innen beschreiben im Sinne der genaueren Differenzierung Subkategorien künstlicher Intelligenz.¹³ Gausling beispielsweise unterscheidet schwache KI (ANI [Artificial Narrow Intelligence]), starke KI (AGI [Artificial General Intelligence])

¹¹ Beranek, Angelika: Auch eine Frage der Menschenrechte: Was ist Künstliche Intelligenz? In: Politikum. Frankfurt am Main: Wochenschau Verlag 2021, S. 7.

¹² Vgl. Dobler (2020), S. 27-28.

¹³ Vgl. u. a. Sukopp, Thomas: Grenzen des Einsatzes von Künstlicher Intelligenz. In: Neue Technologien – neue Kindheiten? Stuttgart: J.B. Metzler 2020. S. 197–225. und Misselhorn, Catrin: Künstliche Intelligenz und Empathie. Vom Leben mit Emotionserkennung, Sexrobotern & Co. Ditzingen: Reclam 2021. sowie Gausling, Tina: KI und DS-GVO im Spannungsverhältnis. In: Ballestrem, Johannes Graf und Ulrike Bär u.a. (Hrsg.): Künstliche Intelligenz. Wiesbaden: Springer Fachmedien Wiesbaden GmbH 2020.

sowie Superintelligenz (ASI [Artificial Superintelligence])).¹⁴ Während schwache künstliche Intelligenz (ANI) eine Hilfsfunktion bei der anwendungsorientierten Lösung bestimmter Probleme bietet, beziehungsweise bei der Beantwortung spezifischer Fragen Entscheidungen forciert, ist starke künstliche Intelligenz (AGI) bislang lediglich ein theoretisches Konzept, dessen Umsetzung sich in Entwicklung befindet.¹⁵ Ein artifizielles System, das mit starker KI ausgestattet ist, wäre nach Gausling, Gandhi und Ehl in der Lage, auf unterschiedliche Weise eine Vielzahl an komplexen Inhalten zu erfassen, diese zu vernetzen und entsprechend menschlichen Fähigkeiten oder gar darüber hinausgehend anwenden zu können. Dies sei allerdings bislang bei keinem artifiziellem System der Fall, wenngleich ein maßgeblicher Zuwachs an Autonomie und Komplexität in nicht allzu ferner Zukunft erwartet wird.¹⁶

ANI und AGI unterscheiden sich bislang offenbar in erster Linie durch das Ausmaß an multidimensionaler Vernetzungsfähigkeit, welche in ihrer ursprünglichen Beschaffenheit allerdings aktuell nur menschlichen Individuen zugestanden wird. Während schwache KI die Vielschichtigkeit des menschlichen Verstandes bloß nachahmt, ist starke KI in der Lage, ähnliche oder bereits überlegene „gedankliche“ Konzepte zu erschaffen. Durch Nick Bostroms Anmerkungen zu einer möglicherweise in der zweiten Hälfte des 21. Jahrhunderts aufkommenden Superintelligenz (ASI) wurde gegen Ende des 20. Jahrhunderts die Vorstellung einer AGI noch übertroffen: ASI bezeichnet demnach „[...] an intellect that is much smarter than the best human brains in practically every field, including scientific creativity, general wisdom and social skills“¹⁷.

Eine ASI dürfte demnach in ihrer angestrebten Autonomie genau dasjenige sein, wovor sich viele Menschen fürchten. In einem vermeintlich realistischen Szenario könnte sie menschlicher Existenz dermaßen überlegen sein, dass sie nicht nur in der Lage wäre, Menschen in sämtlichen Lebensbereichen zu übertreffen, sondern sich auch an das bisher für Menschen spezifische breite Gefühlsspektrum und damit verbundene echte Empathiefähigkeit heranzutasten. All jene Aspekte und Besonderheiten, die bislang Menschen von Robotern und Maschinen unterscheiden, würden hierdurch an Trennschärfe verlieren. Insbesondere in Bezug auf die Möglichkeit, KI bzw. ASI könnten die Emotionen des

¹⁴ Vgl. Gausling (2020), S. 12.

¹⁵ Vgl. Ebd. S. 13.

¹⁶ Vgl. Ebd. S. 13. sowie Gandhi Sharad und Christian Ehl: AI&U – Translating Artificial Intelligence into Business. SV 2017, S. 28.

¹⁷ Bostrom, Nick: How long before superintelligence? In: International Journal of Futures Studies 2 (1998). <https://nickbostrom.com/superintelligence> (18.7.2023)

menschlichen Gegenübers zukünftig so zielsicher deuten, dass eine Interaktion nicht durch die Tatsache überlagert wird, dass es sich bei dem*der Kommunikationspartner*in um ein maschinelles Gegenüber handelt, würde zwar einerseits Vorteile mit sich bringen, andererseits allerdings auch die Frage der Verantwortungszuschreibung deutlich zentraler in den Fokus rücken.

Eine ASI, die beispielsweise im Gespräch mit einem Menschen mit heiklen moralischen, medizinischen oder ethischen Entscheidungen konfrontiert wäre, trüge ein hohes Maß an Verantwortung. Entscheidungen, die ein menschliches Gegenüber entweder basierend auf Erfahrung oder gesteuert durch erlernte Verhaltensweisen treffen würde, würde eine ASI durch ihr den Menschen übertreffendes Maß an „general wisdom and social skills“¹⁸ vermutlich anders festlegen und interpretieren. Ob die Entscheidungen einer ASI allerdings alleinig durch die ihr zugeschriebene Überlegenheit in sämtlichen menschlichen Bereichen einen ähnlich hilfreichen oder erwünschten Effekt hätten wie die, die ein emotional mit dem*der jeweiligen Kommunikationspartner*in verbundener Mensch getroffen hätte, bleibt an dieser Stelle der vorliegenden Arbeit noch ungeklärt. Auch die Frage danach, ob- und inwiefern die Weiterentwicklung ANI-betriebener Systeme zu AGI-Systemen oder gar artifizieller Superintelligenz überhaupt realistisch ist und in welchen Anwendungsfeldern oder Bereichen der Robotik dies besonders in Zukunft problematisch sein könnte, wird im weiteren Verlauf dieser Arbeit geklärt.

2.2.2. Affective Computing

In diesem Kontext ist es unter anderem relevant, nicht nur artifizielle Intelligenz als solche in Konnotation mit psychologisch definierter menschlicher Intelligenz zu betrachten, sondern insbesondere der Unterkategorie der emotionalen künstlichen Intelligenz – dem *affective computing* - Beachtung zu schenken, da dies(e) unter anderem in Hinblick auf die erwartete Weiterentwicklung artifizieller Systeme und deren Einsatz in unterschiedlichen Bereichen und Arbeitsfeldern der Robotik von Bedeutung ist- und weiterhin sein wird.¹⁹ Außer Frage steht, dass ein Großteil menschlichen Verhaltens emotionalen Ursprungs ist.²⁰ Unter anderem aus diesem Grund bedarf ein jedes artifizielles System, das mit Menschen

¹⁸ Bostrom, Nick: How long before superintelligence? In: International Journal of Futures Studies 2 (1998). <https://nickbostrom.com/superintelligence> (18.7.2023)

¹⁹ Insbesondere für Kapitel 7.1. sind die hier erläuterten Inhalte relevant. Da sie allerdings terminologische Klärung verlangen, sind sie strukturell im einführenden Abschnitt der vorliegenden Masterarbeit verortet, um in späteren Teilen der Arbeit ohne Weiteres an bereits behandeltes Wissen anknüpfen zu können.

²⁰ Vgl. Misselhorn, Catrin: Künstliche Intelligenz und Empathie. Vom Leben mit Emotionserkennung, Sexrobotern & Co. Ditzingen: Reclam 2021, S. 12-13.

interagieren soll, grundlegender Kenntnisse über die Beschaffenheit und Auswirkungen von Emotionen.²¹ Vielmehr noch muss ein artifizielles System allerdings in der Lage sein, die Gefühle seines menschlichen Gegenübers auch zu lesen und korrekt zu interpretieren. Hierfür wurde bereits eine Reihe an Möglichkeiten evaluiert und entwickelt.²²

Die Rede ist hierbei unter anderem von gesichtsbasierter und stimmbasierter Emotionserkennung über Bio-Sensoren bis hin zu multimodalen Ansätzen, die mehrere Aspekte berücksichtigen.²³ Gesichtsbasierte Emotionserkennung funktioniert unter anderem mittels des FACS, dem *Facial Action Coding System*, welches mittels genauen Erfassens und Zuordnens kleinster Muskelbewegungen einen Code erstellt, der den vorliegenden Gesichtsausdruck weitestgehend klassifizierbar macht.²⁴ Schiedermeier merkt an, dass es in hohem Maße problematisch ist, dass Gesichtserkennung durch Maschinelles Lernen (ML) angeleitet wird und hierdurch rassistische und rassifizierende Bewertungen seitens der KI entstehen können.²⁵ So unterlief beispielsweise der Gesichtserkennungs-Software von Google 2015 ein großes Missgeschick, als die Software Schwarze Personen²⁶ nicht als menschliche Individuen, sondern als Gorillas einordnete.²⁷

Wird danach gefragt, wer hierfür verantwortlich ist, zeigt sich bereits ein erstes Dilemma, schließlich handelt es sich um eine KI-betriebene Software, die basierend auf ihr implementierten Algorithmen und vermeintlich repräsentativen Daten rassistisches Verhalten zeigte. Jedes artifizielle System, jede KI-Software und jede Maschine wird allerdings von Menschen trainiert und programmiert, weshalb die Frage gestellt werden muss, ob Entwickler*innen der Software für derartige Fehler zur Verantwortung gezogen werden müssen, oder ob der Software als solcher Verantwortung zugeschrieben werden

²¹ Vgl. Misselhorn (2021), S. 12-13.

²² Vgl. Ebd. S. 20.

²³ Vgl. Ebd. S. 20.

²⁴ Vgl. Ebd. S. 20.

²⁵ Vgl. Objective or biased. On the questionable use of Artificial Intelligence for job applications. <https://web.br.de/interaktiv/ki-bewerbung/en/> (18.7.2023) sowie Schiedermeier, Gudrun: Diskriminierende Systeme – Rassismus und Frauenfeindlichkeit in KI-Systemen. In: Brandstetter, Nicole und Ralph-Miklas Dobler u.a. (Hrsg.): Mensch und Künstliche Intelligenz. Herausforderungen für Kultur, Wirtschaft und Gesellschaft. München: UVK Verlag 2021, S. 14.

²⁶ Anmerkung: In der folgenden Arbeit wird zum Zweck einer contra-rassistischen Repräsentation „Schwarz“ großgeschrieben, während „weiß“ kursiv und kleingeschrieben wird. Zur Bedeutung vgl. u.a. Antje Lann Hornscheidt, Adibeli Nduka-Agwu: Der Zusammenhang zwischen Rassismus und Sprache. In: Adibeli Nduka-Agwu u. Antje Lann Hornscheidt (Hrsg.): Rassismus auf gut Deutsch. Ein kritisches Nachschlagewerk zu rassistischen Sprachhandlungen. Frankfurt am Main: Brandes & Apsel 2013, S. 11-33.

²⁷ Vgl. GoogleWatchBlog: Peinlicher Fehler: Google Photos erkennt/erkannte dunkelhäutige Personen als Gorillas. <https://www.googlewatchblog.de/2015/06/peinlicher-fehler-google-photos/> (23.7.2023)

kann. Das obige Beispiel zeigt also bereits sehr deutlich, dass Verantwortungszuschreibung im Bereich der künstlichen Intelligenz und der MMI problematisch ist.²⁸

Stimmbasierte Emotionserkennung geht bei der Analyse ähnlich vor. Hier werden allerdings keine visuellen Aspekte ausgewertet. Vielmehr dienen akustische Besonderheiten und Parameter wie Lautstärke, Tonhöhe, Sprechpausen oder Sprechgeschwindigkeit dazu, gesprochene Daten zu kodieren.²⁹ Wenngleich stimmbasierte Emotionserkennung und Sprachassistentz nicht denselben Fokus haben, sind beide Programme dennoch auf die gleiche Weise für Bias anfällig. Es konnte beispielsweise nachgewiesen werden, dass sie „die Stimme von männlichen *weißen*“³⁰ Amerikanern zu 92% richtig erkennen“, während es „bei *weißen* Amerikanerinnen nur 79%“ und „bei Menschen mit gemischter Abstammung [...] 69%“ sind.³¹

Es kann demnach als höchst problematisch bezeichnet werden, dass Sprachsysteme nicht mittels einer breit gefächerten Auswahl an Menschen mit sprachlichen Auffälligkeiten auf Analysen vorbereitet wurden. In einigen Fällen können auch sprachliche Besonderheiten – wie beispielsweise vermehrtes Stottern oder längere Sprechpausen – dazu führen, dass Daten falsch bewertet werden.³² Wird Stottern möglicherweise von der Software als Zeichen für das Vorliegen einer bestimmten Emotion gewertet, ergibt sich eine erhebliche Problematik bei der richtigen Zuordnung und Interpretation von Sprachdaten, wenn die Person, von der diese stammen, beispielsweise ungeachtet ihrer Emotionen immer stottert oder einzelne Wortteile verschluckt. Durch den Einsatz von Bio-Sensoren kann hierdurch eventuell Abhilfe geschaffen werden.

Während stimmbasierte und gesichtsbasierte Emotionserkennung bei der Identifizierung und Klassifizierung von visuellen oder akustischen Merkmalen technisch analysierend vorgehen und körperliche Parameter nicht erfassen, erlaubt der Einsatz von Bio-Sensoren eine vertiefte Möglichkeit der Wahrnehmung bestehender Emotionen. Der Fokus liegt hierbei insbesondere auf jenen Auffälligkeiten, die den physiologischen Aspekt bei der Entstehung und Aufrechterhaltung eines Gefühls beachten.³³ Wenn es sich bei einem EKG oder bestimmten Features in Smart Watches³⁴ – beispielsweise der Messung bestimmter

²⁸ Zur ausführlicheren Aufarbeitung dieser Thematik siehe insbesondere Kapitel 5 und 7.

²⁹ Vgl. Schiedermeier (2021), S. 25-26.

³⁰ Anm.: kursive Hervorhebung durch die Verfasserin der Arbeit

³¹ Schiedermeier (2021), S. 17 sowie Wong, Henry: Siri, Alexa and unconscious bias: the case for designing fairer AI assistants. In: Design Week 2020.

³² Schiedermeier (2021), S. 15-16.

³³ Vgl. Misselhorn (2021), S. 32.

³⁴ Vgl. Ebd. S. 32.

physiologischer Parameter – zwar nicht zwangsläufig um funktionstüchtige KI handelt, die grundsätzlich Gefühle erkennen und interpretieren kann, so dienen die Ergebnisse bestimmter Messungen dennoch dazu, Emotionen basierend auf körperlichen Merkmalen wie Schweißproduktion, erhöhtem Herzschlag oder Blutdruck zumindest als Abweichung in der Gefühlswelt wahrzunehmen und gegebenenfalls in Kombination mit weiteren Ansätzen deuten zu können.³⁵ Dobler beschreibt in diesem Zusammenhang eine relevante These. So geht er davon aus, dass sich KI in bestimmten Situationen – insbesondere im zwischenmenschlichen Bereich und damit verbundenem kommunikativem Austausch – als eine Instanz darstellt, die in jedweder Form der menschlichen Auffassung überlegen ist und hierdurch nahezu gottesgleich und allwissend erscheint.³⁶ Dies ist ein Trugschluss.

„Wir sollten uns immer bewusst sein, dass es sich nicht um Menschen, sondern um Maschinen handelt, die intelligentes Verhalten nur simulieren.“³⁷ Gerade diese Simulation geht mit Problemen einher, die das normale Ausmaß an Verantwortungsdiffusion weit überschreiten. Nicht nur, weil die Frage danach, wer befähigt sein darf, einer anderen Instanz Verantwortung zuzuschreiben, schwer zu beantworten ist, sondern auch, weil artifizielle Instanzen nicht ausschließlich einer, sondern mehreren wissenschaftlichen Disziplinen zugeordnet werden können. Je nach Klassifizierung bestünde möglicherweise die Chance, ein und derselben Instanz ein unterschiedliches Maß an Verantwortung zuzuschreiben. Da diese Problematik nicht erst in der MMI, sondern bereits im Rahmen der zugehörigen Teilbereichsethiken anfängt, werden diese im Folgenden kurz beleuchtet.

3. Maschinen- und Roboter[ethik]

Der Titel dieser Masterarbeit lautet „Die Utopie der Verantwortungszuschreibung in der Mensch-Maschine-Interaktion“. Dieser präzisen Formulierung zufolge wäre es naheliegend, anzunehmen, die vorliegende philosophisch-ethische Auseinandersetzung bezöge sich lediglich auf Maschinen, den ihnen zugehörigen Fachbereich sowie dessen Komplexität im Kontext einer vermeintlich bestehenden Verantwortungsdiffusion. Dies ist allerdings aus mehreren Gründen eine irreführende Annahme. Die Frage danach, welchem Fachbereich das dieser Arbeit zugrundeliegende Forschungsthema zuzuordnen ist, hängt maßgeblich davon ab, welcher disziplinären Gliederung gefolgt wird.

³⁵ Vgl. Misselhorn (2021), S. 33.

³⁶ Dobler, Ralph-Miklas: „Künstliche Intelligenz“ als Simulation und als Simulakrum. In: Brandstetter, Nicole, und Ralph-Miklas Dobler u.a. (Hrsg.): Mensch und Künstliche Intelligenz. Herausforderungen für Kultur, Wirtschaft und Gesellschaft. München: UVK Verlag 2021, S. 43.

³⁷ Schiedermeier (2021), S. 18.

Janina Loh begreift sowohl die Maschinen- als auch die Roboterethik als Teilbereiche der Angewandten Ethik. Diese und ähnliche Bereichsethiken wiederum werden hinsichtlich ihres Interessensfeldes strukturiert.³⁸ Sowohl Maschinen- als auch Roboterethik beziehen sich auf den „Umgang mit einem nichtmenschlichen Gegenüber“³⁹, während andere Bereichsethiken „ethische Systeme für Sonderbereiche des menschlichen Lebens, in denen Werte vertreten, Normen geltend gemacht und Regeln formuliert werden [...]“ versammeln würden.⁴⁰

Wenn Loh die Roboterethik als Teilbereich der Maschinenethik begreift, lässt sich dies dadurch begründen, dass Maschinen und Roboter nicht dieselben technischen Konstrukte bezeichnen, wenngleich beide Termini alltagssprachlich als Synonyme genutzt werden. Eine Schwierigkeit im terminologischen Kontext ergibt sich bei dem Versuch einer Definition beider Begriffe unter anderem dadurch, dass alle Roboter als Maschinen zu klassifizieren sind, während im Umkehrschluss nicht alle Maschinen als Roboter bezeichnet werden können.⁴¹ Die Frage nach den zentralen Unterschieden zwischen Maschinen und Robotern ist hierbei bei weitem komplexer und vielschichtiger als dies für eine problemorientierte Lösung der Verantwortungszuschreibung sowohl in der Mensch-Roboter-Interaktion als auch in der MMI wünschenswert wäre.

3.1. Maschinen

Loh stützt sich bei der Definition von Maschinen auf die Arbeit von Canguilhem und Strandh, wenn sie eine Maschine als „[...] ein künstliches Gebilde, das aus durch ein Antriebssystem (Motor, Wind, Wasser) bewegten Teilen besteht und Energie umsetzt“⁴² bezeichnet.⁴³ Auch Wolff offeriert eine ähnlich scharf abgegrenzte Definition, wenn er Maschinen als Kompositum mehrerer Einzelteile begreift.⁴⁴

Weder Lohs noch Wolffs Erläuterungen implizieren die Befähigung einer Maschine zu wie auch immer geartetem Lernen, eigenständigem bzw. autonomen Denken oder der Ausbildung ähnlicher Attribute, die im Kontext der MMI als relevant erachtet werden

³⁸ Vgl. Loh (2019), S. 10.

³⁹ Ebd. S. 10.

⁴⁰ Ebd. S. 11.

⁴¹ Vgl. Ebd. S. 10.

⁴² Ebd. S. 16.

⁴³ Zur Vertiefung siehe u.a. Canguilhem, Georges: Maschine und Organismus. In: Die Erkenntnis des Lebens. Berlin 2012, S. 213-227. sowie Strandh, Sigvard: Die Maschine. Geschichte – Elemente – Funktion. Freiburg im Breisgau: Herder 1980.

⁴⁴ Vgl. Wolff, Christian: Vernünftige Gedanken von Gott, der Welt und der Seele des Menschen, und auch allen Dingen überhaupt: den Liebhabern der Wahrheit mitgetheilet [1720]. Hildesheim 2009, S. 337.

könnten. Dies liegt allerdings weniger daran, dass eine Maschine grundlegend nicht lernfähig oder gar denkfähig ist, sondern vielmehr daran, dass Loh im Speziellen den zentralen Unterschied zwischen klassischen Maschinen und Robotern darlegt, indem sie die beiden Termini ursächlich als zu differenzierende Einheiten versteht. Wie bereits erwähnt, kann ein jeder Roboter immer auch als Maschine betrachtet und klassifiziert werden, während dies umgekehrt nicht möglich ist. Unter anderem aus diesem Grund wurde entschieden, im Titel der vorliegenden Arbeit das Wort „Maschine“ als Hyperonym zu verwenden, da hiermit sowohl schlicht nach Loh definierte Maschinen als auch hoch komplexe Roboter einbegriffen sind.

3.2. Roboter

Wenn Loh in ihren Werken verdeutlicht, dass eine Maschine das Potenzial hat, gleichsam ein Roboter zu sein, hierfür jedoch noch weitere Aspekte miteinbezogen und beachtet werden müssen, stellt sich als nächstes die Frage danach, was nun unter einem Roboter zu verstehen ist. An dieser Stelle eignet sich Misselhorns Begriffsdefinition, um Lohs terminologische Bestrebungen zu ergänzen. Sie postuliert, dass ein

[...] Roboter eine elektro-mechanische Maschine ist, die a) über einen eigenständigen Körper und b) über mindestens einen Prozessor verfügt, c) Sensoren hat, die Informationen über die Welt sammeln, d) sowie über Effektoren und Aktoren verfügt, die Signale in mechanische Abläufe übersetzen. Das Verhalten eines Roboters e) ist oder erscheint zumindest autonom, und er kann f) in seine Umgebung hineinwirken beziehungsweise physisch auf sie Einfluss nehmen.⁴⁵

Diese begriffliche Erläuterung impliziert, dass insbesondere die Fähigkeit einer Maschine, autonom zu sein oder so zu erscheinen, ein erstes unausweichliches Erfordernis ist, um eine Maschine als Roboter klassifizieren zu können. Auch die Fähigkeit, einen Einfluss auf die nähere und möglicherweise auch ferne Umwelt auszuüben, scheint bei Misselhorn als eigenes Anforderungsmerkmal auf. Interessant ist, dass die vorliegende Eingrenzung den Anschein macht, es sei ohne weiteres möglich, Roboter so zu definieren, dass bei jeder erforderlichen Klassifizierung keinerlei Zweifel aufkommen, worum es sich bei etwaigen artifiziellen Systemen handelt, solange diese den Kriterien a bis f entsprechen. Dieser Annahme widerspricht Michael Funk aus zwei Gründen. Erstens sei dasjenige, was gemeinhin unter dem Terminus „Roboter“ subsummiert wird, zu vielfältig und komplex, um hierfür lediglich einen Begriff zu verwenden. Zweitens würden sich sämtliche artifiziellen

⁴⁵ Misselhorn, Catrin: Robots as Moral Agents? In: Rövekamp, Frank und Friederike Bosse (Hrsg.): Ethics in Science and Society. German and Japanese Views. München: IUDICIUM Verlag GmbH 2013, S. 43.

Systeme inklusive deren Besonderheiten und Eigenschaften stetig verändern, wodurch eine allgemein gültige Definition unrealistisch bis utopisch sei.⁴⁶

Funk ergänzt den bestehenden Diskurs um weitere Überlegungen, indem er daran erinnert, dass nicht nur die Maschinen- und Roboterethik als Disziplinen gelten, die mit Robotern in Verbindung gebracht werden können, sondern dass auch die Robotik eng mit dem Bestehen von Robotern konnotiert ist.⁴⁷ Auch Bendel verweist hierauf, indem er Roboter zwar als jene Maschinen ansieht, „mit denen sich die Maschinenethik beschäftigt“⁴⁸, er allerdings ergänzt: Die „Robotik muss sich zusammen mit der Maschinenethik (und der Roboterethik) mit der Moral von Robotern auseinandersetzen.“⁴⁹ Spätestens an dieser Stelle wird klar, weshalb zu Beginn des Kapitels darauf hingewiesen wurde, dass der für diese Arbeit relevante Fachbereich keineswegs eindimensional zu betrachten ist. Vielmehr greifen sowohl die Aufgabenbereiche der genannten Disziplinen als auch die zugehörigen Definitionen in Teilen oder gar gänzlich ineinander über.

Dies wird ebenfalls verdeutlicht, wenn in Betracht gezogen wird, wie viele unterschiedliche Erläuterungen zu dem Terminus „Roboter“ aktuell beständig sind. Während Misselhorn eine Beschreibung anbietet, die sich vorwiegend an Kompetenzen orientiert, über die eine Maschine verfügen muss, um als Roboter klassifiziert werden zu können, fokussieren sich andere Definitionen auf die Erklärung jener Tätigkeiten, die ein Roboter durchführt oder die Beschaffenheit des Konstrukts selbst. Auch diese terminologische Herangehensweise ist technisch relevant, wenngleich sie wenig Anhaltspunkte für eine eindeutige Zuordnung oder Differenzierung unterschiedlicher Arten von Robotern bietet. Schließlich gibt es auch Begriffsdefinitionen, die auf eine Beschreibung abzielen, mittels derer sowohl Kompetenzen als auch technische Merkmale und Funktionen festgehalten werden. Hier knüpft Nevejans an, die sich zwar auf eine Definition „intelligenter Roboter“ bezieht, diese allerdings multidimensional aufbaut. “[...] an intelligent robot: acquires autonomy through sensors and/or by exchanging data with its environment (inter-connectivity) and trades and analyses

⁴⁶ Funk, Michael: Computer und Gesellschaft: Roboter und KI als soziale Herausforderung – Grundlagen der Technikethik, Band 3. Wiesbaden: Springer Fachmedien 2022, S. 5.

⁴⁷ Vgl. Funk (2022), S. 5.

⁴⁸ Bendel, Oliver: 300 Keywords Informationsethik. Grundwissen aus Computer-, Netz- und Neue Medien-Ethik sowie Maschinenethik. Wiesbaden: Springer 2016, S. 192.

⁴⁹ Bendel (2016), S. 192.

data; is self-learning (optional criterion); has a physical support; adapts its behaviours and actions to its environment.”⁵⁰ Christaller u.a. bieten wiederum die folgende Definition an:

Roboter sind sensumotorische Maschinen zur Erweiterung der menschlichen Handlungsfähigkeit. Sie bestehen aus mechatronischen Komponenten, Sensoren und rechnerbasierten Kontroll- und Steuerungsfunktionen. Die Komplexität eines Roboters unterscheidet sich deutlich von anderen Maschinen durch die größere Anzahl von Freiheitsgraden und die Vielfalt und den Umfang seiner Verhaltensformen.⁵¹

Was hier bereits vor einigen Jahren als „Erweiterung der menschlichen Handlungsfähigkeit“ bezeichnet wurde, entspricht mittlerweile einem breiten Spektrum an unterschiedlich konzipierten Robotern. Beispielsweise kollaborative Roboter, die in verschiedenen Fachbereichen – unter anderem der Industrie – als Assistenz oder Interaktionspartner menschlicher Arbeitskräfte fungieren können, stehen aktuell hoch im Kurs.⁵² Relevant ist es in diesem Kontext zu erläutern, dass kollaborative Roboter nicht mit sozialen oder Service-Robotern gleichzusetzen sind:

Denn diese werden nicht in Industrieanlagen betrieben, sondern Pflegeheimen, Restaurants oder privaten Wohnräumen. Deren offene Arbeitsumgebungen sind also um ein Vielfaches komplexer. Entsprechend des angestrebten *Nutzens* sprechen wir dann auch von **Pflegerobotern, Servicerobotern oder Haushaltsrobotern**, die wiederum Untergruppen wie **Unterhaltungsroboter, Staubsaugroboter oder Lernroboter** einschließen können.⁵³

In dieser Aufzählung kommt der Begriff „sozialer Roboter“ bzw. „Social Bot“ zwar nicht vor, allerdings ist es in Anbetracht jener Bereiche der Robotik, die im Rahmen dieser Arbeit angesprochen werden, sinnvoll, diesen bereits an dieser Stelle kurz zu erwähnen. Soziale Roboter und Serviceroboter ähneln sich in vielerlei Hinsicht in den ihnen zugeschriebenen Aufgabenbereichen und werden daher auch synonym verwendet, allerdings ist darauf hinzuweisen, dass soziale Roboter insbesondere für empathische dialogische Interaktionen im alltäglichen zwischenmenschlichen Bereich eingesetzt werden, während bei Servicerobotern nicht notwendigerweise dieser Schwerpunkt gesetzt wird.⁵⁴ Insbesondere in Bezug auf die Verantwortungszuschreibung ist diese Differenzierung vonnöten, denn ein Roboter, der für soziale Interaktionen eingesetzt wird, muss zweifelsfrei über andere Fähigkeiten verfügen als eine Maschine, die im Servicebereich tätig ist.

⁵⁰ Nevejans, Nathalie: European Civil Law Rules in Robotics. European Union: October 2016, S. 8. [http://www.europarl.europa.eu/RegData/etudes/STUD/2016/571379/IPOL_STU\(2016\)571379_EN.pdf](http://www.europarl.europa.eu/RegData/etudes/STUD/2016/571379/IPOL_STU(2016)571379_EN.pdf) (8.7.2023)

⁵¹ Christaller, Thomas und Michael Decker u.a.: Robotik: Perspektiven für menschliches Handeln in der zukünftigen Gesellschaft 14. Berlin/Heidelberg: Springer 2013, S. 19; zitiert nach Funk (2022), S. 9.

⁵² Vgl. Funk (2022), S. 10.

⁵³ Ebd. S. 11.

⁵⁴ Ebd. S. 11.

4. Verantwortung

4.1. Kontextualisierung

Der Terminus „Verantwortung“ ist ein sprachliches Konstrukt, das im Alltagssprachlichen Gebrauch und Wortschatz gesellschaftlicher wie auch privater Konversationen und Debatten vorkommt. Wird es umgangssprachlich zumeist eingesetzt, um Schuld für bereits vergangene Ereignisse und deren mehr oder weniger ungünstige, problematische oder gar katastrophale Auswirkungen zuzuschreiben, stellt sich die Frage, ob Verantwortung grundsätzlich erst im Anschluss an eine gesetzte Handlung zugeschrieben werden kann, oder ob Verantwortlichkeit bereits im Vorhinein delegiert werden muss.

Ist ersteres der Fall, ergibt sich die Folgefrage, ob Verantwortungszuschreibung in sich eine Utopie darstellt, ohne dass hierbei überhaupt erst der Kontext der MMI herangezogen oder hergestellt werden muss. Ist zweiteres der Fall, bleibt selbstverständlich zu klären, wie Verantwortung und Verantwortlichkeit zugeschrieben werden können, aber auch, welche Instanzen Verantwortung tragen und welche diese delegieren können. Insbesondere in Anbetracht der in den obigen Kapiteln dargelegten sehr gravierenden Unterschiede und Auffassungen der Bedeutung und Definition von Künstlicher Intelligenz, Maschinen und Robotern jeglicher Art, erhält die Frage danach, ob Verantwortung als solche ein greifbares Konstrukt darstellt, eine neue Dimension.

So ist – wie im Folgenden dargelegt werden wird – nicht nur im Kontext der zwischenmenschlichen Interaktion die Frage nach der Verantwortungszuschreibung beziehungsweise der ursächlichen Bedeutung des Wortes alles andere als eindeutig. Erweitert man den Diskurs nun um nicht-menschliche Instanzen, die jedoch ihrerseits wiederum nicht immer nach wissenschaftlichem Konsens in moralische Handlungssubjekte oder moralische Handlungsobjekte gegliedert werden können, liegt der Folgeschluss nahe, dass die gesamte Debatte der Verantwortungszuschreibung [in der Mensch-Maschine-Interaktion] eine Utopie kreiert, die in ihrem Bestehen insbesondere in Hinblick auf zukünftige Forschung und Weiterentwicklung artifizieller Systeme nicht unproblematisch ist. Aus diesem Grund wird im folgenden Kapitel versucht, einen Überblick über möglichst viele Definitionsversuche und Begriffserläuterungen zu „Verantwortung“ vorzustellen, um diese in weiterer Folge kritisch beleuchten und auf unterschiedliche Bereiche der Robotik und Maschinenethik beziehen zu können.

4.2. Terminologie und Abgrenzungen

Die Existenz eines Verantwortungsbegriffs kann ab der zweiten Hälfte des 15. Jahrhunderts nachgewiesen werden. Eine allgemeine Verwendung der Begrifflichkeit etablierte sich im deutschsprachigen Raum sowie in der französischen Sprache allerdings erst im Laufe des 18. Jahrhunderts, in der italienischen Sprache sogar erst im 19. Jahrhundert.⁵⁵ Wenn davon ausgegangen wird, dass es sich hierbei um ein klar abzugrenzendes und einheitliches Konstrukt handelt, das sprachübergreifend dasselbe bezeichnet, so ereignet sich genau hierbei ein Irrtum, der einer strukturierten Aufklärung bedarf. Die wissenschaftliche Auseinandersetzung mit philosophischer Fachliteratur veranschaulicht deutlich, dass die Vorstellung einer allgemein gültigen Definition des Verantwortungsbegriffs irreführend ist. Vielmehr handelt es sich um ein multidimensionales und kompliziertes Konzept, das in seinen terminologischen Ausführungen nicht unterschiedlicher sein könnte.⁵⁶

Konsens besteht allerdings darin, dass die Zuschreibung von Verantwortung dazu dient, einer Instanz, die für ihr Verhalten zur Rechenschaft gezogen werden kann, die ihr innewohnende Fähigkeit der Rechtfertigung auch tatsächlich abzuverlangen.⁵⁷ Dass eine Person „Rede und Antwort stehen“⁵⁸ muss, ist hierbei eine Erwartung, die sich nicht nur im Werk von Kahle und Weidtmann wiederfindet, sondern unter anderem auch bei Heidbrink, Langbehn und Loh. Heidbrink verweist in diesem Kontext darauf, dass der Verantwortungsbegriff in seiner Intentionalität Ähnlichkeiten zum klassischen philosophischen Pflichtbegriff aufweist, allerdings nicht als Synonym verwendet werden kann, da „[...] sich die Verantwortung eines Akteurs nicht nur auf die Einhaltung einer Norm oder Regel, sondern auch auf die Berücksichtigung der Folgen, die sich aus der Umsetzung einer Norm oder Regel ergeben“⁵⁹, bezieht.

⁵⁵ Ritter, Joachim und Karlfried Gründer u.a. (Hrsg.): Historisches Wörterbuch der Philosophie. Onlineversion Gesamtwerk. Basel: Schwabe AG Verlag. https://www.schwabeonline.ch/schwabe-xaveropp/elibrary/start.xav?start=%2F%2F%2A%5B%40attr_id%3D%27hwph_productpage%27%5D#__elibrary__%2F%2F%5B%40attr_id%3D%27verw.verantwortung%27%5D__1681246470785 (11.4.2023)

⁵⁶ Vgl. Heidbrink, Luger: Definitionen und Voraussetzungen der Verantwortung. In: Heidbrink, Ludger und Claus Langbehn u.a. (Hrsg.): Handbuch Verantwortung. Wiesbaden: Springer Fachmedien Wiesbaden GmbH 2017, S. 4.

⁵⁷ Vgl. Höffe, Otfried: Verantwortung: Begriff und Prinzip. In: Kahle, Reinhard und Niels Weidtmann: Verantwortung. Ein Begriff in seiner Aktualität. Boston: BRILL 2021. S. 3.

⁵⁸ Höffe (2021), S. 3, siehe auch: Heidbrink (2017), S. 4.

⁵⁹ Heidbrink (2017), S. 5, siehe auch Schwartländer, Johannes: Verantwortung. In: Hermann Krings und Hans Michael Baumgarten u.a. (Hrsg.): Handbuch philosophischer Grundbegriffe Band 6. München: Kösel 1974, S. 1577–1588. sowie Ott, Konrad: Verantwortung. In: Grupe, Ommo und Dietmar Mieth (Hrsg.): Lexikon der Ethik im Sport. Schorndorf: Karl Hofmann 1998, S. 578–587.

Pflichten und Verpflichtungen implizieren demnach nicht uneingeschränkt die Absehbarkeit zu erwartender Handlungsfolgen, während dies im Bereich der Verantwortung sehr wohl der Fall ist. Wenngleich Pflicht und Verantwortung keine austauschbaren Termini darstellen, stehen die beiden Konzepte in einer Wechselwirkung, die sich dadurch auszeichnet, dass die Forderung zur Einhaltung bestimmter Pflichten eng an die Zurechnungsfähigkeit einer Instanz gebunden ist und somit gleichsam als eine Voraussetzung für die legitime Möglichkeit einer Zuschreibung von Verantwortung angesehen werden kann.⁶⁰

Angemerkt sei in diesem Zusammenhang auch, dass Verantwortung und deren Herausforderungen eng mit dem Vorhandensein einer intakten Ratio sowie der Artikulationsfähigkeit durch Sprache verbunden sind.⁶¹ Ob mit Sprache lediglich die menschliche Sprache gemeint ist und ob diese wiederum natürlicherweise durch Stimmbandkontraktionen hervorgerufen werden muss, oder ob auch simulierte Äußerungen, die beispielsweise durch ein artifizielles System kreiert werden, als Sprache gelten, wird nicht näher erläutert. Ebenso erscheint an dieser Stelle unklar, wie es sich mit Individuen oder Instanzen verhält, die zwar über praktische Vernunft verfügen, diese allerdings aus unterschiedlichen Gründen nicht sprachlich umsetzen und beweisen können. Müssen diese Instanzen nicht Rede und Antwort für ihr Handeln stehen? Und falls dies bejaht wird, müssen sie es deshalb nicht, weil sie keine sprachliche Artikulationsmöglichkeit besitzen oder weil sie erst gar nicht als zurechnungsfähige Wesen angesehen werden?

Was auf den ersten Blick logisch erscheint, enthält bei genauerer Betrachtung problematische Aspekte. Der Versuch einer Eingrenzung des Verantwortungsbegriffs im Sinne einer Definition liefert allerdings Bausteine, die das Gesamtkonstrukt als solches greifbarer werden lassen. Verantwortung ist nach Höffe dreidimensional zu verstehen und findet auch auf diese Weise gesellschaftliche Anwendung:

Als erstes beansprucht man eine Autorität und Entscheidungsbefugnis, die nicht abgegeben oder sonstwie [sic!] übertragen werden kann [...] Zweitens erhebt man mit dem Ausdruck Anklage [...] und fordert Rechenschaft ein [...] Schließlich steht der Ausdruck für die Aufgabe, für Fehler, die passiert sind, oder für Dinge, die falsch gelaufen sind, die eigene Schuld anzuerkennen.⁶²

Interessant ist, dass von Autorität gesprochen wird, die nicht verlagert werden kann, die also einer bestimmten Instanz aus einem klar definierten Grund zugesprochen wird. Der Grund des Zuspruchs von Verantwortung kann hierbei offenbar einerseits Bezug auf einen neutralen Zustand nehmen, für den möglicherweise Rechenschaft abgelegt werden wird

⁶⁰ Vgl. Heidbrink (2017), S. 6.

⁶¹ Vgl. Ebd. S. 3.

⁶² Ebd. S. 3.

müssen, vor allem scheint aber bei Höffe der Aspekt der Schuld eine zentrale Rolle zu spielen, schließlich liegt das Augenmerk auf Begriffen wie „Anklage“ oder „Fehler“. Beide verweisen auf eine negative Konnotation des Verantwortungsbegriffs. Schuld und Verantwortung synonym zu lesen, scheint dennoch fragwürdig, wenngleich es zum derzeitigen Stand der Betrachtungen nicht auszuschließen ist, dass entstandene Schuld möglicherweise überhaupt erst zu einer rückwirkenden Zuschreibung von Verantwortung führt. Im Folgenden werden unterschiedliche Verantwortungskonzepte vorgestellt, um die Relationen innerhalb der Strukturen der Verantwortungszuschreibung differenziert betrachten und miteinander vergleichen zu können.

4.3. Formen der Verantwortungszuschreibung

4.3.1. Verantwortung bei Otfried Höffe und Claudia Blöser

In Anbetracht der erwähnten Autorität und Entscheidungsbefugnis können Höffes Erläuterungen mithilfe einer Gliederung von Verantwortung in Primär- Sekundär- und Tertiärverantwortung nachvollzogen werden. Ein Verständnis der Primärverantwortung setzt voraus, dass es eine*n eindeutig zu identifizierende*n Handelnde*n gibt, der*die als Verantwortliche*r für bestimmte Aufgaben fungiert - Primärverantwortung ist demnach spezifisch und auf klar definierte Funktionsbereiche beschränkt.⁶³ Anders als bei dem obig abgebildeten groben Definitionsversuch kristallisiert sich durch eine Spaltung des Verantwortungsbegriffs heraus, dass das Konzept nicht automatisch mit Schuld und Irrtum zusammenhängen muss, sondern sich durchaus auf Instanzen beziehen kann, die grundsätzlich aktiv Verantwortung tragen, ohne dass jedoch zwangsläufig die Folgen einer Handlung bereits vollständig absehbar sind. Dem entgegen steht die Idee einer generellen und nicht zu verlagernden Sekundärverantwortung, die genau diese „Zuständigkeit für die Folgen und Nebenfolgen des eigenen Tuns oder Lassens“⁶⁴ in den Fokus rückt: „Die Sekundärverantwortung, »zu der man gezogen wird«, insofern einen passiven Charakter hat [sic!], ist eine Rechenschaftsverantwortung. Aus diesem Grund enthält sie das Potential der Anschuldigung, zumindest der Verdächtigung, [...]“⁶⁵

Die hier angesprochene Sekundärverantwortung konnotiert direkt oder indirekt mit dem Konzept der Tertiärverantwortung, welches immer dann zum Tragen kommt, wenn ein Individuum aufgrund einer abgelegten Rechenschaft entweder seiner Schuld bezichtigt wird

⁶³ Vgl. Höffe (2021), S. 4.

⁶⁴ Ebd. S. 4.

⁶⁵ Ebd. S. 4.

oder aber hiervon freigesprochen wird.⁶⁶ Mehr noch als im Zuge der Sekundärverantwortung haftet nun jedwede*r Handelnde*r für die von ihm*ihr getätigten Entscheidungen und deren Folgen. Eine Gliederung in Primär- Sekundär- und Tertiärverantwortung wird bei Höffe detailliert beschrieben, findet aber auch bei Bendel ihren Niederschlag. Hier wird neben der Terminologie auch der Nutzen besagter Verantwortungstypen diskutiert. Bendel postuliert, dass alle drei Verantwortungsarten miteinander in Verbindung stehen, Primärverantwortung allerdings immer notwendige Voraussetzung für eine Zuschreibung weiterer Verantwortungsdimensionen ist. Des Weiteren begreift Bendel Personen, die Primär- und Sekundärverantwortung zu tragen haben, als „Subjekt[e] der Moral“, während Tertiärverantwortung sie auch als „Subjekt[e] von Recht und Ordnung“ zur Geltung bringt.⁶⁷

Interessant ist, dass als Handlungstragende sowohl bei Höffe als auch bei Bendel lediglich „Personen“ gelten. Von welchem Personenbegriff hierbei ausgegangen wird und ob es philosophisch möglich ist, das vorliegende Verantwortungskonzept auf nicht-menschliche-Entitäten oder digitale Personen anzuwenden, ist fraglich und erweitert den Diskurs um die Problematik einer Verantwortungszuschreibung im Bereich der MMI. Eckert gibt in diesem Kontext zu verstehen, dass Verantwortungskonzeptionen ihrerseits mit unterschiedlichen Auffassungen des Personenbegriffs einhergehen, der Personenbegriff selbst jedoch immer Handlungs- und Schuldfähigkeit impliziert.⁶⁸ Verantwortungskonzepte würden sich demzufolge dahingehend differenzieren lassen, „[...] welche Rolle Intentionalität spielt, ob oder in welchem Maße ein Mensch in seinem Handeln durch freie Entscheidung oder aber durch andere Kräfte – seien dies Hormone, Geister, Karma oder aber die sozialen Bedingungen – geleitet wird und welche Rolle dies für Haftung und Verpflichtung spielt.“⁶⁹

Im Sinne der Vollständigkeit sei an dieser Stelle im Rahmen eines kurzen Exkurses Spiecker gen. Döhmann erwähnt, die die Frage nach der Verantwortungszuschreibung nicht nur philosophisch, sondern vor allem juristisch beleuchtet. Im Kontext der MMI verweist sie nicht zuletzt auf die ungeahnten Gefahren und Risiken, die die fortschreitende systemische Digitalisierung hat, sondern auch auf das fehlende Verständnis für die dadurch entstehenden Probleme bei der Verantwortungs- aber insbesondere der Haftungszuschreibung.⁷⁰

⁶⁶ Vgl. Höffe (2021), S. 4.

⁶⁷ Bendel (2016), S. 233.

⁶⁸ Vgl. Eckert, Julia: Tugend, Recht und Moral: Tendenzen der Verantwortungszuschreibung. In: DZPhil 64(2) 2016, S. 247-248.

⁶⁹ Eckert (2016), S. 247.

⁷⁰ Spiecker gen. Döhmann, Indra: Zur Zukunft systemischer Digitalisierung – Erste Gedanken zur Haftungs- und Verantwortungszuschreibung In: Computer und Recht 32(10) 2016, S. 699.

Rechtliche Maßnahmen zur Kontrolle artifizieller Systeme beispielsweise würden in diesem Kontext einer Erneuerung des Personenbegriffs bedürfen.

Dazu gehört bereits die Grundkonzeption der Zuweisung von Rechten und Pflichten, die an eine natürliche oder juristische Person erfolgt. Eines der tragenden Konzepte des modernen Rechts und eines seiner bemerkenswerten Leistungen ist die Ausgestaltung der juristischen Persönlichkeit. Nicht allein natürliche Personen können Zuschreibungssubjekte für Rechte und Pflichten sein, sondern auch juristische Personen, sowohl des Privat- als auch des Öffentlichen Rechts. Damit geht gleichzeitig einher, dass Rechte und Pflichten stets personal gedacht werden, und dass eine klare Verantwortungszuschreibung besteht.⁷¹

Diese Erläuterungen ermöglichen es, einen Personenbegriff um technische Instanzen zu erweitern. Während menschliche Instanzen automatisch als natürliche Personen gelten, könnten unter dieser Prämisse auch Roboter und ausgewählte Maschinen als [juristische] Personen erfasst werden und im Zuge dessen als Träger*innen der Verantwortung für bestimmte Handlungen bezeichnet werden. Dieser Gedanke wirkt rechtlich schlüssig, ist jedoch philosophisch und ethisch bei weitem komplexer als es auf den ersten Blick den Anschein hat. Schließlich ist das Dilemma der Verantwortungszuschreibung nicht dadurch gelöst, den Personenbegriff zu expandieren. Ich vertrete viel eher die These, dass dieser enger und spezifischer sein müsste, um der bestehenden Utopie der Verantwortungszuschreibung langfristig entgegen zu können. Hier knüpft auch Hans Jonas an, der sich mit der Problematik und den Risiken der technologischen Weiterentwicklung in Bezug auf die ethische Vertretbarkeit einer wie auch immer gearteten Verantwortungszuschreibung auseinandersetzt.

Im Zeichen der Technologie aber hat es die Ethik mit Handlungen zu tun (wiewohl nicht mehr des Einzelsubjekts), die eine beispiellose kausale Reichweite in die Zukunft haben, begleitet von einem Vorwissen, das ebenfalls, wie immer unvollständig, über alles ehemalige weit hinausgeht. Dazu die schiere Größenordnung der Fernwirkungen und oft auch ihre Unumkehrbarkeit. All dies rückt Verantwortung ins Zentrum der Ethik, und zwar mit Zeit- und Raumhorizonten, die denen der Taten entsprechen.⁷²

Hans Jonas dürfte demnach nicht nur der fehlende philosophische Konsens zum Verantwortungsbegriff als unzureichend erscheinen, sondern insbesondere auch eine klar definierbare Dimension fehlen, die benötigt würde, um nicht-menschliche Entitäten, dh. Roboter, Maschinen und andere artifizielle Systeme, überhaupt erst als Verantwortungsträger*innen klassifizieren zu können. Wie in Kapitel 6 – insbesondere Kapitel 6.2. – erläutert werden wird, gibt es auch hierzu eine Reihe bestehender Konzepte und Ideen, die sich in Teilen zwar ähneln, jedoch nicht zweifelsfrei mit den hier diskutierten Verantwortungsbegriffen verknüpft werden können. Am ehesten ist dies möglich, wenn der

⁷¹ Spiecker gen. Döhmman (2016), S. 700.

⁷² Jonas, Hans und Robert Habeck: Das Prinzip Verantwortung: Versuch einer Ethik für die technologische Zivilisation. Berlin: Suhrkamp 2020, S. 8.

Blick nicht alleinig darauf gerichtet wird, was Verantwortung ist, sondern auch, wie sie zu differenzieren ist, und welche Voraussetzungen hiermit verbunden sind.

4.3.2. Verantwortung bei Claudia Blöser

In diesem Zusammenhang sind Blösers Erläuterungen zu Verantwortlichkeit ertragreich. Angelehnt an Immanuel Kant beleuchtet Claudia Blöser nicht in erster Linie die Relation von Verantwortung und Schuld, sondern stellt die Termini Verantwortung und Zurechnung(sfähigkeit) einander gegenüber. Anzumerken ist, dass auch bei Blöser Zurechnung und Schuld eng miteinander korrelieren.⁷³ Dasjenige, was Höffe als „praktische Vernunft“ bezeichnet, kann im Sinne Blösers allerdings als das Vorhandensein eines „freien Willen[s]“ gelesen werden, da jedes Ereignis zwangsläufig auf den freien Willen der*des Ursachers*Ursacherin zurückgeführt wird.⁷⁴ Reaktionen auf bestimmte Handlungen sind demnach nur dann adäquat, wenn zwischen Handelnder*Handelndem und den Folgen einer Handlung das Kriterium der Zurechenbarkeit erfüllt wird.⁷⁵ Auch Blöser kritisiert im Rahmen ihrer Erläuterungen, dass der Verantwortungsbegriff weniger greifbar und unklarer sei, als dies für eine gezielte ethische und philosophische Auseinandersetzung wünschenswert wäre.⁷⁶ Aus diesem Grund zieht sie Kants Begriff der Zurechnung heran, um das Konstrukt „Verantwortung“ detaillierter darzustellen. Vorgeschlagen wird hierbei ein Verständnis von Verantwortung als pro- oder retrospektiv.⁷⁷

Interessanterweise fällt hierbei auf, dass Blösers Gliederung inhaltlich trotz differenzierter Begriffsbenennung jener von Höffe ähnelt. Prospektive Verantwortung bei Blöser umfasst die Verantwortlichkeit einer Instanz für Handlungen oder Ereignisse, die in Zukunft umgesetzt werden- oder stattfinden sollen.⁷⁸ Bei Höffe würde diese Bezeichnung unter den Aspekt der Primärverantwortung fallen. Auch hier sind Folgen von bestimmten Handlungen nur vermeintlich oder gar nicht absehbar, wobei gleichzeitig dennoch bestimmten Personen die Verantwortung für die zu setzenden Handlungen grundsätzlich zugeschrieben wird. Die retrospektive Verantwortung nach Blöser beschreibt nun das gegenteilige Szenario. Verantwortlichkeit bezieht sich hier deutlich auf bereits gesetzte Handlungen, für die

⁷³ Vgl. Blöser, Claudia: Zurechnung bei Kant. Zum Zusammenhang von Person und Handlung in Kants praktischer Philosophie. Boston: De Gruyter 2014, S. 1.

⁷⁴ Vgl. Blöser (2014), S. 1.

⁷⁵ Vgl. Ebd. S. 1.

⁷⁶ Vgl. Ebd. S. 3.

⁷⁷ Vgl. Ebd. S. 3.

⁷⁸ Vgl. Ebd. S. 3.

nachträgliche Verantwortung zugeschrieben wird.⁷⁹ Dieses Konzept ähnelt Höffes Verständnis von Sekundär- und/oder Tertiärverantwortung, allerdings fehlt bei Höffe ein für das Thema dieser Arbeit zentraler Aspekt, den Blöser am Rande erwähnt: „Retrospektive Verantwortung bindet Verantwortung offenbar an Urheberschaft, d. h. Personen müssen normalerweise nur für diejenigen Handlungen einstehen, deren Urheber sie sind. So wird ersichtlich, dass der Begriff der Zurechnung eng mit dem Begriff der Verantwortung in retrospektiver Bedeutung verbunden ist.“⁸⁰

Urheberschaft als Bedingung für die Zuschreibung von Verantwortung scheint eine wichtige Voraussetzung für eine philosophische Auseinandersetzung mit Verantwortlichkeit in der MMI zu sein. Fraglich ist allerdings, wie Urheberschaft im Sinne eines*einer nicht-menschlichen Handelnden zu definieren ist. Darf eine Maschine unter bestimmten Voraussetzungen als Urheber*in der von ihr produzierten Handlungsfolgen angesehen werden, oder bezieht sich Urheberschaft immer auf den tatsächlichen Ursprung des vermeintlichen Denkens? Im Falle eines artifiziellen Systems könnte hierbei diskutiert werden, ob als Urheber*in die Maschine selbst gelten darf, oder ob jene Menschen, die die Maschine mittels des Einsatzes von Algorithmen trainiert haben, als Urheber*innen zu bezeichnen sind. Um diese Frage zu beantworten, werden im weiteren Verlauf dieser Arbeit die Konzeptionen von Robotern und Maschinen als moralische Handlungssubjekte oder moralische Handlungsobjekte herangezogen. Da diese allerdings in Struktur und Aufbau bereits von der Möglichkeit einer Verantwortungszuschreibung in bestimmten Fällen ausgehen, muss noch eingehender geklärt werden, welche Auffassung von Verantwortung hierbei überhaupt zum Tragen kommen könnte und inwiefern diese generalisierbar oder zumindest spezifisch einsetzbar ist.

4.3.3. Verantwortung in der Wissenschaft

Nachdem ein Überblick über die Terminologie der Verantwortung gegeben wurde und zwei geläufige philosophische Konzepte beleuchtet wurden, liegt nun die Frage nahe, ob das philosophisch-theoretische Verständnis von Verantwortung in der Wissenschaft und ethischen Praxis anwendbar ist, oder ob sich hierbei Probleme ergeben, die insbesondere im Bereich der Roboter-, Maschinen- und Technikethik zu Schwierigkeiten führen. In diesem Sinne sei zunächst angemerkt, dass die Wissenschaft als allgemeine Disziplin bis zur frühen Neuzeit ein anderes Ideal verfolgte, als dies jetzt der Fall ist. So wurde lange Zeit das

⁷⁹ Vgl. Blöser (2014), S. 3.

⁸⁰ Ebd. S. 3.

Erlangen eines reinen Wissens als zentrales Element wissenschaftlicher Forschung angesehen, während mittlerweile „eine anwendungsbezogene Forschung, die den subjektiven Zwecken der Menschen dienen soll“⁸¹, im Fokus steht. Nun ist Wissenschaft alles andere als einschneidend und die Arten der Generierung von Wissen und dessen Anwendung unterscheiden sich je nach Disziplin voneinander. Naturwissenschaftliche Erläuterungen und Beweise lassen sich vergleichsweise objektiv durch Berechnungen erbringen, was in geisteswissenschaftlichen Disziplinen vermeintlich nicht möglich ist.⁸²

Wenn nun die Frage nach der Verantwortungszuschreibung im wissenschaftlichen Kontext betrachtet wird, liegt es nahe, die Ergebnisse der Forschung als Maßstab dafür heranzuziehen, welchen Instanzen aus welchen Gründen Verantwortung zugeschrieben werden kann und muss. Gerade im Bereich der technischen Entwicklung artifizieller Programme, Systeme und Roboter ist es allerdings relevant, auch den Prozess der Forschung im Rahmen jeglicher Verantwortungszuschreibung zu beachten.⁸³ Während sich die bisherig dargelegten Verantwortungskonzepte maßgeblich auf die Endresultate bestimmter Entscheidungen und Handlungen beziehen, ergeben sich durch die Verschiebung des Fokus von Resultat auf den Entstehungsprozess neue Perspektiven der Verantwortlichkeit, die insbesondere im Kontext der Maschinen- und Roboterethik relevante Anhaltspunkte für Bearbeitung der bestehenden Utopie einer Verantwortungszuschreibung liefern könnten.

4.4. Verantwortungsmodelle im wissenschaftlichen Kontext

4.4.1. Rollenverantwortung – Interne Verantwortung

Die Rollenverantwortung zum ersten ist an bestimmte Rollen oder Aufgaben geknüpft, die man beispielsweise als Mutter oder Professorin im sozialen oder beruflichen Leben übernimmt. [...] Bei der Kausalhandlungsverantwortung zum zweiten ist man für die durch das eigene Handeln verursachten Folgen verantwortlich. Da ein Wissenschaftler für kausale Folgen oder Auswirkungen nicht nur gegenüber der eigenen Forschungsgemeinschaft, sondern gegenüber Außenstehenden, der Gesellschaft insgesamt oder der Natur Verantwortung übernehmen muss, spricht man hier von externer Verantwortung.⁸⁴

In dieser Darlegung Fenners spiegelt sich deutlich die Komplexität des wissenschaftlichen Verantwortungsbegriffs wider. Im folgenden Kapitel werden im Sinne Fenners die Termini „Rollenverantwortung“, „innerwissenschaftliche Verantwortung“ und „interne

⁸¹ Fenner, Dagmar: Einführung in die Angewandte Ethik. Tübingen: Narr Francke Attempto Verlag² 2022, S. 205.

⁸² Vgl. Fenner (2022), S. 206.

⁸³ Vgl. Nida-Rümelin, Julian: Wissenschaft. In: Rippe, Klaus Peter (Hrsg.): Angewandte Ethik. Die Bereichsethiken und ihre theoretische Fundierung. Ein Handbuch. Stuttgart: Kröner 2005², S. 843-844. sowie Fuchs, Michael: Gute wissenschaftliche Praxis. In: Heinemann, Thomas und Bert Heinrichs u. a. (Hrsg.): Forschungsethik. Eine Einführung. Heidelberg: J.B. Metzler 2010, S. 41, zitiert nach Fenner (2022), S. 209.

⁸⁴ Fenner (2022), S. 209.

Verantwortung“ synonym verwendet. Ebenso sind „Kausalhandlungsverantwortung“, „außerwissenschaftliche Verantwortung“ und „externe Verantwortung“ als semantische Einheiten zu begreifen. Interessant ist, dass Fenner interne und externe Verantwortung voneinander separiert, obgleich beide Verantwortungsmodelle meines Erachtens nur gemeinsam betrachtet den Kern realistischer Verantwortungszuschreibung erfassen können.

Dies sei an einem Beispiel näher erläutert: Fenner impliziert, dass die Rolle eines*einer Wissenschaftlers*Wissenschaftlerin im Sinne einer Funktion losgelöst von den durch die in dieser Funktion getroffenen Entscheidungen und Resultaten betrachtet werden kann, da es sich in diesem Fall nicht mehr um interne, sondern um externe Verantwortung handeln würde. Gleichzeitig nennt sie im Kontext der internen Verantwortung jedoch die Verantwortlichkeit von Wissenschaftler*innen gegenüber der gesamten wissenschaftlichen Gemeinschaft.⁸⁵ Dies scheint auf den ersten Blick widersprüchlich, wird von Fenner allerdings folgendermaßen relativiert: „Interne Verantwortung: Rollenverantwortung gegenüber anderen Wissenschaftlern, seiner Zunft oder der ganzen Forschergemeinschaft („Scientific Community“) ethische Gebote: 1) Wahrheit und Wahrhaftigkeit 2) Objektivität und Unvoreingenommenheit 3) intersubjektive Überprüfbarkeit und systematisierter Zweifel 4) Fairness und Gemeinbesitz“⁸⁶.

Interne Verantwortung dürfte demnach in keinem Zusammenhang mit etwaigen Handlungsfolgen stehen, sondern sich lediglich auf die Bedingungen beziehen, die im Sinne verantwortungsvoller wissenschaftlicher Forschung zu erfüllen sind. Innerwissenschaftliche Verantwortung ist ein Konstrukt, dass offenbar weder prospektiv noch retrospektiv zum Tragen kommt. Hierdurch ergibt sich m.E. die Schwierigkeit, dass hierbei zwar der Terminus „Verantwortung“ gebraucht wird, dieser allerdings lediglich eine Art moralischen Kodex für wissenschaftliche Forschung darstellt und nicht im allgemein gebräuchlichen Verständnis tatsächlicher Verantwortung steht. Anders verhält sich dies bei externer Verantwortung, die viel eher das bezeichnet, was gemeinhin mit Verantwortlichkeit kontextualisiert wird.

4.4.2. Kausalhandlungsverantwortung – Externe Verantwortung

Fenner verdeutlicht, dass externe Verantwortung sich im Gegensatz zu interner Verantwortung über die gesamte Gesellschaft bis hin zur Natur und den Folgen für ebendiese

⁸⁵ Vgl. Fenner (2022), S. 210-211.

⁸⁶ Ebd. S. 217.

erstreckt, da hier die Folgen wissenschaftlicher Forschung im Fokus stehen.⁸⁷ An dieser Stelle kann eine Verbindung zwischen den erläuterten Verantwortungskonzepten bei Blöser und Höffe und jenen von Fenner gezogen werden.

Den verantwortlichen Wissenschaftlern wird dann in dem Sinn eine retrospektive Verantwortung zugesprochen, als sie sich für den angerichteten Schaden im Nachhinein bzw. rückblickend („retrospektiv“) rechtfertigen und dafür die Verantwortung übernehmen sollen. Ganz generell kann jemand aber nur deswegen retrospektiv zur Verantwortung gezogen werden, weil er schon zuvor eine prospektive Verantwortung innehatte, d.h. weil ihm bestimmte rollenspezifische Pflichten und Aufgaben zukamen und er für jemanden oder etwas vorausschauend bzw. der Aussicht nach („prospektiv“) zuständig war. Es besteht also eine Korrespondenzbeziehung zwischen prospektiver, Aufgaben- oder Zuständigkeitsverantwortung und retrospektiver, Rechenschafts- oder Zurechnungsverantwortung.⁸⁸

Aus der Verknüpfung dieser unterschiedlichen Modelle ergibt sich eine Problematik, die zwar zu einer philosophisch-ethischen Bearbeitung der Utopie einer Verantwortungszuschreibung in der MMI führt oder geführt hat, die jedoch keine umsetzbaren Lösungsmöglichkeiten hervorbringt. Wurde schließlich das große Konzept „Verantwortung“ zum Ziel der besseren Greifbarkeit von unterschiedlichen Instanzen in mehrere Einheiten gegliedert, so scheint genau dies bei dem Versuch einer ganzheitlichen Betrachtung des Verantwortungs-Dilemmas für Schwierigkeiten zu sorgen. Eine Differenzierung von Verantwortung kommt offenbar in erster Linie jenen Instanzen zugute, die sich gegen das Tragen ebendieser auf einer oder mehreren Ebenen sträuben.

Ein*e Wissenschaftler*in beispielsweise, die ihre Verantwortung lediglich intern begreift, profitiert bei auftretenden Problemen und Risiken von der eigens getätigten Verantwortungszuschreibung. Im Sinne einer Rollenverantwortung habe er*sie demnach im Sinne der Wissenschaft gehandelt und die ethischen Bedingungen hierfür erfüllt. Verantwortung für die Folgen der eigenen Handlungen kann demnach durch einen Verweis auf spezifische Verantwortungskonzepte abgelehnt werden. Dies ist jedoch höchst problematisch, schließlich „kann die Verantwortung für unmittelbare Folgen der empirisch arbeitenden Wissenschaften, die sich im Forschungsprozess selbst in der direkten Interaktion mit Mit- und Umwelt ergeben“⁸⁹ wohl kaum geleugnet werden.

Möglicherweise hilft an dieser Stelle eine genauere Betrachtung unterschiedlicher Bedingungen weiter, die die Zuschreibung von Verantwortung überhaupt erst ermöglichen.

⁸⁷ Vgl. Fenner (2022), S. 218.

⁸⁸ Ebd. S. 218-219.

⁸⁹ Ebd. S. 220.

5. Problematisierung

5.1. Bedingungen der Verantwortungszuschreibung

Heidbrink hebt in Anlehnung an Immanuel Kant Zurechnungsfähigkeit als zentrale Bedingung für Verantwortlichkeit hervor.⁹⁰ Auch bei Thomas von Aquin und Aristoteles finden sich Richtlinien, die in der Frage nach Zurechenbarkeit und Zurechnungsfähigkeit relevant sind:

Hierzu zählt die Freiwilligkeit, denn nur derjenige, der ohne Nötigung handelt, kann für seine Taten zur Rechenschaft gezogen werden. Neben der Freiwilligkeit (hekou'sion) gehören auch Wissentlichkeit und Willentlichkeit zu den Bedingungen der Zurechenbarkeit von Handlungen. Nur wer unter der Kenntnis absehbarer Umstände und aus eigenem Entschluss (prohaíresis) handelt, lässt sich für sein Tun verantwortlich machen. Dabei ist es dem Handelnden nicht gestattet, sich auf sein Unwissen herauszureden. Er besitzt vielmehr die Sorgfaltspflicht, sich über die genauen Umstände seines Handelns zu informieren.⁹¹

Freiwilligkeit und Wissentlichkeit scheinen auf den ersten Blick Kriterien zu sein, die im wissenschaftlichen Kontext zwangsläufig gegeben sein sollten. Dies kommt allerdings nur dann zum Tragen, wenn es sich bei den Handelnden tatsächlich um reale Personen handelt und nicht um Maschinen oder Roboter, die mit realen Personen interagieren. In diesem Fall würde und wird die Frage der Freiwilligkeit und der Wissentlichkeit eine sehr differenzierte Betrachtung erfordern und die These, dass bei der Zuschreibung von Verantwortung in der MMI eine Utopie vorliegt, bestätigen. Diese Problematik wird später ausführlich und strukturiert abgehandelt, muss der Vollständigkeit halber jedoch bereits hier erwähnt und angemerkt werden. In Anbetracht der Tatsache, dass Unwissen gemeinhin als Kriterium für den Versuch einer Abwehr von Verantwortung gebraucht wird, ist zu überlegen, ob dies legitim ist oder nicht. Aristoteles erinnert in diesem Zusammenhang daran, dass auch ein Mensch Verantwortung für sein Handeln zu tragen hat, „der aus Willensschwäche (akrasia) gegen die rechtmäßige Ordnung verstößt und es trotz besserer Einsicht unterlässt, auf den Pfad der Tugend (areté) zurückzukehren.“⁹²

Alltagssprachliche Aussagen wie „Ich wusste es nicht besser“, „Ich dachte, es ist nicht so schlimm“ oder „Ich glaubte wirklich, da kann nichts passieren“, rechtfertigen demnach in keiner Weise unverantwortliches Handeln, selbst wenn der Inhalt der Aussagen auf subjektiver Ebene korrekt ist. Eine Ablehnung von Verantwortlichkeit lässt sich hierbei nicht nur durch Aristoteles Bewertung von Unwissenheit und Willensschwäche zurückweisen,

⁹⁰ Vgl. Heidbrink (2017), S. 8.

⁹¹ Ebd. S. 9.

⁹² Vgl. Bien, Günther (Hrsg.): Aristoteles. Nikomachische Ethik. Hamburg: Felix Meiner⁴ 1985, S. 44–58, zitiert nach Heidbrink (2017), S. 9.

sondern kann auch durch die zuvor genannte zentrale Bedingung der Zurechnungsfähigkeit gestützt werden. Dies ist unter anderem deshalb interessant, weil eine Person ab jenem Moment, indem sie Person als zurechnungsfähig bezeichnet wird, das Privileg der Rechtfertigung und Entschuldbarkeit durch Unwissenheit verliert. Genau dieses Privileg kommt jedoch nicht-menschlichen Instanzen automatisch zu, wenngleich fragwürdig ist, ob diese tatsächlich weniger dazu in der Lage sind, Handlungsfolgen abzusehen als Wissenschaftler*innen, die durch ihre Arbeit zwar vorhersehbare Folgen erwarten, diese jedoch nicht immer ohne Weiteres in ihrem vollen Ausmaß begreifen können. So verhält es sich m.E. nicht nur bei Wissenschaftler*innen, sondern bei allen Menschen, die nach bestem Wissen und Gewissen handeln und dennoch manchmal Fehler begehen, für die sie zur Verantwortung gezogen werden, während der Fehler eines artifiziellen Systems im Zweifelsfall vermutlich häufig in einem diffusen Feld aus fragwürdiger Verantwortungsdiffusion untergehen würde.

Um ein Zwischenfazit zu ziehen: Die Zuschreibung von Verantwortung hat die Bedingungen der Freiheit, Kausalität und Intentionalität zur Voraussetzung. Die Kriterien und Maßstäbe, die für die Zuschreibung von Verantwortung erforderlich sind, hängen in der Praxis von den konkreten Umständen und Rahmenbedingungen ab, unter denen Akteure ihre Handlungen vollziehen. Vorsatz und Absicht, Fähigkeiten und Kenntnisse, Aufgaben und Rollen, das Arbeitsumfeld und gesetzliche Regelungen sind Faktoren, die bei der Verantwortungsattribution berücksichtigt werden müssen.⁹³

Heidbrink räumt an dieser Stelle die Relevanz einer Beachtung der äußeren Umstände ein. Genau an diesem Punkt ist es möglich, die im Laufe dieses Kapitels gestellten Fragen nun auf die Maschinen- und Technikethik und die damit verbundene MMI zu beziehen, da hierdurch ein Vergleich gewährleistet wird, der die Utopie der Verantwortungszuschreibung im technischen Bereich sichtbar macht und aufzeigt, dass das Delegieren von Verantwortung in der zwischenmenschlichen Interaktion strukturalistisch nachvollziehbar ist, während sich in Bezug auf artifizielle Systeme gravierende Probleme ergeben.

5.2. Das Problem der Verantwortungszuschreibung in der Maschinen- und Technikethik

Fenner beschreibt Verantwortung als das zentrale Prinzip der Technikethik, legt aber zugleich dar, weshalb dieses Prinzip zum derzeitigen Stand nicht ertragreich oder gar funktional ist. So sei ein Prinzip, das genutzt werden müsse, um Verantwortung auf artifizielle Systeme übertragen zu können, momentan ein „[...] inhaltsleeres Prinzip ohne normative Standards/Wertmaßstäbe“⁹⁴. Hierfür gibt es mehrere Gründe. Zunächst ist es

⁹³ Heidbrink (2017), S. 10.

⁹⁴ Fenner (2022), S. 262.

relevant, zu begreifen, dass hinsichtlich der Verantwortungszuschreibung in der Maschinen-, Roboter- und Technikethik bei weitem striktere Kriterien benötigt werden als derzeit überhaupt verfügbar sind. „In einschlägiger Fachliteratur wird es vermieden, im Zusammenhang von Künstlicher Intelligenz von „Handlungen/Handlungsgründen“ und „Entscheidungen“ zu sprechen. Stattdessen wird von „Operationen“ bzw. „Ursachen“ gesprochen.“⁹⁵ Hierdurch wird auf subtile Weise die Frage nach der Verantwortungszuschreibung allerdings nicht einfacher, sondern noch komplexer. Durch die Ablehnung von Begrifflichkeiten, die Maschinen und Roboter im Sinne ihrer Handlungsfähigkeit mit menschlichen Instanzen auf eine Ebene stellen, liegt die Frage nahe, ob technische artifizielle Systeme überhaupt zu einer wie auch immer gearteten Verantwortungsübernahme fähig sind und wenn ja, wie über die Zurechnungsfähigkeit dieser Systeme entschieden wird.

Josifović postuliert, dass im Bereich der Technikethik ein „Accountability gap“⁹⁶ besteht, der insbesondere durch die autonome Lernfähigkeit bestimmter KI-gestützter Systeme hervorgerufen wird. Selbstlernende Algorithmen beispielsweise werden mit einer riesigen Menge an Daten trainiert, die es ihnen möglich macht, basierend auf ebendiesen Inhalten mittels eines menschlichen Auftrags autonom zu operieren.⁹⁷ Wenngleich dies für eine Reihe an Operationen praktikabel erscheint, ergeben sich hieraus auch Lücken in der Rückverfolgbarkeit getroffener Entscheidungen.

Somit tritt beim Einsatz selbstlernender Algorithmen ein verantwortungstheoretisch relevantes Problem auf, das darin besteht, dass die menschlichen Akteure, die sie ursprünglich programmiert haben, und solche, in deren Auftrag sie operieren, nicht vorhersagen können, nach welchem Rationale [sic!] ein Algorithmus neue Klassen generieren und operieren, also zukünftige „Entscheidungen“ treffen wird. Mehr noch: Oftmals ist es selbst im Nachhinein nicht möglich, dies nachzuvollziehen. Diese Unmöglichkeit ergibt sich nicht infolge irgendeiner Nachlässigkeit bei der Programmierung oder infolge irgendwelcher Verantwortungslosigkeit beim Einsatz Künstlicher Intelligenz, sondern schlicht aus der Tatsache, dass Algorithmen mit gewaltigen, chaotischen Datenmengen operieren, die sie eigenständig strukturieren.⁹⁸

Schwierigkeiten ergeben sich nun nicht in dem Moment der Entscheidungsfindung oder einer Vorbereitung ebendieser, sondern entstehen im Zweifelsfall erst dann, wenn getroffene Entscheidungen bzw. „Operationen“ nicht die gewünschten Ergebnisse liefern oder gar Schaden verursachen. Wird nun danach gefragt, welche Instanzen für diese Fehler und

⁹⁵ Josifović, Saša: Denn wir wissen nicht, was sie tun. Das Problem der Verantwortungszuschreibung beim Einsatz selbstlernender Algorithmen. In: Philosophisches Jahrbuch (Freiburg) 127(1) 2020, S. 48.

⁹⁶ Josifović (2020), S. 50.

⁹⁷ Vgl. Ebd. S. 49.

⁹⁸ Ebd. S. 50.

Schäden zur Verantwortung gezogen werden müssen, ergeben sich mehrere Möglichkeiten, die – wie sich zeigen wird – unbefriedigend und unzureichend sind.

Zunächst muss hierfür die Neutralitätsthese der Technik betrachtet werden. Diese besagt folgendes:

Technische Geräte und Verfahren (Technik (1)) sind ethisch neutral. These 2: Alle Handlungen des Entwickelns und Herstellens technischer Artefakte (Technik (2)) sind ethisch neutral. These 3: Das Wissen um Herstellung, Gebrauch und Reparatur technischer Artefakte (Technik (3)) ist im Unterschied zur Anwendung dieses Wissens ethisch neutral.⁹⁹

Bis zum Ende des 20. Jahrhunderts war die Annahme weit verbreitet, Technik sei als solche wertneutral.¹⁰⁰ Dies mag unter anderem daran liegen, dass Technik als Hyperonym für vielerlei technische Errungenschaften verstanden werden kann. Wird Technik als Prozess der Herstellung technischer Gerätschaften begriffen, ist es durchaus nachvollziehbar, diesen Prozess als neutral zu bezeichnen.¹⁰¹ Gilt Technik allerdings als das Ergebnis eines Herstellungsprozesses, so impliziert die Neutralitätsthese, auch die Produkte technischer Verfahren seien neutral. Dies ist kritisch zu reflektieren. Julliard postuliert, technische Geräte seien als solche nicht ethisch gut oder schlecht, allerdings ist hierbei durchaus zu beachten, dass Maschinen sowohl zu guten als auch zu ethisch fragwürdigen und üblen Zwecken gebraucht werden können.¹⁰² Hieraus ergibt sich wiederum die Frage, ob ein Diskurs, der die Neutralitätsthese zu bekräftigen oder zu widerlegen sucht, überhaupt relevant ist oder ob es vielmehr um die Frage nach den Folgen vermeintlicher Neutralität und hiermit verbunden – um Verantwortlichkeit – geht. Technik oder technische Gerätschaften als Verantwortungstragende zu identifizieren, scheint angesichts ihrer Neutralität nicht zielführend oder realistisch und ist demnach ein unbefriedigender Ansatz.

Richtet sich der Fokus auf diejenigen Instanzen, die an der Herstellung technischer Gerätschaften beteiligt sind, so liegt folgendes Problem vor: Sowohl Ingenieur*innen als auch Techniker*innen folgen bei der Produktion technischer Systeme lediglich den Anforderungen ihres Berufs. Sie deshalb für mögliche Folgen ihrer Arbeit verantwortlich zu machen, scheint daher nicht gerechtfertigt, noch dazu, weil viele Maschinen entgegen ihrem vorgesehenen Zweck verwendet werden können. In diesem Fall erschiene es absurd, die Hersteller*innen für wie auch immer gearteten Missbrauch verantwortlich zu machen. Zu

⁹⁹ Fenner (2022), S. 271.

¹⁰⁰ Vgl. Grunwald, Armin: Überblick über die Technikethik. In: Grunwald, Armin und Rafaela Hillerbrand (Hrsg.): Handbuch Technikethik. Stuttgart/Weimar: J.B. Metzler 2021², S. 3.

¹⁰¹ Vgl. Fenner (2022), S. 270.

¹⁰² Vgl. Julliard, Yannick: Ethische Technikgestaltung. Technikethik aus Sicht eines Ingenieurs. Frankfurt am Main: Peter Lang 2003, S. 67-68, zitiert nach Fenner (2022), S. 270-271.

behaupten, die Verantwortung für etwaige Risiken und Fehler läge nur bei Nutzer*innen, wirkt allerdings genauso wenig plausibel, schließlich wissen Hersteller*innen um bestimmte Nutzungsrisiken und sind zumindest dafür verantwortlich, potenzielle Käufer*innen über diese zu informieren.¹⁰³ Der technologische Determinismus hat letztlich dazu geführt, dass bis zur zweiten Hälfte des 20. Jahrhunderts die Entstehung einer Technikethik, die sich mit obig dargelegten Problemen befasst, behindert wurde, indem die These vertreten wurde, dass technische Entwicklungen einer nicht beeinflussbaren Eigendynamik unterlägen.¹⁰⁴ Hierdurch wurde gewissermaßen das Problem der Verantwortungszuschreibung ausgelagert und an eine nicht definierbare oder greifbare Instanz abgegeben. Dies mag in Hinblick auf nicht KI-gestützte Technik funktioniert haben, ist jedoch eine Sichtweise, die spätestens bei der Betrachtung artifizieller [selbstlernender] Systeme an ihre Grenzen gerät. Unter anderem aus diesem Grund wurden in der Technikethik zwei Verantwortungskonzepte ausgearbeitet, die auch im Kontext der Mensch-Maschine-Interaktion zur Anwendung kommen können.

5.3. Herstellungsorientierte vs. gebrauchorientierte Verantwortung

Im Sinne der Notwendigkeit spezifischer Verantwortungskonzepte in der Technikethik erinnert Skorupinski daran, dass sämtliche Entwicklungsprozesse technischer Systeme und Gerätschaften hochkomplex und multidimensional sind.¹⁰⁵ Im Falle von auftretenden Schwierigkeiten oder negativen Folgen ist es daher nicht direkt möglich, eine*n Ingenieur*in oder Produzent*in zur Verantwortung zu ziehen, schließlich sind an der Technikherstellung teilweise tausende Menschen beteiligt, die im Rahmen unterschiedlicher Entwicklungsstadien sehr spezifische und differenzierte Aufgaben übernehmen.¹⁰⁶ Das herstellungsorientierte Verantwortungskonzept sieht dennoch die Konstrukteur*innen etwaiger Technik als jene Instanzen an, die im Falle von Schäden und negativen Folgen planmäßiger Verwendung Verantwortung zu tragen haben.¹⁰⁷

Ropohl erwähnt in diesem Kontext die Adäquatheit des Modells in Anbetracht der Entwicklung ethisch höchst problematischer Gegenstände. Wird beispielsweise ein Gerät bewusst zum Zwecke der Folter entwickelt, obwohl die Hersteller*innen ein ausgeprägtes

¹⁰³ Vgl. Fenner (2022), S. 271.

¹⁰⁴ Vgl. dazu Grunwald, Armin: Technikethik. In: Düwell, Marcus und Christoph Hübenthal u.a. (Hrsg.): Handbuch Ethik. Stuttgart/Weimar: J.B. Metzler 2011³, S. 281. sowie Skorupinski, Barbara: Technikethik – Technikfolgenabschätzung. In: Maring, Matthias (Hrsg.): Ethisch-Philosophisches Grundlagenstudium 2. Münster: LIT 2005, S. 181-182, zitiert nach Fenner (2022), S. 277.

¹⁰⁵ Vgl. Skorupinski (2005), S. 182.

¹⁰⁶ Vgl. Fenner (2022), S. 278-279.

¹⁰⁷ Vgl. Ebd. S. 278.

Bewusstsein über die Gefahren und Folgen besitzen, so ist das herstellungsorientierte Verantwortungskonzept in jedem Fall nachvollziehbar und vertretbar.¹⁰⁸ Im Falle der Produktion ethisch unproblematischer Gerätschaften erscheint das Konzept jedoch lückenhaft.

Häufig werden Ingenieure auch noch für die richtige Inbetriebnahme, die adäquate Nutzung und die sachgerechte Entsorgung verantwortlich gemacht (vgl. Hubig u. a. 28): Bei multifunktionalen technischen Geräten oder bei Geräten mit vermeidbaren Nebenwirkungen sollten sie die Nutzer über die möglichen Schäden und Gefahren bei der Verwendung aufklären und allfällige angezeigte Maßnahmen zur Vermeidung eines Fehlgebrauchs einleiten [...] ¹⁰⁹

Das gebrauchtorientierte Verantwortungskonzept verlagert wiederum die Verantwortung nur auf die Nutzer*innen, die mittels ihrer Entscheidung des Kaufs bestimmter Gerätschaften selbst für dessen Gebrauch und damit verbundene Folgen haften.¹¹⁰ Weder das herstellungs- noch das gebrauchtorientierte Verantwortungskonzept wirken tatsächlich anwendbar. Als nächstes sollen daher die Konzepte moralischer Akteurschaft beleuchtet werden, die in der Roboter- und Maschinenethik eine zentrale Rolle spielen.

6. Arbeitsfelder der Roboterethik

6.1. Roboter als moralische Handlungssubjekte

6.1.1. Definition moralischer Akteursschaft (moral agency)

Im Rahmen der Maschinen- Roboter- und Technikethik scheint in der Literatur der Begriff der *moral agency* vermehrt auf und kann sinngemäß als moralische Akteursschaft oder Handlungsfähigkeit übersetzt werden.¹¹¹ Was auf den ersten Blick in facheinschlägiger Literatur wie ein abgerundetes Konzept dargestellt wird, beinhaltet bei genauerer Betrachtung einige Aspekte, die die gezielte Auseinandersetzung mit der Verantwortungsdelegation an Maschinen und Roboter nicht zwangsläufig erleichtern, sondern teils sogar erschweren. Unter anderem liegt dies daran, dass die Terminologie von *agency* als solche bereits sehr voraussetzungsreich ist. Von unterschiedlichen Verfasser*innen werden zahlreiche Konzepte zur Erläuterung moralischer Akteursschaft vorgestellt, die ihrerseits eine Reihe an Bedingungen implizieren. Diese Bedingungen wiederum sind – wie im Laufe dieses Kapitels aufgezeigt wird, sehr unterschiedlich und überdies nicht ausschließlich auf artifizielle Systeme oder aber menschliche Handelnde

¹⁰⁸ Vgl. Ropohl, Günter: Neue Wege, die Technik zu verantworten. In: Lenk, Hans und Günter Ropohl (Hrsg.): Technik und Ethik. Stuttgart: Reclam 1993², S. 160.

¹⁰⁹ Fenner (2022), S. 279.

¹¹⁰ Vgl. Ebd. S. 279.

¹¹¹ Loh (2020), S. 138.

beschränkt, sondern können gleichermaßen als hybrid anwendbares Gebilde gelesen werden. Im Sinne einer strukturierten Betrachtung dessen, was fachlich als *moral agency* gilt, sei daher zunächst das Grundgerüst von Akteursschaft betrachtet. Hierdurch wird nicht nur gewährleistet, dass diverse Formen von *agency* nicht als Synonyme verstanden und genutzt werden, sondern insbesondere, dass deren semantische Unterschiede auch in Hinblick einer Auseinandersetzung mit Verantwortungszuschreibung etwaiger künstlicher, natürlicher oder moralischer Akteur*innen beachtet werden. Während bei einigen Wissenschaftler*innen Unterscheidungen im Rahmen einer Definition moralischer Akteursschaft keine Erwähnung finden, fokussiert Himma genau auf diese terminologische Lücke. Er geht intensiv der Frage nach, ob die realistische Möglichkeit besteht, künstliche Akteur*innen gleichsam als moralische Akteur*innen zu verstehen.¹¹² Hierfür differenziert er die unterschiedlichen Arten von Akteursschaft und grenzt sie weiters von natürlicher Akteursschaft ab:

One can distinguish natural agents from artificial agents. Some agents are natural in the sense that their existence can be explained by biological considerations; people and dogs are natural agents insofar as they exist in consequence of biological reproductive capacities – and are hence biologically alive. Some agents might be artificial in the sense that they are manufactured by intentional agents out of pre-existing materials external to the manufacturers; such agents are artifacts. Highly sophisticated computers might be artificial agents; they are clearly artificial and would be artificial agents if they satisfy the criteria for agency – in particular, if they are capable of instantiating intentional states that cause actions.¹¹³

Himma erinnert dennoch daran, dass die Trennschärfe zwischen den unterschiedlichen Termini nicht immer vollständig gegeben ist. So gäbe es beispielsweise durchaus Akteur*innen – zum Beispiel Klone – die unter bestimmten Voraussetzungen gleichsam natürlich und künstlich sein können.¹¹⁴ Bevor nun die Frage nach der Bedeutung moralischer Akteursschaft beantwortet werden kann, ist es relevant, nicht nur die Formen ebendieser zu kennen, sondern auch den semantischen Wortgehalt zu begreifen. „The idea of agency is conceptually associated with the idea of being capable of doing something that counts as an act or action. As a conceptual matter, X is an agent if and only if X is capable of performing actions.“¹¹⁵

Eine ähnliche Vorstellung vertritt auch Black, die hiermit jedoch nicht Akteursschaft als Grundgerüst meint, sondern postuliert, **moralische** Akteursschaft beziehe sich grundlegend auf die Fähigkeit eines Individuums, über das eigene Verhalten zu entscheiden und zu

¹¹² Vgl. Himma, Kenneth Einar: 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 (2009), S. 19.

¹¹³ Himma (2009), S. 19.

¹¹⁴ Vgl. Ebd. S. 21.

¹¹⁵ Ebd. S. 19.

begreifen, dass dieses eine Auswirkung auf das Befinden anderer Individuen hat. Mehr noch geht Black davon aus, die Fähigkeit der Schadensvermeidung sei hier bereits inbegriffen.¹¹⁶ Diese Auffassung impliziert allerdings, dass *moral agency* Verantwortung sowohl voraussetzt als auch notwendigerweise die Möglichkeit einer Zuschreibung ebendieser für realistisch hält. Diese Sichtweise wird in der Fachliteratur nur teils unterstützt. So wird moralische Akteursschaft beispielsweise als Vorbote für moralische Verantwortung begriffen, jedoch nicht damit gleichgesetzt oder als miteinander konnotiert verstanden.¹¹⁷

Himmas These dürfte jener von Black als komplementär entgegengestellt werden. Himma verneint zwar keineswegs, dass bestimmte Voraussetzungen notwendige Bedingung für die Zuschreibung moralischer Akteursschaft sind, allerdings nimmt er Abstand davon, *agency* und *moral agency* an dieselben Bedingungen zu knüpfen. Gerade die Fähigkeit, moralische Maßstäbe zu erkennen, zu begreifen und umzusetzen, beziehungsweise auf unterschiedliche Situationen anzuwenden, scheint nicht notwendigerweise eine Voraussetzung für moralische Akteursschaft zu sein, sondern vielmehr ein Unterscheidungsmerkmal zu Akteurschaft als solcher. Was Handlungssubjekte von moralischen Handlungssubjekten unterscheidet, ist demnach in erster Linie, dass moralisch Handelnde über moralische Verpflichtungen und Richtlinien verfügen, während dies bei Handelnden anderer Art nicht der Fall ist.¹¹⁸

Eine Definition von moralischer Handlungsfähigkeit findet sich häufig nur im Rahmen einer Aufzählung von Bedingungen für ebendiese, ohne jedoch die Begrifflichkeit selbst konkret ab- und einzugrenzen. Blacks Versuch, diese Problematik zu umgehen, ergibt folgenden Definitionsvorschlag: “Moral agency is the ability both to refrain from acting immorally or inhumanely, regardless of situational pressures, and conversely to engage in prosocial behavior.”¹¹⁹ *Moral agency*, verstanden als Fähigkeit, problematisches Verhalten zu vermeiden und auch in emotional fordernden Situationen oder solchen, die eine*n menschliche*n Akteur*in massiv unter Druck setzen würden, moralisch richtig und sozial zu agieren, scheint interessanterweise gleichermaßen realistisch wie auch utopisch. Realistisch wirkt diese Anforderung in Bezug auf jene Akteur*innen, die aufgrund ihrer

¹¹⁶ Vgl. Black, Jessica E.: An Introduction to the Moral Agency Scale. Individual Differences in Moral Agency and Their Relationship to Related Moral Constructs, Free Will, and Blame Attribution. In: *Social psychology* 47(6) 2016, S. 295.

¹¹⁷ Vgl. Malle, Bertram F. und Guglielmo, Steve u.a.: Moral, cognitive, and social: The nature of blame. In: Forgas, Joseph und Klaus Fiedler u.a. (Hrsg.): *Social thinking and interpersonal behaviour*. Philadelphia, PA: Psychology Press 2012, S. 311–329. sowie Pasupathi, Monisha und Cecilia Wainryb: Developing moral agency through narrative. In: *Human Development* 53, 2010, S. 55–80.

¹¹⁸ Vgl. Himma (2009), S. 21.

¹¹⁹ Black (2016), S. 296.

Beschaffenheit als Roboter oder Maschinen, der ihnen implementierten Algorithmen sowie deren fehlender Fähigkeit, menschliche Emotionen zu empfinden, auch in hochriskanten oder gefährlichen Situationen ruhig zu bleiben und so zu handeln, wie es gemeinhin als “prosocial” angesehen würde. In Bezug auf menschliche Handelnde hingegen wirkt Blacks Beschreibung moralischer Akteursschaft unrealistisch bis unmöglich, da anzunehmen ist, dass Menschen als Träger*innen unterschiedlichster emotionaler Zustände nicht immer und vor allem nicht unabhängig von den Herausforderungen bestimmter Situationen zu moralisch und gesellschaftlich anerkanntem Verhalten in der Lage sind. Während bei Black keine detailliertere Auseinandersetzung mit der Problematik moralischer Fehlbarkeit stattfindet, ergibt sich bei Himma eine engmaschige Betrachtung, die einem Verständnis moralischer Handlungsfähigkeit in vielerlei Hinsicht zuträglich ist.

Himma stimmt zu, dass Akteur*innen, die den Unterschied zwischen richtig und falsch nicht begreifen, nicht als moralische Handlungssubjekte deklariert werden können und ihnen somit auch keine Verantwortung für das von ihnen ausgeführte Verhalten zugeschrieben werden kann. Die relevante Differenzierung zu Blacks Auffassung von *moral agency* liegt allerdings darin, dass Himma die Fähigkeit, die Unterschiede zwischen richtig und falsch zu **kennen** als Bedingung für moralische Akteursschaft voraussetzt, während Black die fehlerfreie Umsetzung dieser Bedingung als notwendig erachtet.¹²⁰ Himma gesteht auch moralischen Handlungssubjekten die Wahrscheinlichkeit einer bestimmten Fehlerquote ein. Sie postuliert, dass nicht nur manche, sondern nahezu alle erwachsenen zurechnungsfähigen menschlichen Individuen in bestimmten Situationen „notwithstanding their own views to the contrary“¹²¹ nicht immer wissen, welche Handlungen als richtig und welche als falsch zu klassifizieren sind.

An dieser Stelle könnte nun wiederum in Anbetracht der Definition von *moral agency* Blacks, das Konzept des freien Willens Anwendung finden. Die Annahme, *moral agency* basiere auf der Fähigkeit, durch das Vorhandensein eines freien Willens [moralisch richtige] Entscheidungen zu treffen und diese gezielt in ein Verhalten umzusetzen, das vorsätzlich ein bestimmtes Ergebnis nach sich zieht, liegt nahe. Black dementiert dies allerdings, indem sie darauf verweist, Akteursschaft stehe zwar in einem Verhältnis zu dem Konzept des freien Willens, sei jedoch nicht hiermit gleichzusetzen. Den Unterschied zwischen „free will“ und *agency* begreift sie wie folgt: “Free will refers to the ability to choose between two or more

¹²⁰ Vgl. Himma (2009), S. 23.

¹²¹ Ebd. S. 23.

actions without any limitations or constraints (see Libet, 1999; O'Connor, 2014). Agency refers to the ability to intentionally act in order to cause a given outcome (see Bandura, 2006; Malle et al., 2012).”¹²²

Das Wort “intentionally” verweist hierbei auf eine wichtige Funktion. Ein*e Akteur*in, der*die als moralisches Handlungssubjekt gilt, dürfte sich demnach von jenen Akteur*innen, die als moralisches Handlungsobjekt (siehe Kap. 6.3.) gelten, durch die Möglichkeit unterscheiden, bewusst, gezielt und absichtlich eine Handlung oder Reaktion einer anderen vorzuziehen. Dies impliziert nicht nur ein komplexes und funktionales Denkvermögen, sondern auch die hiermit verbundene Möglichkeit der Verantwortungszuschreibung. Parthemore und Whitby verstehen diese als Bedingung für die Bezeichnung eines*einer wie auch immer beschaffenen Handelnden als *moral agent*.

6.1.2. Bedingungen für moralische Handlungssubjekte

Wie bereits im Rahmen des Versuchs einer Begriffsdefinition von *moral agency* sichtbar geworden ist, werden die Bedingungen, welche erfüllt sein müssen, um eine*n Handelnde*n als moralisches Handlungssubjekt deklarieren zu können, in der gängigen Fachliteratur divergent dargestellt, was sich letztendlich durch das Definieren moralischer Handlungssubjekte **über** Bedingungen und Fähigkeiten ergibt. Im folgenden Abschnitt werden unterschiedliche Perspektiven vorgestellt, die ihrerseits auf differenzierte Weise versuchen, zu erläutern, was ein moralisches Handlungssubjekt von einem solchen Handlungsobjekt unterscheidet und inwiefern bestimmte Fähigkeiten natürlicher oder artifizieller Handelnder diese Differenzierung rechtfertigen. Während unter anderem bei Himma und Black der Fokus auf teils sehr komplexen Bedingungen für *moral agency* liegt, kommt bei Parthemore und Whitby eine vermeintlich fundamentale Forderung hinzu: “We take it as granted that that agent [Anm.: a moral agent] must be embedded or situated in a particular spatiotemporal context and embodied in a particular physical form: one that shapes and constrains its interactions with its environment.”¹²³

Es wird also postuliert, dass das Gebundensein eines wie auch immer beschaffenen Intellekts an eine körperliche Form notwendige Bedingung für die Möglichkeit einer Verantwortungszuschreibung und damit auch der gewissenhaften Bezeichnung eines

¹²² Himma (2009), S. 296.

¹²³ Parthemore, Joel und Blay Whitby: What makes any agent a moral agent? Reflections on machine consciousness and moral agency. In: International Journal of Machine Consciousness 5(2) 2013, S. 1.

Systems als moralische*r Akteur*in sei. Interessanterweise erleichtert dies jedoch die Auseinandersetzung mit der Zuordnung artifizierter Systeme als moralische Handlungssubjekte nicht, sondern erschwert sie sogar. Wenngleich verständlich ist, dass im Sinne einer sprachlich und semantisch korrekten Bezeichnung nur der*diejenige handeln kann, der*die auch über eine physische Form verfügt, ist es dennoch fraglich, ob diese Anforderung an artifizelle Akteur*innen nicht das ursächliche Problem – nämlich die Fähigkeit, Entscheidungen zu treffen und diese teils auch ohne Vorhandensein einer körperlichen Erscheinung umzusetzen – außer Acht lässt.

Bei Karpov wird diese Problematik umgangen, indem nicht der Körper als Bedingung für die Bedeutsamkeit eines Akteurs*einer Akteurin und dessen*deren Handlungen betrachtet wird, sondern vielmehr das zugrunde liegende Gefühl für das eigene Selbst. So sieht Karpov das Erleben des eigenen Selbst im Sinn eines Ich-Verständnisses als zentral an. Dieses Ich müsse in einem geordneten Verhältnis zu der es umgebenden Welt stehen und über die Fähigkeit verfügen, sich selbst mit anderen Handelnden zu vergleichen.¹²⁴ Genauer erläutert wird diese These mittels eines detaillierten Modells, welches fünf Charakteristika benennt und beschreibt, deren Gegebenheit eine*n Akteur*in zu einem moralischen Handlungssubjekt macht. Hierbei handelt es sich um „Language“, „Motive“, „Feelings“, „Sympathy“ und „Responsibility“.¹²⁵ Sprache als Erfordernis für Moral impliziert, dass nur Menschen und gegebenenfalls artifizelle Systeme, Maschinen und Roboter überhaupt hinreichende Qualifikationen als moralische Handlungssubjekte aufweisen, während dies bei Tieren nicht der Fall ist, selbst wenn diese über eine eigene Art der Kommunikation verfügen.¹²⁶ Weiters kommt dem Vorhandensein eines oder mehrerer Motive, das/die ein bestimmtes Verhalten hervorrufen, Bedeutung zu.¹²⁷

Ein weiteres Charakteristikum, welches Karpov für moralische Handlungssubjekte als zentral erachtet, sind Gefühle, wobei in diesem Kontext folgendes angemerkt wird: „If, by feelings, we understand a certain process that reflects a subjective evaluative attitude to some objects, then those processes that occur in an emotional-needs architecture can rightfully be called feelings. A robot can really feel.“¹²⁸ Diese Darstellung ist durchaus interessant und ergibt im Rahmen einer Simplifizierung des Gefühlsbegriffs eine realistische

¹²⁴ Vgl. Karpov, Valery E.: Can a robot be a moral patient? In: Kuznetsov, Sergei O. und Aleksandr I. Panov u.a. (Hrsg.): Artificial Intelligence. Springer Nature Switzerland AG 2020, S. 63.

¹²⁵ Karpov (2020), S. 67-68.

¹²⁶ Vgl. Ebd. S. 67.

¹²⁷ Ebd. S. 67.

¹²⁸ Ebd. S. 68.

Betrachtungsmöglichkeit. Wenngleich bei Himma der Terminus „Gefühle“ (engl. „feelings“) nicht aufscheint, so wird dort allerdings häufig die Relevanz von „intentional states (i.e., mental states that are about something else, like a desire for X)“¹²⁹ erläutert. Gefühle als solche dürften zwar eher notwendige als hinreichende Voraussetzung für das Vorhandensein bestimmter emotionaler Zustände, Wünsche und Sehnsüchte sein, können jedoch im Sinne eines Vergleichs unterschiedlicher Bedingungen für *moral agency* bei unterschiedlichen Wissenschaftler*innen durchaus als Parallele angesehen werden.

Ähnlich verhält es sich mit der Fähigkeit artifizieller Systeme zu der Empfindung von Sympathie, die nach Karpov ein weiteres Merkmal moralischer Handlungssubjekte darstellt und ihrerseits ein ausgeprägtes Reflexionsvermögen voraussetzt. Die Fähigkeit, zwischen dem eigenen empfundenen oder wahrgenommenen Selbstkonzept und jenem des Gegenübers zu unterscheiden und basierend auf den erkannten Gemeinsamkeiten oder Unterschieden Sympathie oder Antipathie zu empfinden, dürfte demnach ein Ergebnis dessen sein, was gemeinhin als Denkfähigkeit bezeichnet werden würde.

Weitere Ansätze beinhalten eine Kombination dieser oder ähnlicher Aspekte als Kernmerkmale moralischer Akteursschaft, während manche von Karpov postulierten Charakteristika unter Ausnahme des Selbst-Konzepts bzw. Ich-Gefühls und der Relevanz von Verantwortung keine oder lediglich eine untergeordnete Rolle spielen. Hierauf wird im Folgenden näher eingegangen. Zunächst sei jedoch das fünfte und letzte Merkmal erwähnt, das Karpov als unabdingbar für *moral agency* hält – Die Fähigkeit, Verantwortung für das eigenen Tun und Handeln zu übernehmen.¹³⁰ Genau an dieser Stelle verweise ich erneut darauf, dass, wie bereits dargelegt, Verantwortungszuschreibung im Kontext nicht-menschlicher Instanzen höchst problematisch ist und nicht ohne Weiteres als Merkmal moralischer Akteursschaft angenommen werden kann, solange nicht klar definiert ist, von welcher Form der Verantwortung ausgegangen wird und ob diese einem artifiziellen System überhaupt zugeschrieben werden kann.

Parthemore und Whitby erinnern daran, dass die Delegation von *moral agency* an artifizielle Handlungsträger*innen nicht von allen Wissenschaftler*innen befürwortet wird, sondern teils gar als utopisch bezeichnet wird.¹³¹ Die Begründung hierfür basiert jedoch nicht auf möglicherweise für *moral agency* notwendigen fehlenden Fähigkeiten artifizieller Systeme,

¹²⁹ Himma (2009), S. 20.

¹³⁰ Vgl. Karpov (2020), S. 68.

¹³¹ Vgl. Parthemore und Whitby (2013), S. 2.

sondern vielmehr an deren ursächlicher Beschaffenheit als biologisch nicht-lebendige und nicht-menschliche Wesenheiten.¹³² Ein artifizielles System, welches über sämtliche Fähigkeiten verfügt, die eine menschliche Instanz zweifelsfrei zu einem moralischen Handlungssubjekt machen würden, würde demnach nicht an seinen intellektuellen oder praktischen Fähigkeiten scheitern, sondern lediglich daran, dass diese nicht biologisch, sondern technisch hervorgebracht wurden und funktionieren. Dass diese Auffassung nicht gängiger Konsens ist, wurde bereits mittels der Unterscheidung diverser Formen von Akteursschaft nach Himma gezeigt.

Parthemore und Whitby setzen im Sinne Blacks und Karpovs bestimmte Merkmale als Bedingung für moralische Akteursschaft voraus. Insbesondere das Vorhandensein eines Selbst-Konzepts und der damit verbundenen Möglichkeit kritischer Selbst-Reflexion stehen hierbei an oberster Stelle.¹³³ Auch die Verantwortlichkeit der artifiziiellen Instanz für das von ihr ausgeführte Verhalten und die hiermit einhergehende dringende Notwendigkeit, dieser Instanz Verantwortung zuschreiben zu können, begrenzen den Terminus des moralischen Handlungssubjekts. Interessanterweise scheint bei Parthemore und Whitby noch eine weitere Kategorie auf, die in ähnlicher Sach- und Fachliteratur nicht zur Anwendung kommt. So wird hier deutlich zwischen dem Konstrukt des „conceptual agent: i.e., an agent that possesses and employs concepts“¹³⁴ und jenem des „moral agent“ unterschieden. Verdeutlicht wird an dieser Stelle, dass ein „conceptual agent“ nicht immer auch ein moralisches Handlungssubjekt sein muss, während es sich umgekehrt zwangsläufig so verhält. Moralische Handlungssubjekte dürften demnach eine Subkategorie von „conceptual agents“ darstellen.¹³⁵ Diese Differenzierung und gleichzeitig Konnotierung von Kategorien erinnert an die in früheren Teilen dieser Arbeit erwähnte Erläuterung Lohs, dass alle Roboter Maschinen, nicht aber alle Maschinen Roboter seien.¹³⁶

Interessant ist, dass im Rahmen der Diskussion rund um moralische Akteursschaft die Frage nach der Verbindung ebendieser zu unterschiedlichen Strömungen der Moralphilosophie aufkommt. In Hinblick darauf wird erläutert, es sei bei der Betrachtung von *moral agency* nicht grundlegend, wodurch sich moralische Handlungen bemessen lassen oder inwiefern (a)moralisch Handelnde in Verbindung mit bestimmten philosophischen Grundgedanken

¹³² Vgl. Parthemore und Whitby (2013), S. 2.

¹³³ Vgl. Ebd. S. 2.

¹³⁴ Ebd. S. 2.

¹³⁵ Ebd. S. 2.

¹³⁶ Vgl. Loh (2019), S. 10.

stehen, sondern vielmehr, auf welche Weise die Handelnden selbst konstituiert und konzipiert sein müssen, um als moralische Handlungssubjekte wahrgenommen werden zu können.¹³⁷ Dies erscheint aus mehreren Gründen problematisch.

Erstens verfehlt eine von den Grundsätzen der Moralphilosophie losgelöste Betrachtung moralischer Akteursschaft die ursprüngliche Absicht, zu determinieren, welche Instanzen als moralisch Handelnde geltend gemacht werden müssen, können und dürfen. Zweitens scheint es paradox, von moralisch Handelnden zu sprechen, ohne jedoch zu definieren, inwiefern der Terminus „moralisch“ für die Zuschreibung von „*moral*“ *agency* überhaupt relevant ist. Drittens wirft diese Sichtweise Fragen auf, die eine Aufweichung der Utopie der Verantwortungszuschreibung im Bereich artifizierter Systeme nicht nur debilitiert, sondern sogar noch verschärft. Schließlich kann moralische Verantwortung wohl kaum dem*derjenigen zugeschrieben werden, der*die ihrerseits zwar als „moralisches“ Handlungssubjekt identifiziert wurde, jedoch nicht aufgrund seiner*ihrer Fähigkeit, bestehende moralische Konzepte in Hinblick auf – und angelehnt an bestimmte Richtungen der Moralphilosophie zu hinterfragen und anzuwenden, sondern lediglich aufgrund des Vorhandenseins bestimmter kognitiver oder – im Falle von Maschinen – artifizell instruierter Charakteristika.

Da für die Verifizierung der dieser Arbeit zugrunde liegenden These, die Verantwortungszuschreibung in der MMI sei utopisch, eine saubere Argumentation und literaturgestützte Beweisführung nötig ist, müssen die Charakteristika, die als zentral für moralische Akteursschaft zählen, konkret hinterfragt werden. Hierfür eignet sich eine genaue Betrachtung jener Konzepte, die von Floridi und Sanders, sowie Sullins und Moor erarbeitet wurden.

6.2. Konzepte moralischer Akteursschaft

6.2.1. J.P. Sullins vs. L. Floridi und J.W. Sanders

Folgend werden die Konzepte J.P. Sullins sowie L. Floridi und J.W. Sanders gegenübergestellt. Um ein leichteres Verständnis des folgenden Abschnitts zu gewährleisten und die Vergleiche beider Konzepte schlüssig durchführen zu können, sei vorab bereits zusammenfassend erwähnt, welche Charakteristika für *moral agency* im Rahmen des jeweiligen Konzepts als zentral erachtet werden. Bei Sullins handelt es sich hierbei um

¹³⁷ Vgl. Parthemore und Whitby (2013), S. 3.

Autonomie, Intentionalität und Verantwortung, während Floridi und Sanders Autonomie, Interaktivität und Anpassungsfähigkeit nennen.¹³⁸ Im Gegensatz zu einigen bisher erläuterten Bedingungen für moralische Akteursschaft, ist sowohl der Ansatz von Sullins als auch jener von Floridi und Sanders explizit auf *moral agency* in Bezug auf nicht-menschliche Entitäten ausgerichtet. In den bisher dargelegten Konstrukten von Himma, Black und Karpov wurde bereits vermehrt auf das notwendige Vorhandensein eines Selbst-Konzepts sowie ausgeprägter Denkfähigkeit verwiesen wurde. Sullins erweitert diesen Diskurs um die folgenden Aspekte, die sein Konzept moralischer Akteursschaft zusammenfassen:

I detail three requirements for a robot to be seen as a moral agent. The first is achieved when the robot is significantly autonomous from any programmers or operators of the machine. The second is when one can analyze or explain the robot's behavior only by ascribing to it some predisposition or 'intention' to do good or harm. And finally, robot moral agency requires the robot to behave in a way that shows and understanding of responsibility to some other moral agent.¹³⁹

Bevor diese drei Charakteristika im Sinne Sullins' genauer erläutert werden können, ist es zunächst relevant, sie in ihrer Terminologie zu begreifen. Autonomie beispielsweise ist ein Konstrukt, das philosophisch mehrere Aspekte abdeckt. Im *Metzler Lexikon der Philosophie* wird Autonomie insbesondere im Kontext der Moralphilosophie mit Selbstbestimmtheit menschlicher Entitäten gleichgesetzt. Diese Selbstbestimmtheit könne ein Individuum lediglich erreichen, sofern es in der Lage ist, die Ergebnisse gefühlseigener Zustände wie Sehnsüchte oder Begierden zu kontrollieren.¹⁴⁰ Eine Erläuterung dessen ist in Zusammenhang mit dem Postulat Sullins', Autonomie sei ein Merkmal moralischer Akteursschaft, dringend notwendig, da der Autonomiebegriff ursprünglich mit einer Reihe von Aspekten verknüpft ist, die dem Menschen vorbehalten sind, bei Sullins allerdings limitiert und entschärft zur Anwendung kommen. Ein artifizielles System kann nach dieser Beschreibung von Autonomie folglich zwar autonomes Verhalten zeigen, nie aber im philosophischen Sinne autonom **sein**.

Ähnlich verhält es sich bei L. Floridi und J.W. Sanders: "Autonomy means that the agent is able to change state without direct response to interaction: it can perform internal transitions to change its state."¹⁴¹ In diesem Kontext werden bei Floridi und Sanders zwar nicht die

¹³⁸ Die deutsche Übersetzung der Termini "autonomy", "intention" und "responsibility" sowie "interactivity" und "adaptability" ist angelehnt an Loh (2020), S. 48-53.

¹³⁹ Sullins, John P.: When Is a Robot a Moral Agent? In: International Review of Information Ethics 6(12) 2006, S. 23.

¹⁴⁰ Metzler Lexikon Philosophie. Autonomie. <https://www.spektrum.de/lexikon/philosophie/autonomie/255>, (8.7. 2023)

¹⁴¹ Floridi, Luciano und J.W. Sanders: On the Morality of Artificial Agents. In: Wallach, Wendel und Peter Asaro (Hrsg.): Machine Ethics and Robot Ethics. London: Routledge 2017, S. 325.

Programmierer*innen des Roboters oder der Maschine erwähnt, so wie dies bei Sullins der Fall ist, jedoch liegt es nahe, dass die Annahme einer Unabhängigkeit von äußeren Reizen und die Möglichkeit einer Maschine, ohne externe Beeinflussung eigenständiges Verhalten zu kreieren, genau dasjenige autonome Verhalten beschreibt, welches auch Sullins als notwendige Bedingung für jedes moralische Handlungssubjekt voraussetzt. Interessant ist, dass auch das zweite Merkmal, welches Floridi und Sanders als zentral für *moral agency* erachten, in ähnlicher Form bei Sullins aufscheint. Floridi und Sanders sprechen von Interaktivität, während bei Sullins von Intentionalität die Rede ist.

In Hinblick auf das Kriterium der Intentionalität, verstanden als „Zielgerichtetheit des Handelns oder der Gefühle“¹⁴² ist eine Verwendung im Kontext artifizieller Akteur*innen weit weniger trennscharf als dies bei dem Konstrukt der Autonomie der Fall ist. Während, wie im Laufe dieser Arbeit bereits erwähnt, zwar kein Konsens darüber besteht, ob Roboter und Maschinen fühlen können oder Emotionen bloß nachahmen, ist die Fähigkeit des Handelns als eigenständige Tätigkeit nicht in Frage zu stellen. Fragwürdig ist viel eher, ob eine Verknüpfung des geforderten Autonomiebegriffs und dem Charakteristikum der Intentionalität möglich ist und somit das Kriterium der Verantwortung [im Sinne der Möglichkeit einer Verantwortungszuschreibung] erfüllt werden kann.

Kehrt man zurück zu Sullins' Betrachtung der von ihm postulierten Erfordernisse für *moral agency*, so wird sogleich klar, dass sein Konzept von Autonomie adaptiert ist, um zu gewährleisten, dass es auf nicht-menschliche Instanzen anwendbar ist. „I mean to use the term ‘autonomy,’ in the engineering sense, simply that the machine is not under the direct control of any other agent or user.“¹⁴³ Sullins betont in diesem Kontext, dass diese Form der Autonomie einer Vielzahl an Entitäten zugeschrieben werden kann, was jedoch nicht dazu verleiten darf, zu implizieren, diese seien hierdurch gleichsam automatisch moralische Handlungssubjekte. Als solche ließen sich Instanzen lediglich dann bezeichnen, sofern Kriterium zwei und drei – Intentionalität und Verantwortung – ebenfalls erfüllt sind.¹⁴⁴

Die Zielgerichtetheit intentionalen Handelns basiert für Sullins insbesondere auf der Interaktion des Roboters oder der Maschine und seiner*ihrer Programmierung mit der wie auch immer beständigen Umwelt. Solange eine Maschine so handelt, dass ihre Aktionen als vorsätzlich und berechenbar gelten, kann sie laut Sullins als moralisches Handlungssubjekt

¹⁴² Metzler Lexikon Philosophie. Intentionalität.

<https://www.spektrum.de/lexikon/philosophie/intentionalitaet/993> (8.7. 2023)

¹⁴³ Sullins (2006), S. 28.

¹⁴⁴ Vgl. Ebd. S. 28.

bezeichnet werden.¹⁴⁵ Bezüglich der Frage nach einer absoluten Entsprechung der von ihm beschriebenen Intentionalität mit dem gängigen Verständnis ebendieser, schreibt Sullins: “There is no requirement that the actions really are intentional in a philosophically rigorous way, nor that the actions are derived from a will that is free on all levels of abstraction. All that is needed is that [...] there is a comparable level of personal intentionality and free will between all the agents involved.”¹⁴⁶

Floridi und Sanders, die Interaktivität anstelle von Intentionalität als zentral erachten, beschreiben diese wie folgt: „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 [...]“¹⁴⁷ Wenngleich Interaktivität und Intentionalität nicht als Synonyme zu begreifen sind, so liegt dennoch zwischen Sullins’ Konzept und jenem von Floridi und Sanders eine Parallele vor. Möglicherweise kann Interaktivität, wie sie bei Floridi und Sanders beschrieben wird, als Voraussetzung für das Kriterium der Intentionalität nach Sullins gelesen werden, insbesondere, wenn das dritte Charakteristikum – Anpassungsfähigkeit - nach Floridi und Sanders betrachtet wird: „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 [...] as learning its own mode of operation in a way which depends critically on its experience.”¹⁴⁸ Die Fähigkeit eines Roboters, basierend auf eigener Erfahrung die ihm implementierten Regeln und Ideen zu adaptieren, kann m.E. sowohl als eigenständige Voraussetzung für moralische Handlungssubjekte genannt werden als auch in Konnotation mit dem sowohl von Sullins als auch Floridi und Sanders postulierten Merkmal der Autonomie verknüpft werden. Ein Roboter, welcher basierend auf erlebten Erfahrungen die eigenen [moralischen] Vorstellungen und Handlungsvorschriften ändern kann¹⁴⁹, verfügt grundsätzlich über ein bestimmtes Maß an Autonomie, welches wiederum in Kombination mit den beiden anderen geforderten Kriterien zwangsläufig hinreichend für *moral agency* ist.

Sullins’ sowie Floridi und Sanders’ Konzepte unterscheiden sich demnach nur in einem Punkt maßgeblich. Während Sullins nur jene Maschinen als moralische Akteur*innen bezeichnen würde, die für ihr Handeln verantwortlich gemacht werden können, ist dies für Floridi und Sanders kein Erfordernis. In Anbetracht der im Rahmen dieser Arbeit bereits

¹⁴⁵ Vgl. Sullins (2006), S. 28.

¹⁴⁶ Ebd. S. 28.

¹⁴⁷ Floridi und Sanders (2017), S. 325.

¹⁴⁸ Ebd. S. 326.

¹⁴⁹ Vgl. Ebd. S. 326.

dargelegten verschiedenen Konzepte von Verantwortung liegt die Frage nahe, auf welches Verantwortungskonzept sich Sullins bezieht. Tatsächlich begreift dieser „Verantwortung“ offenbar lediglich als die Fähigkeit einer Maschine, sich selbst nicht als losgelöst von ihrer Umwelt zu betrachten, sondern ihre Handlungen so zu wählen oder anzupassen, als sei sie sich ihrer Verantwortung gegenüber anderen moralischen Akteur*innen bewusst.¹⁵⁰ Weshalb diese und ähnliche Konzepte im Kontext der Verantwortungszuschreibung in der MMI nicht hinreichend sind, wird in späteren Teilen dieser Arbeit diskutiert. Vorerst ist es relevant, darzulegen, wie unterschiedlich die Bedingungen für *moral agency* beschrieben werden und wie ähnlich die ihnen zugrunde liegenden Konzepte teils sind, wenngleich nur selten dieselben Termini verwendet werden. Die Frage, ob Roboter und Maschinen unter bestimmten Voraussetzungen als *moral agents* bezeichnet werden können, wird sowohl von Sullins als auch von Floridi und Sanders positiv beantwortet.

6.2.2. Moral agency nach C. Misselhorn

Auch Catrin Misselhorn vertritt die Auffassung, es sei durchaus möglich, nicht-menschliche bzw. artifizielle Entitäten als moralische Handlungssubjekte zu begreifen. Hierfür sei allerdings ein zweidimensionales Dimensionsfeld vonnöten, welches aufzeigt, unter welchen Bedingungen Akteur*innen als *moral agents* klassifiziert werden können:

The first dimension is that being an agent involves the capacity to act for reasons (Davidson 1980; Dretske 1995). The second dimension is that agents are in some sense self-originating sources of their doings, i.e., that their actions are not determined by external factors, involve some sort of flexibility, or are under the control of the agent.¹⁵¹

An dieser Stelle lässt sich eine Parallele zu Himma finden, wenngleich dort nicht von “capacity to act for reasons“, sondern von “moral reasoning“ im Sinne eines Verständnisses für moralische Konzepte und nicht von “self-origination“, sondern von “consciousness“¹⁵² im Sinne eines ausgeprägten Bewusstseins als Voraussetzungen für *moral agency* die Rede ist. Misselhorns zweidimensionales Konzept erinnert allerdings auch in Teilen an das Charakteristikum der Autonomie, welches sowohl Sullins als auch Floridi und Sanders als Voraussetzung für *moral agency* implizieren. So ist mit “self-origination“ bei Misselhorn die Fähigkeit eines artifiziellen Systems gemeint, ohne Außeneinwirkung Handlungen zu beschließen und durchzuführen. Bei Sullins und Floridi und Sanders wird ebendies unter dem Terminus Autonomie subsummiert. Auch das Merkmal “ability to act for reasons“

¹⁵⁰ Vgl. Sullins (2006), S. 28.

¹⁵¹ Vgl. Misselhorn (2013), S. 44.

¹⁵² Vgl. Ebd. S. 24.

erinnert semantisch sowohl an das Kriterium der Interaktivität bei Floridi und Sanders als auch an jenes der Intentionalität bei Sullins. Diese Annahme wird tatsächlich durch einen Hinweis Misselhorns gestützt: “[...] moral agents do not just act for reasons, but [...] these reasons are the result of a process of deliberation. Understood this way, moral agency requires so-called higher-order intentionality (Dennett 1981), i.e., the capacity to reflect on, evaluate and adjust one's own first-order mental states [...]”¹⁵³ Der Begriff der Intentionalität scheint hier in Misselhorns detaillierten Erläuterungen demnach tatsächlich auf. Auch das Verb “adjust“ wird in diesem Zusammenhang erwähnt, was wiederum als eine Parallele zu dem Charakteristikum der “Adaptability“ nach Floridi und Sanders gelesen werden kann. Interessanterweise scheint Misselhorns hybrides Konzept moralischer Akteursschaft demnach in seinem Kern einen Großteil jener Kriterien zu umfassen, welche im Rahmen anderer Konstrukte bereits unter einer anderen Bezeichnung zusammengefasst sind.

Betrachtet man die bisher erläuterten Bedingungen moralischer Akteursschaft sowohl losgelöst von den ihnen zugrunde liegenden Konzepten als auch im Kontext ebendieser, so lassen sich zweifelsfrei eine Reihe an Gemeinsamkeiten erkennen. Fraglich ist allerdings immer noch, inwiefern sich die unterschiedlichen postulierten Kriterien für *moral agency* tatsächlich praktisch anwenden lassen und wenn ja – ob alle moralischen Handlungssubjekte auf demselben Niveau moralische Akteur*innen sind, oder ob sich allein durch die teils unterschiedlichen Aufgaben und Funktionsbereiche der jeweiligen artifiziellen Systeme Unterschiede im Grad ihrer Akteursschaft ergeben. Im Folgenden wird hierzu der Ansatz J.H. Moors beleuchtet, welcher genau diese Frage aufnimmt und mittels einer Klassifizierung moralischer Akteur*innen ein differenziertes Verständnis moralischer bzw. ethischer Handlungsfähigkeit ermöglicht.

6.2.3. Moral agency nach J.H. Moor

J.H. Moor präsentiert anders als Floridi, Sanders, Sullins und Misselhorn ein Konzept moralischer Akteursschaft, welches eine hierarchische Gliederung vermeintlicher moralischer und ethischer Akteur*innen ermöglicht. Das Fehlen eines oder mehrerer Merkmale, die eine*n Akteur*in im Kontext anderer Konzepte entweder zu einem moralischen Handlungssubjekt machen, oder aber ihm automatisch die Fähigkeit moralischer Akteursschaft absprechen würden, wird bei Moor anders bewertet.¹⁵⁴ So können unterschiedliche Instanzen in vier verschiedene Typen moralischer Akteur*innen eingeteilt

¹⁵³ Misselhorn (2013), S. 47.

¹⁵⁴ Vgl. Loh 2020, S. 54-55.

werden. Moor unterscheidet “ethical impact agents“, “implicit ethical agents“, “explicit ethical agents“, und “full ethical agents”¹⁵⁵ Der leichten Lesbarkeit halber wird im Folgenden die Übersetzung von Loh herangezogen, welche Moors Kategorien als „Akteur*innen mit ethischem Einfluss“, „implizit ethische Akteur*innen“, „explizit ethische Akteur*innen“ sowie „ethische Akteur*innen im vollen Sinne“¹⁵⁶ bezeichnet. Moors Konzept moralischer Akteursschaft basiert offenbar auf einer klar strukturierten Abstufung bestimmter Handelnder. Es scheint daher viel weniger das Ausmaß an Fähigkeiten, die ein*e Akteur*in hat, eine Rolle zu spielen, als vielmehr der Grad an ethischem Einfluss, der mit diesen Fähigkeiten oder dem hierdurch ausgelösten Verhalten einhergeht. Ein künstliches Kompositum beispielsweise, welches selbst keine moralischen und ethischen Entscheidungen treffen kann, aber dennoch durch seinen Einsatz ethischen Einfluss besitzt, würde schlicht aufgrund dieses Einflusses als Akteur*in mit ethischem Einfluss und somit als moralisches Handlungsobjekt gelten. Moor erwähnt in diesem Kontext beispielsweise artifizielle Systeme – sogenannte “robotic camel jockeys”¹⁵⁷, die in Qatar anstelle von Kindersklaven eingesetzt werden, um Kamele zu reiten. Die Jockeys selbst verfügen zwar nur über wenige Funktionen, ihr ethischer Einfluss ist unter der Prämisse, dass hierdurch menschliche Kinder der Sklaverei entkommen können, dennoch groß.

Implizit ethische Akteur*innen verfügen über ähnlich simples moralisch relevantes Potenzial, unterscheiden sich jedoch hinsichtlich ihrer Fähigkeiten von den eben genannten Akteur*innen mit ethischem Einfluss, indem sie in der Lage sind, durch die ihnen implementierten Algorithmen unethisches Verhalten zu vermeiden und somit im Idealfall moralisch erwünschte Handlungen zu setzen. Ein Beispiel für eine*n implizit ethische*n Akteur*in ist demnach beispielsweise „[...] 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.”¹⁵⁸ Anders verhält es sich mit explizit ethischen Akteur*innen, die wohl im Sinne Sanders’, Floridis und Sullins’ am ehesten den Kriterien für *moral agency* gerecht würden, während sowohl *ethical impact agents* als auch *implicit ethical agents* wohl am ehesten als moralische Handlungsobjekte eingestuft werden würden. Explizit ethische Akteur*innen stellen bei Moor tatsächlich die höchste Stufe moralischer Akteursschaft da, die artifizielle Systeme

¹⁵⁵ Moor, James H.: The Nature, Importance, and Difficulty of Machine Ethics. In: IEEE Intelligent Systems, 21(4) 2006, S. 19-20.

¹⁵⁶ Loh (2020), S. 55-57.

¹⁵⁷ Moor (2006), S. 19.

¹⁵⁸ Ebd. S. 19.

erlangen können, da Moor davon ausgeht, dass – wie sogleich erläutert wird – lediglich Menschen als ethische Akteur*innen im vollen Sinne zu betrachten sind.¹⁵⁹

Loh merkt an dieser Stelle an, dass explizit ethische Akteursschaft „quasimoralisch“ zu sein scheint, da Moor diese zwar Robotern und Maschinen zugesteht, jedoch trotz ähnlicher Erfordernisse für Akteur*innen im vollen Sinne keinem Roboter ebendiesen Status erlauben würde.¹⁶⁰ Explizit ethische Akteursschaft zeichnet sich durch „die Kompetenzen, ethische Urteile zu fällen und zu evaluieren, autonom in intransparenten Alltagssituationen zu entscheiden und mit unvorhergesehenen Faktoren umzugehen [...]“¹⁶¹ aus. Für *full ethical agents* setzt Moor folgendes voraus: „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. We typically regard humans as having consciousness, intentionality, and free will.“¹⁶² Anhand dieser Erläuterungen liegt es nahe, zu postulieren, dass Moor die volle ethische Akteursschaft artifizieller Systeme lediglich deshalb für unerreichbar hält, weil artifizielle Systeme nicht biologisch menschlich sind. Interessanterweise findet sich hier eine Parallele zu Parthemore und Whitby, die im Kontext moralischer Akteursschaft ähnliche Limitationen erwähnen. Für die Frage nach der Verantwortungszuschreibung in der MMI ist Moors Ansatz hinsichtlich seiner Komplexität besonders interessant, gerade in Anbetracht der Tatsache, dass zwei der vier von ihm aufgestellten Thesen zu *moral agency* von anderen genannten Wissenschaftler*innen zweifelsfrei abgelehnt werden würden. Vielmehr würden sie allerdings dem Konzept von *moral patiency* entsprechen, welches dazu dient, Roboter, die nicht oder nur unzureichend die Kriterien für moralische Handlungssubjekte erfüllen, als moralische Handlungsobjekte darzustellen.

6.3. Roboter als moralische Handlungsobjekte

Betrachtet man jene Literatur, die sich mit Robotern als moralischen Handlungssubjekten auseinandersetzt und jene, die sich damit befasst, unter welchen Bedingungen Roboter und Maschinen als moralische Handlungsobjekte zu bezeichnen sind, fällt auf, dass bei weitem weniger Literatur zu dem Konzept von *moral patiency* zu finden ist, als dies im Kontext von *moral agency* der Fall ist. Dies dürfte vermutlich insbesondere daran liegen, dass mit *moral patiency* signifikant weniger problematische Fragestellungen verbunden sind, aber auch dadurch begründet sein, dass viele Wissenschaftler*innen *moral patiency* mit *moral agency*

¹⁵⁹ Moor (2006), S. 20.

¹⁶⁰ Loh (2020), S. 57.

¹⁶¹ Ebd. S. 57.

¹⁶² Moor (2006), S. 20.

konnotiert sehen. *Moral patiency* sei demnach ein Terminus beziehungsweise ein Konzept, das kontextuell betrachtet wird und daher nicht mit einer Reihe eigenständiger Überlegungen bedacht werden muss.¹⁶³ Artificielle Systeme, die als moralische Handlungsobjekte gelten, tragen grundsätzlich – wenn überhaupt – nur einen Bruchteil der Verantwortung, über deren Zuschreibung bei moralischen Handlungsobjekten debattiert wird. *Moral patiency* scheint damit einerseits ergänzend zu *moral agency* zu verstehen sein, andererseits jedoch auch komplementär hierzu zu stehen. Peter und Danaher verstehen moralische Handlungsobjekte wie folgt:

A moral patient is [...] a being who possesses some moral status— i.e. is owed moral duties and obligations, and is capable of suffering moral harms and experiencing moral benefits—but who does not take ownership over the moral content of its own existence. To put it another way, a moral patient is a being upon whom well-being (and other valuable states) are bestowed, but which does not (or cannot) take an interest in the autonomous formulation and pursuit of its own moral goals.¹⁶⁴

Das Recht auf einen moralischen Status dürfte bestimmte Maschinen im Bereich artifizieller Systeme von Gegenständen unterscheiden. Während ein Objekt, verstanden als unbelebter Gegenstand, keinerlei Recht auf einen bestimmten Umgang hat, verhält es sich bei Maschinen und Robotern unter bestimmten Voraussetzungen anders. Mit ihnen darf demnach nicht nach Belieben verfahren werden, da ihnen eine Form von moralischem Status zukommt. Die Frage danach, woran sich bemisst, ob jene Instanzen, die mit besagten Gegenständen, Maschinen oder Robotern in Interaktion stehen, diesen gegenüber moralisch handeln müssen, wird interessanterweise kontrovers diskutiert. Die Auffassung, dass mit moralischen Handlungsobjekten auf eine moralisch vertretbare Weise umgegangen werden muss, wenngleich diese selbst kein moralisch angemessenes Verhalten an den Tag legen können, wird beispielsweise dadurch gerechtfertigt, dass sie in einem möglicherweise bestehenden Naheverhältnis zu menschlichen Akteur*innen stehen und sich hierdurch sowie durch die Konsequenz, sie für menschliche Instanzen bedeutungstragend sind, auch Rechte und Werte besitzen.¹⁶⁵ „[...] Maschinen werden eben nicht um ihrer selbst willen als moralische Handlungsobjekte identifiziert, sondern aufgrund ihres Nutzens für die Menschen. Roboter sind damit zwar *moral patients*, allerdings aus instrumentellen Gründen, nämlich zum Schutz des moralisch angemessenen menschlichen Charakters.“¹⁶⁶

¹⁶³ Vgl. Loh (2020), S. 72.

¹⁶⁴ Danaher, John: The rise of the robots and the crisis of moral patiency. In: AI & society. London: Springer 2017, S. 132. sowie Peter, Fabienne: Pure epistemic proceduralism. In: Episteme 5(1) 2008, S. 36.

¹⁶⁵ Vgl. Loh (2020), S. 75.

¹⁶⁶ Loh (2020), S. 86.

Die Debatte darüber, ob nun alle Maschinen und/oder Roboter grundsätzlich das Potenzial haben, als moralische Handlungsobjekte klassifiziert zu werden, oder ob dieses Privileg nur manchen artifiziellen Systemen zukommt, zeichnet sich durch Dissens aus. Fest steht, dass all jene Roboter, die als moralische Handlungssubjekte (*moral agents*) gelten, automatisch auch moralische Handlungsobjekte sind.¹⁶⁷ Umgekehrt verhält es sich allerdings nicht so. *Moral patiency* dürfte demnach als eine Art Basis für echte moralische Akteursschaft und damit verbundene Verantwortlichkeit gelten. Interessant ist hierbei insbesondere, dass bei moralischen Handlungssubjekten in erster Linie deren Fähigkeiten und internalisierten Programme sowie Algorithmen Beachtung finden, während im Kontext von *moral patiency* weit mehr über die äußerliche Form der zugehörigen Roboter und Maschinen diskutiert wird. Dies hat folgenden Grund:

Nicht nur, aber besonders im Kontext der Frage nach der Verantwortungszuschreibung in der MMI ist es relevant, die Fähigkeit der menschlichen Psyche, unbelebten Gegenständen emotionalen Wert zuzuschreiben, zu beleuchten.¹⁶⁸ Durch eine wie auch immer geartete emotionale Bindung an ein wie auch immer beschaffenes Gegenüber wird gleichsam die Basis für das Eingehen einer nicht egalitären Beziehung gebildet. Ob es sich bei diesem Gegenüber um unbelebte Gegenstände wie beispielsweise Stofftiere, Bücher oder kleine Skulpturen handelt, ob um Maschinen wie Computer, Tablets oder Laptops oder um artifizielle Systeme, die aufgrund ihrer Fähigkeiten den Anschein vermitteln, real zu sein, ist vorerst irrelevant. Tatsache ist, dass Menschen in der Lage sind, sich emotional an vermeintlich leblose Dinge zu binden und diesen ungeahnten Wert zuzuschreiben. Die Gründe hierfür dürften individuell unterschiedlich sein. Die Rede ist bei Loh sowohl von bestimmten Gefühlen und Assoziationen, die mittels eines Gegenstandes in uns ausgelöst werden, bis hin zu emotional aufgeladenen Objekten, die starke Erinnerungen in uns auslösen oder momentane impulsive Gefühlszustände hervorbringen.¹⁶⁹

Diese Auffassung ist angelehnt an das Konzept des Anthropomorphismus, welches sich mit der Vermenschlichung nicht-lebendiger Gegenstände befasst. Fragt man sich zunächst, inwiefern Anthropomorphisierung mit *moral patiency* konnotiert ist, so wird man schnell feststellen, dass bei Überlegungen zu moralischen Handlungsobjekten die Frage nach dem Umgang mit ebendiesen zentral ist. Wie bereits erwähnt, genießen all jene Instanzen, die als Subjekte der Moral gesehen und anerkannt werden, einen besonderen Status, der ihnen

¹⁶⁷ Vgl. Loh (2020), S. 72-74.

¹⁶⁸ Vgl. Ebd. S. 75-76.

¹⁶⁹ Vgl. Ebd. S. 82.

gewisse Rechte und Möglichkeiten verleiht, die „Objekte“ normalerweise nicht haben oder lediglich dann erhalten, wenn Menschen sie ihnen aus individuell unterschiedlichen Gründen verleihen. Loh vertritt die These, dass Menschen grundsätzlich zu einer Identifikation mit nicht-menschlichen Instanzen fähig sind und dass diese Identifikation umso stärker ist, je deutlicher im nicht-menschlichen und nicht-lebendigen Gegenüber menschliche Charakteristika erkannt werden.¹⁷⁰ Dies bedeutet einerseits, dass ein anthropomorphes Design beziehungsweise ein Design, welches in Menschen zumindest eine Assoziation mit menschlichen Merkmalen hervorruft, gezielt eingesetzt werden kann, um in menschlichen Instanzen Gefühle für unbelebte Materie hervorzurufen. Andererseits bedeutet dies auch, dass im Kontext moralischer Handlungsobjekte vermutlich genau deshalb über bestimmte Pflichten gegenüber ebendiesen diskutiert wird, weil anthropomorph gestaltete Objekte, Roboter und Maschinen andere Emotionen in Menschen auslösen als jene, die lediglich als Gegenstand oder künstliches Design wahrgenommen werden. Inwiefern dies in Hinblick auf die Verantwortungszuschreibung in der MMI relevant ist, soll in den folgenden Abschnitten dieser Arbeit diskutiert werden.

7. Die Dichotomie der Verantwortungszuschreibung

7.1. Utopische Zuschreibungen in Therapie- Medizin und Pflegerobotik

Um die bislang genannten Konzepte, Thesen und Überlegungen auf konkrete Beispiele anwenden zu können, wurden für diese Arbeit zwei Bereiche der Robotik gewählt, die sich für eine Darlegung der Utopie der Verantwortungszuschreibung in der MMI besonders gut eignen. Der Teilbereich der Therapie- Medizin und Pflegerobotik beschäftigt sich hierbei nicht nur mit den Aufgabenbereichen unterschiedlicher Maschinen und Roboter. Insbesondere wird im Rahmen dieser Bereichsethik intensiv diskutiert, wie vielschichtig der Einsatz von Maschinen in allen zugehörigen Arbeitsfeldern ist. Wenngleich beispielsweise Exoskelette und autonome Rollstühle ebenso zu Robotern im Gesundheitsbereich zählen¹⁷¹, richtet sich die folgende philosophische Betrachtung nicht auf solche und ähnliche Maschinen, sondern auf jene, deren Interaktionsmuster im Rahmen der Frage nach Verantwortungsdiffusion- und Zuschreibung in der MMI relevant sind oder dies in Zukunft sein werden. Zu nennen sind im Zuge dessen nicht nur Therapie- und Pflegeroboter, sondern auch Operationsroboter, welche Chirurg*innen bei ihrer Arbeit unterstützen können.

¹⁷⁰ Vgl. Loh (2020), S. 78.

¹⁷¹ Vgl. Bendel, Oliver: Roboter im Gesundheitsbereich. In: Bendel, Oliver (Hrsg.): Pflegeroboter. Windisch: Springer Gabler 2018, S. 196.

7.1.1. DaVinci

Die Sorge, dass ein Operationsroboter bestimmte chirurgische Eingriffe tatsächlich autonom durchführen kann, ist bislang unbegründet. So dienen Operationsroboter nicht etwa als Ersatz für ausgebildete Chirurg*innen und Ärzt*innen, sondern stehen diesen lediglich bei der Durchführung bestimmter Maßnahmen im Rahmen von operativen Eingriffen unterstützend zur Seite.¹⁷² Insbesondere die Möglichkeit, kleinste Bewegungen präzise zu setzen und hierdurch die Genauigkeit bei der Durchführung der Operation zu erhöhen, kann als klarer Vorteil gewertet werden. Des Weiteren ist anzumerken, dass Operationsroboter - beispielsweise das DaVinci-System von *Intuitive Surgical* - nicht autonom oder teilautonom fungieren, sondern als Teleroboter von Chirurg*innen gesteuert werden.¹⁷³ Die Operation selbst wird daher trotz der Involviertheit eines artifiziellen Systems von menschlichen Instanzen angeleitet, überwacht und durchgeführt. Roboter sind lediglich Assistent*innen, die mögliche menschliche Fehler ausgleichen können.¹⁷⁴

Im Rahmen einer Beleuchtung der Verantwortungsfrage entstehen beim Einsatz von nicht-autonomen Operationsrobotern demnach bei weitem weniger Probleme als dies beim Einsatz anderer artifizieller Systeme in ähnlichen Arbeitsfeldern der Fall ist. Dennoch kann die Frage der Verantwortungszuschreibung auch im Falle von Telerobotern [wie bspw. dem DaVinci-System] nicht eindimensional betrachtet und beantwortet werden. An dieser Stelle kommen die im Laufe dieser Arbeit geschilderten Verantwortungskonzepte sowie die hiermit konnotierte Unterscheidung von moralischen Handlungssubjekten und solchen Handlungsobjekten zum Tragen. Das DaVinci-System wird nun im Ausschlussverfahren in Hinblick auf die Konzepte von Sullins, Floridi und Sanders sowie Misselhorn und Moors beleuchtet.

Da der DaVinci Roboter nicht autonom arbeiten kann, sondern auf die Steuerung von außen angewiesen ist, ist das erste für Sullins relevante Charakteristikum für moralische Akteursschaft ausgeschlossen. Auch die für *moral agents* relevante Intentionalität kann im Falle einer nicht-autonomen Maschine als nicht gegeben gesehen werden. Hiermit würde nach Sullins automatisch auch das dritte Kriterium für moralische Akteursschaft – Verantwortlichkeit - unerfüllt bleiben. Somit dürfte der DaVinci-Roboter von Sullins wohl

¹⁷² Bendel (2018), S. 201.

¹⁷³ Vgl. Ebd. S. 201.

¹⁷⁴ Ebd. S. 201; Vgl. auch Bekey, George A: Current trends in robotics: Technology and ethics. In: Lin, Patrick und Keith Abney u.a. (Hrsg.): Robot ethics. The ethical and social implications of robotics. Cambridge: MIT Press 2012, S. 23.

höchstens als moralisches Handlungsobjekt, jedoch niemals als moralischer Akteur bezeichnet werden. Betrachtet man den Roboter im Kontext von Floridi und Sanders, so müssten die Kriterien „Autonomie, Interaktivität und Anpassungsfähigkeit“ gegeben sein, um moralische Akteursschaft zuschreiben zu können. Kriterium 1 ist, wie soeben im Kontext Sullins’ erläutert, nicht erfüllt. Das Merkmal der Interaktivität [Kriterium 2] hingegen kann im Falle eines Roboters, welcher explizit für die Interaktion mit menschlichen Instanzen kreiert wurde, als gegeben betrachtet werden. Anpassungsfähigkeit, das dritte Kriterium, welches Floridi und Sanders als notwendige Bedingung für moralische Handlungssubjekte voraussetzen, ist zwar vermeintlich gegeben, indem angenommen werden könnte, jedes artifizielle System, welches Handlungen unterschiedlich und je nach Situation performen kann, sei anpassungsfähig, allerdings ist dies wie bereits im zugehörigen Kapitel 6.2.1. erläutert, nicht dasjenige, was Floridi und Sanders unter “Adaptability” verstehen. Vielmehr würde diese voraussetzen, dass besagtes artifizielles System eine eigene Handlungsstrategie besitzt, die basierend auf gesammelten Erfahrungen abgeändert und angepasst werden kann.¹⁷⁵ Dies ist beim DaVinci-Roboter nicht der Fall. Somit ist nach dem Konzept von Floridi und Sanders lediglich eines von drei Merkmalen moralischer Akteursschaft erfüllt. Das DaVinci-System klassifiziert sich demnach auch im Kontext dieser Theorie im besten Fall als moralisches Handlungsobjekt, nicht jedoch als moralisches Handlungssubjekt. Eine Betrachtung des Roboters im Sinne Misselhorn ist an dieser Stelle obsolet, da durch die bereits gezeigten Überlegungen klar gezeigt werden konnte, dass das DaVinci System weder über “capacity to act for reasons” verfügt, noch über eine wie auch immer geartete Form der “self-origination” und somit auch von ihr keineswegs als *moral agent* eingestuft werden könnte. Zu guter Letzt kann das DaVinci-Operationssystem im Rahmen des mehrdimensionalen Konzepts von Moor höchstens als Akteur mit ethischem Einfluss bezeichnet werden und fällt somit eindeutig in den Bereich des moralischen Handlungsobjekts.

Streng genommen trägt er folglich keinerlei moralische und/oder ethische Verantwortung. Genau an dieser Stelle ergibt sich das Problem der Verantwortungszuschreibung in der MMI. Gegeben ist ein artifizielles System, welches zwar für ethisch und moralisch relevante Zwecke eingesetzt wird und somit ethischen Einfluss hat, sich dieses Einflusses jedoch nicht bewusst ist. Es fungiert als Hilfsmittel für die eigentlichen ethischen Akteur*innen, in diesem Fall Chirurg*innen, Ärzt*innen und gegebenenfalls die menschlichen

¹⁷⁵ Floridi und Sanders (2017), S. 326.

Operationsassistent*innen. Hieraus ergibt sich die folgende Frage: Wenn das DaVinci-System nicht als moralische*r Akteur*in angesehen werden kann und ihm*ihr daher die Fähigkeit, Verantwortung zu übernehmen, nicht zugeschrieben werden kann, wie wird in diesem Fall die Verantwortlichkeit für etwaige riskante Fehler [mit problematischen Folgen] geklärt? Wer haftet für Schäden, die bei einem*einer Patient*in entstehen, wenn durch die Assistenz eines Roboters operative Eingriffe nicht wie geplant ablaufen? Im Falle des DaVinci-Operationssystems kann auf das bereits geschilderte Konzept der herstellungsorientierten und gebrauchorientierten Verantwortung zurückgegriffen werden, um zu verdeutlichen, dass die Zuschreibung von Verantwortung im Fall artifizieller Systeme nicht realistisch ist. Folgt man der These, die Konstrukteur*innen von Maschinen seien alleinig für etwaige entstehende Probleme und Risiken verantwortlich, so ergibt sich die Problematik, dass die Konstrukteur*innen keine direkt zu identifizierenden Personen sind, sondern eine breite Masse an unterschiedlichen Instanzen, die an der Herstellung beteiligt waren.¹⁷⁶

Lediglich den Nutzer*innen des Operationssystems Verantwortung für mögliche Fehler zuzuschreiben, erscheint jedoch genauso wenig realistisch, schließlich sind Ärzt*innen zwar im besten Fall Nutznießer*innen vorangeschrittener Technik und können hierdurch die Ergebnisse bestimmter Eingriffe für ihre Patient*innen verbessern, allerdings kann wohl kaum ein technischer Fehler dazu führen, dass Chirurg*innen für die daraus entstandenen Schäden verantwortlich gemacht werden. Der *Accountability Gap* ist also, wie sich bereits an der Simulation eines vergleichsweise eindimensionalen artifiziellen Systems gezeigt hat, groß und darf nicht außer Acht gelassen werden, wenn die MMI philosophisch und ethisch betrachtet wird. Nun gibt es im Bereich der Therapie- Medizin und Pflegerobotik allerdings nicht nur von außen steuerbare Systeme, sondern auch solche, die tatsächlich oder zumindest vermeintlich autonom agieren können. Meistens handelt es sich hierbei um sogenannte Therapie- oder Pflegeroboter.

Therapieroboter unterstützen therapeutische Maßnahmen oder wenden selbst, häufig als autonome Maschinen, solche an [...] Sie sind mit ihrem Aussehen und in ihrer Körperlichkeit wie traditionelle Therapiegeräte präsent, machen aber darüber hinaus selbst Übungen mit Gelähmten, unterhalten Betagte und fordern Demente und Autisten mit Fragen und Spielen heraus. Manche verfügen über mimische, gestische und sprachliche Fähigkeiten und sind in einem bestimmten Umfang denk- und lernfähig [...] ¹⁷⁷

Anders als das eben genannte Beispiel eines Teleroboters, sind Therapieroboter demnach in der Lage, komplexe Interaktionen mit Menschen durchzuführen. Hierdurch vermitteln sie

¹⁷⁶ Vgl. Fenner (2022), S. 278-279.

¹⁷⁷ Bendel (2018), S. 202.

nicht nur den Anschein, über menschliche Fähigkeiten zu verfügen, sondern täuschen überdies in weiten Teilen ein ausgeprägtes Denkvermögen vor. Da Denkvermögen in hohem Maße mit Autonomie konnotiert ist und diese wiederum direkt mit den Kriterien für moralische Akteursschaft in Verbindung steht, liegt die Frage nahe, ob bestehende Verantwortungskonzepte auf solche artifiziellen Systeme problemlos angewendet werden können. Dies soll im Folgenden geprüft werden.

7.1.2. Paro

Paro, ein Therapieroboter, der vorwiegend in der Arbeit mit an Demenz erkrankten Menschen zum Einsatz kommt, ist einer Robbe nachempfunden und verfügt über spezielle Features, die es ihr erlauben, die Interaktionen mit den ihr zugeteilten Patient*innen zu überblicken. So kann Paro beispielsweise auf eine Kontaktaufnahme reagieren, indem ihr das Wissen über ihren eigenen „Namen“ implementiert wurde. Des Weiteren „erinnert“¹⁷⁸ sie sich an die wie auch immer geartete Behandlung zu speichern, die ihr Patient*innen zu Teil werden lassen. Insbesondere kann die artifizielle Robbe auch unterschiedliche Geräusche und eine Reihe an Bewegungen ausführen, was den Anschein vermittelt, sie hätte Gefühle und eine gezielte Reizwahrnehmung.¹⁷⁹ Wenngleich dies ein Fehlschluss ist, ermöglicht die Interaktion mit Paro vielen demenziell erkrankten Menschen eine Begegnung mit einem vermeintlich realen Tier, das sie jedoch keineswegs verletzen können, selbst wenn die Robbe bei Fehlverhalten den Anschein vermitteln würde, Gefühle zu empfinden. Als therapeutische*r Begleiter*in ist Paro daher ideal für die Interaktion mit einem menschlichen Gegenüber geeignet.

Nachteile der Interaktion mit einem solchen Therapieroboter dürften kaum vorhanden sein, was die Frage nach der Verantwortungszuschreibung vermeintlich obsolet macht. Dennoch ist es interessant, zu überlegen, inwiefern Paro als *moral agent* oder als *moral patient* klassifiziert werden kann, schließlich wäre es durchaus etwas Besonders, ein tierähnliches artifizielles System als Verantwortungsträger*in einzustufen, wenn bei realen Tieren eine solche Zuschreibung nicht in Frage käme. Janina Loh hat sich bereits damit auseinandergesetzt, bestehende Ansätze auf unterschiedliche artifizielle Systeme anzuwenden und dabei festgestellt, dass Paro zwar keinem Konzept nach als moralisches Handlungssubjekt verstanden werden kann, jedoch auch nicht automatisch als moralisches

¹⁷⁸ Bendel (2018), S. 202.

¹⁷⁹ Ebd. S. 202.

Handlungsobjekt einzuordnen ist.¹⁸⁰ Diese Klassifizierung ist nach Loh sehr stark davon abhängig, welche Definition von *moral patiency* herangezogen wird, um ein bestehendes artifizielles System einzuordnen. Da in dieser Arbeit die Frage der Verantwortungszuschreibung zentral ist, wurden eingehend verschiedene bestehende Konzepte zu moralischer Akteursschaft [moral agency] erläutert, während die Auseinandersetzung mit der Idee moralischer Handlungsobjekte vergleichsweise kurzgehalten wurde. Dies resultiert aus der Tatsache, dass *moral patients* keine Verantwortungsträger*innen im eigentlichen Sinne sind und deren Betrachtung im Rahmen dieser Arbeit in erster Linie der Vollständigkeit dient, um den Status moralischer Handlungsobjekte hervorzuheben und hierdurch die Verantwortungszuschreibung und deren Problematik vordergründig darlegen zu können. Ungeachtet der Klassifizierung eines artifiziiellen Systems wie der Roboterrobbe Paro, wird bei genauerer Betrachtung kenntlich, dass die Problematik der Verantwortungszuschreibung in der MMI nicht erst dort beginnt, wo Roboter und Maschinen als moralische Handlungsobjekte deklariert werden, sondern bereits dort entsteht, wo unklar ist, inwiefern der *Accountability Gap*, der zwangsläufig durch die Vielfalt an Definitionen sowohl moralischer Handlungsobjekte als auch solcher Objekte entsteht, ausbalanciert werden soll.

7.1.3. Care-O-Bot und TUG

Um die Dichotomie und damit auch die Utopie der Verantwortungszuschreibung in der Therapie- Medizin- und Pflegerobotik zu verdeutlichen, sei eine dritte Form artifiziieller Systeme beleuchtet, welche eindeutig die Verantwortlichkeit moralischer Entscheidungen und Handlungen trägt, diese jedoch eigentlich aufgrund der ihr implementierten Fähigkeiten eigentlich nicht tragen dürfte. Es handelt sich hierbei um sogenannte „Pflegeroboter“ wie den Care-O-bot vom Fraunhofer IPA sowie TUG von Aethon.¹⁸¹ Beide Maschinen sind dazu befähigt, Medikamente zu transportieren, diese auszugeben und sich eigenständig sowie zielsicher zu bewegen.¹⁸² An dieser Stelle erlangt die Betrachtung der Verantwortungszuschreibung eine neue Dimension, die nicht unproblematisch ist. Stahl und Coeckelbergh beschreiben diese wie folgt:

[...] within the human–robot interaction and within the healthcare situation, there seems to be a problem: the robot is given (more) autonomy, in the sense of doing tasks by itself without human

¹⁸⁰ Vgl. Loh (2020), S. 92.

¹⁸¹ Vgl. Bendel (2018), S. 206. sowie Fraunhofer-Institut für Produktionstechnik und Automatisierung. Care O-bot 4. www.care-o-bot.de (8.7.2023) und Aethon. TUG. www.aethon.com (8.7.2023)

¹⁸² Vgl. Bendel (2018), S. 206.

intervention, but does not seem to have the capacity of moral agency: it can do all kinds of things, but unlike humans does not have the capacity to reflect on the ethical quality of what it does.¹⁸³

Diese Stellungnahme fasst die These dieser Masterarbeit bezüglich einer bestehenden Utopie der Verantwortungszuschreibung in der MMI prägnant zusammen. Insbesondere die Tatsache, dass eine rasante Weiterentwicklung bereits bestehender artifizieller Prototypen von Robotern erfolgt¹⁸⁴, die nicht nur auf den Teilbereich der Therapie- Medizin- und Pflegerobotik beschränkt ist, sondern in allen Bereichsethiken stattfindet, die sich mit künstlicher Intelligenz und damit ausgestatteten mehr oder weniger autonomen Systemen stattfindet, ist hierbei zu beachten.

Kehrt man zurück zu Care-O-Bot und TUG, ist das obige Zitat von Stahl und Coeckelbergh besonders treffend. Beide Pflegeroboter erhalten durch die ihnen anvertraute Aufgabe, Medikamente nicht nur zu transportieren, sondern auch abzugeben und für deren Einnahme zu sorgen, ein großes Maß an Verantwortung. Dieses ergibt sich schlichtweg durch das Risiko, dass falsche Medikamente verteilt oder richtige Medikamente falsch ausgegeben und/oder von Patient*innen unsachgemäß eingenommen werden. Ein Pflegeroboter ist im Gegensatz zu einer menschlichen Instanz nicht in der Lage, die Medikamenteneinnahme tatsächlich zu überblicken oder reflektiert zu betrachten. Wird demnach ein Medikament zwar verteilt, jedoch nicht eingenommen, so kann dies genauso problematisch sein wie die Einnahme eines Medikaments, welches vom Pflegeroboter irrtümlich falsch zugestellt wurde. An dieser Stelle ist demnach nicht nur zu hinterfragen, ob ein Roboter dieser Art überhaupt Verantwortung tragen kann, sondern in erster Linie steht wie bei allen artifiziellen Systemen, die mit dergleichen Aufgaben betraut werden, die Frage nach der moralischen Handlungsfähigkeit im Raum. Lediglich die Klassifizierung eines Systems als *moral agent* würde die Zuschreibung moralischer Verantwortung gewährleisten.

Betrachtet man die beiden Maschinen hinsichtlich ihrer moralischen Akteursschaft, fällt folgendes auf: Die Tätigkeiten, die von Care-O-Bot und TUG übernommen werden, erfordern ein gewisses Maß an Autonomie. Wie dieser Terminus zu verstehen ist, wurde in früheren Teilen dieser Arbeit geklärt. Um die Utopie der Verantwortungszuschreibung in der MMI anhand des konkreten gegebenen Beispiels zu begreifen, sei jedoch die zugehörige Ergänzung von Stahl und Coeckelbergh in Betracht gezogen, die insbesondere auf das Kriterium der Autonomie in Bezug auf Maschinen gerichtet ist. „Autonomy means here that

¹⁸³ Stahl, Bernd Carsten und Mark Coeckelbergh: Ethics of healthcare robotics: Towards responsible research and innovation. In: Robotics and Autonomous Systems 86, 2016, S. 153.

¹⁸⁴ Vgl. Stahl und Coeckelbergh (2016), S. 152.

the robot is designed to carry out tasks without continuous human guidance and assistance, preferably in an unstructured environment.”¹⁸⁵ Alleine durch die Forderung eines von menschlichen Instanzen unabhängigen Verhaltens, wird klar, dass die genannten Pflegeroboter ein wichtiges Kriterium moralischer Akteursschaft – nämlich Autonomie – tatsächlich nicht erfüllen. Sowohl Sullins als auch Floridi und Sanders stützen den soeben geschilderten Autonomiebegriff, wobei Sullins „Autonomie“ um einiges flexibler versteht als Floridi und Sanders oder Stahl und Coeckelbergh.

Wir erinnern uns an dieser Stelle daran, dass Sullins Autonomie als das Fehlen direkter Kontrolle einer Instanz über die Maschine beschreibt.¹⁸⁶ Hierbei stellt sich besonders bei dem aufrichtigen Versuch, artifizielle Systeme als *moral agents* zu klassifizieren, die Frage, wie flexibel “direct control”¹⁸⁷ zu verstehen ist. Grundsätzlich ist im Falle von Pflegerobotern wie dem Care-O-Bot oder TUG nicht davon auszugehen, dass das Kriterium der Autonomie erfüllt ist – unabhängig davon, welches Konzept von Autonomie herangezogen wird. Sullins, welcher Intentionalität als zweites Kriterium moralischer Akteursschaft festlegt, könnte dieses im Falle der genannten Pflegeroboter durchaus als gegeben betrachten, schließlich sind die Handlungen beider Systeme durch die von außen gegebenen Kontrollmechanismen, bspw. der Programmierung oder Implementierung bestimmter Algorithmen, zielgerichtet. Da Sullins Verantwortungsfähigkeit als drittes Charakteristikum für *moral agency* voraussetzt, Care-O-Bot und TUG jedoch maximal ein Kriterium für moralische Akteursschaft erfüllen, ist zu schlussfolgern, dass die beiden genannten Maschinen keine moralischen Handlungssubjekte sind.

Floridi und Sanders dürften dies möglicherweise anders bewerten. Wenngleich das Kriterium der Autonomie nicht erfüllt ist, verhält es sich mit den Charakteristika der Anpassungsfähigkeit und Interaktivität anders. Interaktivität setzt bei Floridi und Sanders lediglich voraus, dass ein System mit einer ihm entgegentretenden weiteren Instanz interagieren kann.¹⁸⁸ Dies ist bei Pflegerobotern zweifellos der Fall. Darüber, ob diese Systeme auch über Anpassungsfähigkeit im eigentlichen Sinne verfügen, kann disputiert werden. Da Floridi und Sanders “Adaptability” als jene Kompetenz begreifen, die es einem*einer Handelnden erlaubt, basierend auf gemachten Erfahrungen das eigene Verhalten zu ändern¹⁸⁹, ist es naheliegend, TUG und Care-O-Bot diese Fähigkeit

¹⁸⁵ Stahl und Coeckelbergh (2016), S. 153.

¹⁸⁶ Sullins (2006), S. 23.

¹⁸⁷ Ebd. S. 28.

¹⁸⁸ Vgl. Floridi und Sanders (2017), S. 325.

¹⁸⁹ Vgl. Ebd. S. 326.

abzusprechen. Beide Roboter können zwar ein bestimmtes Verhalten ausführen, ihre Lernfähigkeit im Sinne einer Anpassung eigener Handlungen basierend auf gemachten Fehlern ist jedoch nicht ohne Weiteres möglich, schließlich wurden beide Systeme vorrangig für bestimmte Zwecke programmiert und sind dadurch nicht- oder nur minimal- auf problematische und unvorhersehbare Situationen vorbereitet. Auch im Kontext von Floridi und Sanders erlangt hiermit keiner der beiden Roboter den Status eines moralischen Handlungssubjekts.

Eine Anwendung des Modells von *moral agency*, welches auf Moor basiert, ermöglicht eine Sichtweise, die der von Floridi, Sanders und Sullins entgegensteht. Anders als ebendiese geht Moor nicht von einer binären Klassifizierung aus, die Akteur*innen entweder zu moralischen Handlungssubjekten oder aber [moralischen] Handlungsobjekten macht, sondern fokussiert wie bereits in 6.2.3. erläutert, auf ein hierarchisches Modell moralischer Akteursschaft. Da Pflegeroboter sich zwar vermeintlich „autonom“ bewegen, jedoch lediglich ihnen implementierte Aufgaben übernehmen und ausführen können, wäre eine Klassifizierung ebendieser als „implizit ethische Akteur*innen“¹⁹⁰ adäquat. Als solche entsprechen sie jedoch ebenso wenig den Kriterien für tatsächliche moralische Handlungssubjekte wie bei Sullins sowie Floridi und Sanders. Bei bisher gängigen Modellen von Pflegerobotern handelt es sich gemeinhin bisher um Maschinen, die aufgrund ihrer Fähigkeiten nicht als *moral agents* gelten und daher im Grunde genommen nicht zurechnungs- und verantwortungsfähig sind. Das Problem entsteht nun genau an dieser Stelle.

Roboter und Maschinen, die eigentlich keine Verantwortung übernehmen können und dürfen, bekommen von menschlichen Instanzen Aufgaben zugewiesen, die genau diese Fähigkeit voraussetzen. Ereignen sich nun bei der Durchführung mehr oder weniger problematische Fehler, müssten zwangsläufig die mit den Maschinen arbeitenden Menschen zur Verantwortung gezogen werden, da die Maschinen selbst als moralische Handlungsobjekte keine Verantwortung übernehmen können. Anders würde es sich verhalten, wäre es möglich, ein oder mehrere Therapie- und/oder Pflegesysteme tatsächlich als moralische Handlungssubjekte zu konzipieren. In diesem Fall wäre die Verantwortungszuschreibung zwar immer noch utopisch, jedoch bei weitem weniger problematisch als es der Fall ist, wenn einem artifiziellen System, welches im Grunde genommen nicht weiß, was es tut, Aufgaben anvertraut werden, die dringend

¹⁹⁰ Vgl. Loh (2020), S. 55-57.

Reflexionsfähigkeit und kognitives Verständnis komplexer Situationen voraussetzen würden.

Beachtlich ist, dass trotz dieser Problematik die Weiterentwicklung artifizierter Systeme [nicht nur im Gesundheitsbereich] vorangetrieben wird, ohne jedoch realistische Lösungen zu entwickeln, die den Umgang mit solchen und ähnlichen Dilemmata und Utopien regeln würden. Vielmehr scheint der Diskurs rund um den bestehenden *Accountability Gap* zwar erweitert zu werden, jedoch lediglich im Sinne eines Verweises auf Probleme und Risiken, nicht jedoch im Sinne eines Anhaltens unverhältnismäßig rapider Weiterentwicklung mit dem Ziel, gängige Probleme zu lösen, bevor neue artifizielle Systeme Schwierigkeiten ungeahnten Ausmaßes mit sich bringen können.

Die Fragestellungen, die mich im Rahmen dieser Arbeit beschäftigt haben, finden ihren Niederschlag im Kontext von Verantwortung und Zurechnungsfähigkeit demnach auf unterschiedliche Art und Weise in verschiedenen Teilbereichen der Roboter- und Technikethik. Um aufzuzeigen, dass nicht nur im Bereich der Therapie- Medizin- und Pflegerobotik eine Utopie der Verantwortungszuschreibung besteht, soll ergänzend ein Exkurs in die Militärrobotik gemacht werden.

7.2. Utopische Zuschreibungen in der Militärrobotik

Die Militärrobotik, der eine Reihe an letalen Waffensystemen entsprungen sind, zählt laut Misselhorn als Bereich der Maschinenethik, der höchst kontrovers diskutiert und teils heftig kritisiert wird.¹⁹¹ Zu erwähnen ist, dass Loh in der Militärrobotik eine Überlagerung zweier Bereichsethiken sieht, die durchaus unterschiedliche Schwerpunkte haben. Einerseits gelte die Militärrobotik als Teilbereich der Roboterethik, in welcher eine Beschäftigung mit Robotern als vermeintlich moralischen Interaktionspartner*innen stattfindet, andererseits kommen gemeinhin auch Aspekte zum Tragen, die in der Militär- und Kriegsethik adressiert werden.¹⁹² Militärroboter können grundsätzlich entsprechend dem Maß an Autonomie, über das sie verfügen, klassifiziert werden. Die menschliche Instanz, die mit einem Militärroboter interagiert, ist demnach entweder “in the loop, on the loop oder out of the loop“.¹⁹³

Bei In-the-Loop-Systemen bedient ein Mensch das System und fällt sämtliche Entscheidungen, und sei es per Fernbedienung. On-the-Loop-Systeme hingegen sind zwar programmiert, sie können aber in Echtzeit unabhängig von menschlichem Eingreifen operieren. Dem Menschen obliegt jedoch weiterhin die Überwachung, und er hat jederzeit die Möglichkeit, zu intervenieren. Out-of-the-Loop-

¹⁹¹ Vgl. Misselhorn, Catrin: Grundfragen der Maschinenethik. Stuttgart: Reclam⁴ 2019, S. 155.

¹⁹² Loh (2020), S. 31.

¹⁹³ Vgl. Ebd. S. 32.

Systeme verhalten sich wie On-the-Loop-Systeme, doch besteht keine menschliche Kontroll- und Eingriffsmöglichkeit mehr.¹⁹⁴

Sowohl *In-the-Loop-Systeme* als auch *On-the-Loop-Systeme* erfordern daher ein hohes Maß an Kontrolle, das zwar in Hinblick auf das Risiko, das mit dem Einsatz militärischer Roboter einhergeht, beruhigend ist, jedoch für Kriegszwecke den Nachteil hat, dass diese Roboter auf die Kommunikation zwischen sich selbst und Menschen angewiesen sind und hierdurch bei weitem langsamer in ihrer Entscheidungsfindung sind, als dies bei vollautonomen Waffensystemen der Fall ist.¹⁹⁵ Je langsamer die Entscheidungen getroffen werden, umso größer ist jedoch auch die Chance, „vom Feind entdeckt, gestört oder außer Gefecht gesetzt zu werden.“¹⁹⁶ *Out-of-the-Loop-Systeme* hingegen sind völlig unabhängig von menschlichen Interaktionspartner*innen. Sie sind aufgrund der ihnen implementierten Algorithmen und kontinuierlichen Trainings in der Lage, Entscheidungen in einer derartigen Schnelligkeit zu treffen, die die menschlich-kognitiven Fähigkeiten weit übersteigt.¹⁹⁷

Hochproblematisch ist hierbei unter anderem, dass das Training solcher Systeme wie bei allen Robotern durch Daten erfolgt, die von menschlichen Instanzen gewählt, oder von Algorithmen selektiert wurden, die jedoch ihrerseits nur selektieren können, weil sie zuvor ebenfalls von Menschen programmiert wurden.¹⁹⁸ Dass Menschen nicht unfehlbar sind und daher auch von ihnen durchgeführte Programmierungen Fehler und Vorurteile beinhalten können, steht wohl außer Frage. Komplexe und riskante Situationen ergeben sich in jedem Moment, in dem ein autonomes Waffensystem aufgrund seiner Programmierung die Entscheidung darüber trifft, wer als Feind kategorisiert und bekämpft werden soll. Diese Entscheidungen – sowohl jene der Klassifizierung- als auch der Vernichtung – obliegen im Falle von *Out-of-the-Loop-Systemen* einzig und allein den jeweiligen Maschinen. Wenngleich sie von Menschen trainiert wurden, können ebendiese Menschen nun nicht mehr in Entscheidungen und Handlungen ihrer Waffensysteme eingreifen.

Stellt man sich die Frage, was dies für die Verantwortungszuschreibung in der MMI bedeutet, ergeben sich mehrere problematische Aspekte. Einerseits handelt es sich bei Militärrobotern grundsätzlich um Systeme, deren Umgang mit Verantwortung ein anderes Ausmaß erfordert als dies beispielsweise bei Therapie- und Pflegerobotern der Fall ist. Dies ergibt sich schlichtweg durch die Tatsache, dass die Entscheidungen, die von militärischen

¹⁹⁴ Misselhorn (2019), S. 158.

¹⁹⁵ Ebd. S. 158.

¹⁹⁶ Ebd. S. 158.

¹⁹⁷ Vgl. Ebd. S. 158.

¹⁹⁸ Vgl. Loh (2020), S. 33.

Systemen getroffen werden, den sicheren Tod für das als Feind klassifizierte menschliche Gegenüber bedeuten können. Abgesehen davon, dass es höchst problematisch ist, solche Entscheidungen artifiziellen Systemen zu überlassen, stellt sich im selben Zuge die Frage danach, ob die als autonom deklarierten Systeme tatsächlich verantwortungs- und zurechnungsfähig sind. Dass *Out-of-the-Loop-Systeme* autonom fungieren können, bedeutet noch lange nicht, dass sie im Falle von Fehlentscheidungen tatsächlich zur Verantwortung gezogen werden können.

Genau an dieser Stelle beginnt im Bereich militärischer Systeme die Utopie der Verantwortungszuschreibung ein Ausmaß anzunehmen, welches problematischer nicht sein könnte. Gegeben ist eine Reihe an mehr oder weniger kontrollierbaren Waffensystemen, deren Aufgabe es ist, militärische Zwecke zum Vorteil bestimmter menschlicher Instanzen zu erfüllen. Solange Menschen nun die von ihnen trainierten Systeme überwachen, wie dies bei *In-the-Loop-* und *On-the-Loop-Systemen* der Fall ist, kann Verantwortlichkeit vermeintlich stets der menschlichen Instanz überschrieben werden, da sie es ist, die den Roboter steuert, überwacht und kontrolliert. Im Falle von vollautonomen Kriegssystemen, die selbstständig entscheiden, ein Gegenüber zu töten oder zu verschonen, besteht jedoch keinerlei Möglichkeit, das Verantwortungsproblem zu lösen. Es handelt sich in diesem Fall schlichtweg um eine bestehende Lücke, die zum jetzigen Zeitpunkt der Entwicklungen offensichtlich nicht mehr bewältigt und geschlossen werden kann. Ein autonomes System, das in der Lage ist, losgelöst von menschlichen Instanzen Entscheidungen zu treffen und danach zu handeln, klassifiziert sich nach den bisher dargelegten Konzepten moralischer Akteursschaft eindeutig als *moral agent* – als moralisches Handlungssubjekt. Als solches trägt es zwangsläufig Verantwortung für das von ihm durchgesetzte Handeln. Fraglich ist nun, ob dies überhaupt möglich ist. Das autonome System wurde wie bereits erwähnt von Menschen designt, programmiert und trainiert und führt exakt dasjenige aus, was Menschen zuvor als Handlungsmöglichkeiten implementiert haben. Kann ein solches System nun tatsächlich für dasjenige verantwortlich gemacht werden, was es tut? Ist es zulässig, aufgrund bestehender Konzepte zu moralischer Akteursschaft anzunehmen, dass ein System tatsächlich zurechnungsfähig ist, weil es menschengemachte Kriterien erfüllt, die es vermeintlich zu einem Handlungssubjekt machen, welches moralische Entscheidungen treffen kann und darf? Ich meine, dass dies ein Fehlschluss ist. Gerade im Bereich der Militärrobotik existieren Roboter, die nicht als moralische Handlungssubjekte klassifiziert werden können, weil ihnen hierfür schlichtweg die nötigen Charakteristika fehlen. In diesem Fall besteht die Utopie der Verantwortungszuschreibung darin, dass diese Maschinen

dennoch für Zwecke genutzt werden, die ein hohes Maß an Verantwortung voraussetzen würden. Da sie von Menschen gesteuert und kontrolliert werden, ergeben sich im Fall nicht-autonomer Systeme ähnliche Probleme wie sie zuvor im Kontext der Therapie- Medizin und Pflegerobotik erläutert wurden.

Handelt es sich nun bei Militärrobotern allerdings um autonome Systeme, so gelten sie als moralische Handlungssubjekte und müssten in der Lage sein, die Verantwortung für die von ihnen getroffenen Entscheidungen und Handlungen zu übernehmen. Doch wie kann ein artifizielles System Verantwortung übernehmen, wenn dasjenige, was es tut, oder getan hat nicht das ist, was seinem eigenen „Denken“ entspringt? Da es sich bei artifiziiellen Systemen um menschlich designte Instanzen handelt, bleibt der bereits mehrfach geschilderte *Accountability Gap* nicht nur bestehen, sondern vergrößert sich insbesondere durch die Ernsthaftigkeit jener Entscheidungen, die Militär- und Kriegeroboter treffen müssen, um ein Vielfaches. Nicht nur utopisch, sondern vielmehr absurd ist es, dass eine Reihe an Konzepten und Ideen zu moralischer Akteursschaft und Verantwortung besteht, jedoch noch kein Konzept entwickelt wurde, welches sich tatsächlich auf mehr oder weniger autonome Systeme anwenden lässt. Wie Hans Jonas treffend formuliert:

Aus der erweiterten Zukunftsdimension heutiger Verantwortung ergibt sich das abschließende Thema: die Utopie. Die weltweite technologische Fortschrittsdynamik birgt als solche einen impliziten Utopismus in sich, der Tendenz, wenn nicht dem Programm nach. Und die eine schon existierende Ethik mit globaler Zukunftssicht, der Marxismus, hat eben im Bunde mit der Technik die Utopie zum ausdrücklichen Ziel erhoben.¹⁹⁹

Genau diese Utopie darf, soll und muss m.E. im Philosophie- und/oder Ethikunterricht der Sekundarstufe II besprochen werden. Es ist mehr als relevant, die junge Generation, die sich in den jeweiligen Berufsfeldern maßgeblich an der Weiterentwicklung artifiziieller Systeme beteiligen wird, bereits im Schulalter darauf hinzuweisen, welche Unstimmigkeiten, Gefahren und Probleme die vermeintlich mögliche Verantwortungszuschreibung in der MMI beinhaltet.

8. Fachdidaktische Erläuterungen

8.1. Kontextualisierung

Die aktuell geltende Fassung des Lehrplans der allgemeinbildenden höheren Schulen, welche vom Rechtsinformationssystem des Bundes (RIS) zur Verfügung gestellt wird, umfasst neben einzelnen fachspezifischen Lehrplänen im ersten Teil auch allgemeine

¹⁹⁹ Jonas und Habeck (2020), S. 9.

Bildungsziele, die ihrerseits in Bildungsbereiche gegliedert sind. Für die vorliegende Arbeit ist unter Anbetracht der behandelten Inhalte insbesondere eine fachdidaktische Betrachtung im Kontext des Bereichs „Natur und Technik“ interessant. Dessen Bildungsziel ist wie folgt definiert: „Der Unterricht hat [...] grundlegendes Wissen, Entscheidungsfähigkeit und Handlungskompetenz zu vermitteln. Die Schülerinnen und Schüler sind zu befähigen, sich mit Wertvorstellungen und ethischen Fragen im Zusammenhang mit Natur und Technik sowie Mensch und Umwelt auseinander zu setzen.“²⁰⁰ Auch der gesetzliche Auftrag, der an Lehrpersonen herangetragen wird, enthält ähnliche Forderungen:

Die allgemein bildende [sic!] höhere Schule hat im Sinne des § 2 des Schulorganisationsgesetzes an der Heranbildung der jungen Menschen mitzuwirken, nämlich beim Erwerb von Wissen, bei der Entwicklung von Kompetenzen und bei der Vermittlung von Werten. Dabei ist die Bereitschaft zum selbstständigen Denken und zur kritischen Reflexion besonders zu fördern.²⁰¹

Neben der Vermittlung von Wissen scheinen demnach auch Kompetenzerwerb und Förderung der kritischen Auseinandersetzung mit erworbenem Wissen eine zentrale Rolle bei der erfolgreichen Ausbildung junger Menschen zu spielen. Dass bereits im allgemeinen Teil des Lehrplans das Augenmerk in diesem Zusammenhang auf der Beschäftigung mit ethischen Fragestellungen und Wertevermittlung liegt, eignet sich, um die Problematik der Verantwortungszuschreibung in der MMI im Rahmen einer Unterrichtsreihe philosophisch zu erarbeiten.

Im Lehrplan der AHS scheint häufig der Begriff „Kompetenz“ auf. Wie bereits viele weitere sprachliche Konstrukte, die in dieser Arbeit abgehandelt wurden, ist auch dieser Begriff vielfältig interpretierbar. In Anlehnung an die Definitionen von Klieme und Rösch lässt sich Kompetenz als Synonym für eine zu erwerbende Fähigkeit, Fertigkeit oder fachspezifisches Können einsetzen.²⁰² Diesem Verständnis nach erscheint es nachvollziehbar, warum das Streben nach kompetenzorientiertem Unterricht derart groß ist und auch in den Lehrplänen der AHS seinen Niederschlag findet. Gelingt es einer Lehrperson, Schüler*innen nicht nur theoretisches Wissen, sondern auch das zugehörige Knowhow zu vermitteln, wirkt dies für Außenstehende – unter anderem Eltern – vermutlich wie eine Bestätigung der erfolgreichen Umsetzung des jeweiligen Unterrichts. Das Erwerben von Kompetenzen scheint nach außen

²⁰⁰ Vgl. Rechtsinformation des Bundes: Lehrplan der allgemeinbildenden höheren Schulen. <https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=10008568&FassungVom=2017-08-31> (13.3.2023)

²⁰¹ Ebd. (13.3.2023)

²⁰² Siehe u.a.: Klieme, Eckhard und Hermann Avenarius u.a.: Zur Entwicklung nationaler Bildungsstandards: eine Expertise. Bonn, Berlin: Bundesministerium für Bildung und Forschung 2009. sowie Rösch, Anita: Kompetenzorientierung im Philosophie- und Ethikunterricht: Entwicklung eines Kompetenzmodells für die Fächergruppe Philosophie, Praktische Philosophie, Ethik, Werte und Normen, LER. Berlin: LIT 2009.

hin messbar zu sein. Fraglich ist in diesem Kontext allerdings, ob hierdurch dem Weg des Kompetenzerwerbs für die gewünschte „Vermittlung von Werten“, die „Bereitschaft zum selbstständigen Denken und zur kritischen Reflexion“²⁰³ nicht zu wenig Bedeutung beigemessen wird. Schließlich ist das Verfügen über eine Kompetenz nicht gleichbedeutend mit der Fähigkeit, diese kritisch einsetzen und nutzen zu können.

Ein meine Gedanken fortführender Ansatz kann bei Carsten Roeger gefunden werden. Dieser weist darauf hin, dass die Frage nach dem Kompetenzerwerb im Philosophieunterricht anders ausgelegt oder gar beantwortet werden müsse als in anderen Unterrichtsgegenständen und postuliert, das eigentliche Ziel des Philosophieunterrichts sei das Philosophieren selbst, nicht der Erwerb bestimmter Kompetenzen.²⁰⁴ Roeger argumentiert hierbei wie folgt: „Authentisches selbstständiges Philosophieren liegt nur dann vor, wenn noch [...] keine feststehende Definition für den Unterricht angelegt wird, welche dann lediglich nur hergeleitet, nachvollzogen oder angewandt wird.“²⁰⁵

Die Bestrebung Roegers, Unterricht so zu strukturieren, dass das selbstständige Denken ohne jedwede direktive Führung durch die Lehrperson angeregt wird und tatsächlich komplexe inhaltlich korrekte Konzepte und Definitionen hervorbringt, scheint auf den ersten Blick etwas weit hergeholt. Es wird keineswegs in Frage gestellt, dass Schüler*innen der Sekundarstufe II dazu aufgefordert werden können und sollen, eigenständige Gedankengänge formulieren und nachvollziehen zu lernen. Dass dies allerdings ohne ein vorgefertigtes Lehr- und Lernkonzept möglich ist, ist nicht nur aufgrund der sehr geringen Zeitspanne, die für tatsächlichen Philosophieunterricht wöchentlich zur Verfügung steht, zu bezweifeln, sondern scheint auch in Anbetracht der Tatsache schwierig, dass Jugendliche, selbst wenn sie für ihr Alter ein besonders hohes Abstraktionsniveau erreicht haben oder über großes Fachwissen verfügen, in den meisten Fällen dennoch die Anleitung und Unterstützung einer Lehrperson benötigen dürften, um die eigenen Überlegungen philosophisch aufbereiten zu können.

Selbst wenn deutlich mehr zeitliche Ressourcen für die philosophische Bildung junger Menschen Teil des Curriculums allgemeinbildender höherer Schulen wären, ist in Frage zu stellen, inwieweit diese Ressourcen für „authentisches Philosophieren“ genutzt werden

²⁰³ Rechtsinformation des Bundes: Lehrplan der allgemeinbildenden höheren Schulen. <https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=10008568&FassungVom=2017-08-31> (13.3.2023)

²⁰⁴ Vgl. Roeger, Carsten: Philosophieunterricht zwischen Kompetenzorientierung und philosophischer Bildung 1(2). Opladen: Verlag Barbara Budrich, 2016, S. 183.

²⁰⁵ Roeger (2016), S. 193.

könnten, ohne hierbei die Lehraufträge und Ziele, die im Rahmen des Lehrplans festgelegt sind, sowie die hiermit verbundene Themenvielfalt in den Hintergrund zu rücken und die Vermittlung von Wissen zu vernachlässigen. Dennoch beinhaltet Roegers Aussage einen relevanten Aspekt in Bezug auf die von mir postulierte These, der Weg hin zum Kompetenzerwerb müsse stärker beachtet werden, stellt er doch seinerseits die These in den Raum, das Philosophieren selbst sei ein zentraler Baustein in der Entwicklung bestimmter Fertigkeiten.

In Bezug auf die Auseinandersetzung mit verschiedenen Aspekten der Roboterethik und der KI-Forschung, insbesondere deren Zusammenhang zu Verantwortungsdiffusion und Verantwortungszuschreibung reicht nach eigenem Ermessen ein eindimensionaler Zugang zur Thematik nicht aus, um Schüler*innen der Sekundarstufe II ein möglichst komplexes Verständnis der vorliegenden Problematik zu ermöglichen. Aus diesem Grund orientieren sich die in Kapitel 6.2.1 und 6.2.3. abgedruckten Stundenbilder an einer Kombination unterschiedlicher Herangehensweisen, die in Bezug auf die fachdidaktische Umsetzung komplexer Themenfelder anwendbar scheinen. Dies ist idealerweise als Versuch zu werten, Philosophieunterricht neu zu denken oder zumindest nicht in festgefahrenen Strukturen zu verharren, die der Entwicklung junger Menschen nicht dienlich sind. Interessanterweise hält die Beschäftigung mit KI/AI, MMI und unterschiedlichen Erscheinungsformen bestimmter Maschinen und Roboter auch in die Lehrbücher des Unterrichtsfachs Deutsch Einzug.²⁰⁶ Dies ist unter anderem erwähnenswert, weil das Unterrichtsfach Deutsch nicht zwangsläufig mit philosophischen Fragestellungen oder Themenbereichen assoziiert wird, diese allerdings mittlerweile beinahe selbstverständlich in bestimmte Aufgabenkomplexe des Deutschunterrichts mit einfließen und dadurch die Schüler*innen in gewisser Weise automatisch dazu anregen, eigenständig zu philosophieren.

In Zusammenhang mit Utopie und Dystopie wird Schüler*innen der Sekundärstufe II im Deutschbuch „Wortwechsel 4“ beispielsweise eine Art philosophische Denkübung abverlangt, die sich genauso gut in einem Lehrbuch des Unterrichtsfachs Philosophie abbilden ließe.²⁰⁷ Wenngleich sie dort vermutlich auf den Erwerb anderer Sach- und Fachkompetenzen abzielen würde, so ist die Aufforderung zur eigenständigen Auseinandersetzung mit philosophisch komplexen Sachverhalten m.E. positiv zu bewerten. MMI und Robotik scheinen mittlerweile offenbar zu zentralem gesellschaftlichen Status

²⁰⁶ Vgl. Mayer-Steflic, Karin und Verena Toschner u.a.: Wortwechsel 4. Sprachbuch für die 8. Klasse AHS. Linz: Veritas 2019, S. 37.

²⁰⁷ Vgl. Mayer-Steflic u.a. (2019), S. 37.

gelangt zu sein, der längst nicht mehr nur in den Bereichen der Philosophie und Ethik anzutreffen ist.

An dieser Stelle lässt sich eine Querverbindung zu Klafki schaffen, welcher im Zuge der „Konzentration auf epochaltypische Schlüsselprobleme unserer Gegenwart und der vermutlichen Zukunft“²⁰⁸ unter anderem auch auf die „Gefahren und die Möglichkeiten der neuen technischen Steuerungs-, Informations- und Kommunikationsmedien“ hinweist.²⁰⁹ Unterricht und dessen Struktur sollen nach Klafki demnach idealerweise ebenjenen Problemfeldern gewidmet sein, die gesellschaftlich von uneingeschränkter Relevanz sind.²¹⁰ Klafki rechtfertigt dies, indem er auf die dringende Aufgabe von Bildungseinrichtungen verweist, „Aufklärung über noch nicht Bewußtgewordenes [sic!], noch Unbefragtes [sic!], noch vermeintlich Selbstverständliches zu leisten, sofern daraus Folgen für heutiges und späteres Handeln, vor allem Folgen für andere Menschen resultieren oder resultieren können.“²¹¹ In dieser Forderung ähnelt Klafkis Konzept in seinem Kernziel jenem von Dewey, welcher den Begriff des problemorientierten Lehrens und Lernens [POL] geprägt hat. Auch hier stehen relevante gesellschaftliche Themen im Fokus des Unterrichts, „die in Kleingruppen mit tutorieller Unterstützung gelöst werden mit dem Ziel, transferfähiges Wissen und fachspezifische Lern- und Denkstrategien zu erwerben.“²¹²

Die im Rahmen dieser Arbeit geplante Unterrichtsreihe versucht weder dem klassischen Zugang der Kompetenzorientierung noch dem problemorientierten Zugang einen Vorzug zu geben, sondern beide Ansätze bestenfalls miteinander zu kombinieren, um den höchstmöglichen Lernzuwachs für Schüler*innen generieren zu können.

8.2. Unterrichtsplanung

Die folgenden vier Unterrichtsstunden sind für die Durchführung in einer allgemeinbildenden höheren Schule im Unterrichtsfach [Psychologie,] Philosophie konzipiert und können sowohl für eine achte Klasse im regulären Fachunterricht als auch für das Wahlpflichtfach Philosophie implementiert werden. In Anbetracht der Erfahrungen, die im Rahmen eigener Unterrichtstätigkeit an einer bilingualen AHS gemacht wurden, eignet

²⁰⁸ Klafki, Wolfgang: Allgemeinbildung in der Grundschule und der Bildungsauftrag des Sachunterrichts. In: Widerstreit Sachunterricht 4 (2005), S. 4.

²⁰⁹ Klafki (2005), S. 4.

²¹⁰ Vgl. Lehner, Michael und Inga Gryl u.a.: Problemorientierter Unterricht ≠ Problemorientierter Unterricht. In: GW-Unterricht 1 (2021), S. 41.

²¹¹ Klafki (2005), S. 5.

²¹² Reusser, Kurt: Problemorientiertes Lernen.- Tiefenstruktur, Gestaltungsformen, Wirkung. In: Beiträge zur Lehrerbildung 23(2) 2005, S. 159.

sich eine Zeitspanne von sechs bis acht Unterrichtsstunden, um einen philosophisch-ethischen Themenkomplex ausführlich zu erarbeiten. Da die vorliegende Arbeit allerdings vorwiegend theoretisch konzeptioniert ist, werden im Folgenden lediglich zwei Stundenplanungen erläutert, um einen Überblick darüber zu geben, inwiefern die bestehende Theorie in die Praxis transferiert werden kann.

Die folgenden Stundenkonzepte wurden basierend auf der Annahme zweier Doppelstunden zu je 100 Minuten erstellt, können jedoch nach eigenem Ermessen sowohl gekürzt als auch beliebig um etwaige erwünschte Aspekte ergänzt werden. Der Lehrplanbezug lässt sich dem 2. Semester der 8. Klasse – insbesondere Modul 2 „Grundfragen der Ethik“²¹³ – zuordnen.

Im Rahmen meiner finalen Praxisphase des Masterstudiums Lehramt fand sowohl mit Schüler*innen als auch mit betreuenden Lehrpersonen reger Austausch über die philosophischen Interessen und Wünsche der schulinternen Jugendlichen statt. Hierbei kristallisierte sich insbesondere heraus, dass das Interesse an künstlicher Intelligenz und Robotik, sowie jenen Problemfeldern, die hiermit in Zusammenhang stehen, im Regelunterricht sowie im Wahlpflichtfach hoch war. Unter anderem aus diesem Grund stieß der Vorschlag, Unterrichtsstunden zu diesen Themenbereichen abzuhalten, bei den betreuenden Mentor*innen und auch Schüler*innen auf große Zustimmung. Roegers These, dass die Neugier junger Menschen zentrale Bedingung für einen erfolgreichen Lernprozess ist²¹⁴, dient als Ausgangspunkt dieser fachdidaktischen Auseinandersetzung. Während die bereits im schulischen Kontext durchgeführte Unterrichtsreihe auf eine intensive Auseinandersetzung mit Robotik, Roboter- und Maschinenethik abzielte, sollen die nun folgenden neu konzipierten Unterrichtseinheiten den Aspekt der Verantwortung in den Fokus rücken, da dieser einige Schüler*innen sehr interessierte, die zur Verfügung gestellte Unterrichtszeit jedoch nicht ausreichte, um überhaupt einen ersten Schritt in Richtung einer komplexen Verknüpfung bereits erlernter Inhalte mit der Frage nach der Verantwortungszuschreibung in der MMI zu machen.

Für die folgenden Stundenkonzeptionen wird demnach vorausgesetzt, dass Schüler*innen bereits mit grundlegenden Begriffen und Konzepten der Roboter- und Maschinenethik bekannt gemacht wurden und ein zentrales Verständnis für die Klassifizierung artifizieller Systeme als moralische Handlungsträger*innen besitzen.

²¹³ Rechtsinformation des Bundes: Lehrplan der allgemeinbildenden höheren Schulen. <https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=10008568&FassungVom=2017-08-31> (13.3.2023)

²¹⁴ Vgl. Roeger (2016), S. 186.

8.2.1. Stundenbild 1

Zeit	Phase	Inhalte	Methode
10 Min.	Einstieg	Mentimeter → Schüler*innen verschriftlichen Assoziationen zum Begriff „Verantwortung“	Mentimeter
20 Min.	Hinführung 1	- Lehrperson erarbeitet im Plenum die genannten Begriffe mit den SuS²¹⁵ - etwaige Fragen werden beantwortet - Unklarheiten werden geklärt - SuS bekommen Theorie-Input seitens der LP²¹⁶	LSG
10 Min.	Hinführung 2	SuS werden in Kleingruppen aufgeteilt und versuchen eine eigene Definition von Verantwortung auszuarbeiten	GA
15 Min.	Problematisierung	- SuS stellen ihre Definitionen vor - Lehrperson und SuS besprechen Unterschiede und Gemeinsamkeiten	Frontal LSG
20 Min.	Transfer	Jede Kleingruppe erhält die Definition einer anderen Gruppe und bespricht anhand dieser jene Fallbeispiele, die von der LP zur Verfügung gestellt werden	GA
25 Min.	Sicherung	SuS und Lehrperson gehen gemeinsam die Vignetten durch und verschriftlichen Ergebnisse sowie Beobachtungen	Plenumsdiskussion

8.2.2. Themenspezifischer Bezug

Ein Verständnis des Verantwortungsbegriffs ist auch ohne dessen Bezug zur MMI bereits ein hochkomplexes Unterfangen. Aus diesem Grund soll eine Annäherung im Sinne des in Kap. 6.1. „Kontextualisierung“ erläuterten Ansatzes von Roeger stattfinden, indem der Versuch gewagt wird, Schüler*innen „tatsächliches Philosophieren“ zu ermöglichen. Das

²¹⁵Anmerkung: Wenn im Rahmen dieser Arbeit in konzipierten Stundenbildern die Abkürzung „SuS“ vorkommt, so ist diese stets als genderinklusive Bezeichnung für Schüler*innen jedweden Geschlechts zu lesen.

²¹⁶ Anmerkung: „LP“ steht stellvertretend als genderinklusive Bezeichnung für Lehrende jedweden Geschlechts

konzipierte Stundenbild ist hierfür wie folgt umzusetzen: Nach einer Begrüßung der Schüler*innen, der Klärung etwaiger organisatorischer Angelegenheiten sowie der Erläuterung des Stundenfahrplans, sind die Schüler*innen dazu aufgefordert, mittels *Mentimeter* ihre Assoziationen zum Begriff „Verantwortung“ zu verschriftlichen, ohne dass der Terminus vorher bereits detailliert im Unterricht besprochen wurde. So soll eine möglichst unvoreingenommene erste Auseinandersetzung mit einem philosophisch sehr komplexen Konzept gewährleistet werden. Im Anschluss an diese erste Gedankensammlung werden die Ergebnisse des *Mentimeter* mit den Jugendlichen besprochen. Hierbei wird im Idealfall bereits eine Richtung sichtbar, hinsichtlich derer die Schüler*innen Verantwortung begreifen. Die Lehrperson steht den Jugendlichen an dieser Stelle als fachliche*r Dialogpartner*in gegenüber, der*die Fragen und Unklarheiten konstruktiv lösen und den Schüler*innen ein gutes Basiswissen vermitteln kann.²¹⁷

Tatsächlich ist die Rolle der Lehrperson(en) im Rahmen des Unterrichts nicht unumstritten, was sich bereits in den Ausführungen zu kompetenz- und/oder problemorientiertem Unterricht gezeigt hat. Grundsätzlich wird erwartet, dass Lehrpersonen sich „[...] auf ständig wechselnde Bedingungen einstellen und adaptiv handeln können, und dies in einem Feld, wo es selten eine richtige Lösung, sondern i.d.R. mehrere Alternativen gibt [...]“²¹⁸ Wie diese flexible Handlungsfähigkeit allerdings auszusehen hat, ist nicht immer klar. Collings, Brown und Newman haben in diesem Kontext den Ansatz des “Cognitive Apprenticeship” geprägt, in welchem die unterschiedlichen Möglichkeiten der Lehrperson, sich konstruktiv in den Lernprozess der Schüler*innen zu integrieren, aufgezeigt werden.²¹⁹ Hierauf wird in den folgenden Erläuterungen des Öfteren eingegangen, um ein möglichst vollständiges Bild der zu haltenden Stunden vermitteln zu können.

Insbesondere wenn im vorliegenden Stundenbild von Theorieinput seitens der Lehrperson die Rede ist, ist es essenziell, zu betonen, dass die Schüler*innen durch den Theorieinput der Lehrperson keineswegs in ihrer gedanklichen Vielfalt eingeschränkt werden sollen. Vielmehr soll der Diskurs um Verantwortung geöffnet und erweitert werden, um einen authentischen philosophischen Prozess anzuregen. Da sich gezeigt hat, dass für viele Schüler*innen Gruppenarbeiten ertragreich sind, wird die Klassengemeinschaft in Kleingruppen zu je drei bis vier Personen aufgeteilt. Jede Gruppe erhält denselben

²¹⁷ Vgl. Reusser (2005), S. 170.

²¹⁸ Ebd. S. 175.

²¹⁹ Vgl. Collins, Allan und John Seely Brown u.a.: Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In: Resnick, Lauren B. (Hrsg.): Knowing, learning, and instruction: Essays in the honour of Robert Glaser. Hillsdale, New Jersey: Erlbaum 1989, S. 453–494. sowie Reusser (2005), S. 170.

Arbeitsauftrag – den Terminus „Verantwortung“ basierend auf den bisher diskutierten Inhalten und erworbenem Wissen bestmöglich zu definieren. Die Lehrperson steht den Schüler*innen hierbei lediglich assistierend zur Seite und dient als „personales Lerngerüst“.²²⁰ Auch bei der Besprechung der Erarbeitungen ist die Rolle der Lehrperson beobachtend. Werturteile oder gar Beurteilungen sind an dieser Stelle nicht förderlich. Ein gemeinsamer kritischer Austausch über die unterschiedlichen Definitionen und deren Tragweite ist hingegen erwünscht und gewollt.

Nachdem Definitionen besprochen und möglicherweise einige Gemeinsamkeiten und Unterschiede festgestellt wurden, wird in einem weiteren Schritt die Definition jeder Gruppe an eine andere Gruppe weitergereicht. Da die Arbeit mit Vignetten eine der gängigen „Kleinformen des problemorientierten Unterrichts“²²¹ ist, wird in der vorliegenden Unterrichtssequenz genau hiermit gearbeitet, um die Lernenden mittels realitätsnaher und deren Lebenswelt entsprechenden Fallbeispielen „zur Bearbeitung von «hochwertigen» fachlichen Problemen“²²² zu motivieren. Sämtliche Kleingruppen erhalten hierfür ein Handout mit Vignetten, auf die die jeweilige Verantwortungsdefinition bestmöglich angewandt werden soll. Dieser Schritt dient bereits der Vorbereitung auf die zweite Doppelstunde zum Themenkomplex der Verantwortungszuschreibung. Relevant ist insbesondere, dass die ausgewählten und zur Verfügung gestellten Vignetten lediglich Situationen abbilden, die zwischen menschlichen Parteien stattfinden. In der Unterrichtseinheit, die in Kapitel 6.2.3. abgebildet ist, wird hingegen auf Beispiele eingegangen, in denen Verantwortlichkeit in einem größeren Kontext und nicht lediglich auf menschliche Individuen beschränkt diskutiert wird. All jenes, was im Lehrer*innen-Schüler*innen-Gespräch zur Sprache kommt, soll eigenständig verschriftlicht werden und als Ausgangspunkt für die kommende Unterrichtssequenz genutzt werden.

8.2.3. Stundenbild 2

Zeit	Phase	Inhalte	Methode
10 Min.	Einstieg	Wiederholung der vergangenen Unterrichtseinheit	Frontal

²²⁰ Reusser (2005), S. 170.

²²¹ Ebd. S. 171.

²²² Ebd. S. 171.

20 Min.	Hinführung 1	Theorieinput seitens der LP zu Unterschieden und Differenzen von Zurechenbarkeit und Verantwortung	Frontal
10 Min.	Hinführung 2	Überarbeitung der in der vorherigen Stunde verfassten Definitionen	GA Plenum
30 Min.	Problematisierung und Transfer	Vignettenbetrachtung mit Schwerpunkt auf artifiziellen Systemen – Utopie der Verantwortungszuschreibung?	LSG Plenum
20 Min.	Transfer	SuS versuchen mögliche Ansätze zur Lösung der bestehenden Problematik zu finden und verschriftlichen diese im Rahmen einer Partner*innenarbeit	PA
10 Min.	Sicherung	Konzepte werden vorgestellt und diskutiert	Präsentation

8.2.4. Themenspezifischer Bezug

Stundenbild 2 steht in unmittelbarem Zusammenhang mit Stundenbild 1. Für die Umsetzung in der Sekundarstufe II ist es relevant, dass die im Rahmen der ersten Doppelstunde durchgeführte problemorientierte Auseinandersetzung stattgefunden hat und abgeschlossen wurde. Im Gegensatz zu Stundenbild 1 ist die vorliegende Unterrichtssequenz nun nicht nur problemorientiert, sondern versucht den kompetenzorientierten Zugang Klafkis mit jenem von Dewey zu kombinieren, indem ein ausgewogenes Verhältnis zwischen zu erwerbenden Kompetenzen und eigenständigem Philosophieren angestrebt wird. Was dies für die Umsetzung im Unterricht bedeutet, lässt sich wie folgt situieren:

Kompetenzorientierung als eine Form der Lernzielorientierung ist mit dem zweckrationalen Ansatz vereinbar, Unterricht vom Ende her zu planen. Man legt nach Hilbert Meyer (1993, S. 151) das gewünschte Ergebnis fest und plant anschließend den Weg dorthin, unter Berücksichtigung der zur Verfügung stehenden Ressourcen, wie Unterrichtszeit, Lerngruppe, Ausstattung oder die fachlichen Fertigkeiten und das Wissen der Lehrerin und des Lehrers.²²³

Im Gegensatz zu Stundenbild 1 ist nun also tatsächlich das gewünschte Endziel des Unterrichts bereits vorhanden. So sollen die Schüler*innen dazu befähigt werden, in der Verantwortungszuschreibung im Rahmen der Mensch-Maschine-Interaktion eine Utopie auszumachen und diese in ihren Kernaspekten zu begreifen. Das Ziel, dass Schüler*innen die Utopie der Verantwortungszuschreibung erkennen, ist also bereits zu Beginn des

²²³ Roeger (2016), S. 205.

Unterrichts vorgegeben, wird allerdings von der Lehrperson nicht als solches kommuniziert. Auf diese Weise ergibt sich für Schüler*innen idealerweise die Möglichkeit, durch eine Kombination problemorientierten Lernens und Kompetenzorientierung ein tiefgreifendes Verständnis für die behandelte Materie zu erlangen.²²⁴

Zu Beginn der vorliegenden Unterrichtssequenz sollen die erlernten Inhalte zunächst mündlich wiederholt werden, um das Vorwissen der Schüler*innen zu reaktivieren. Im Anschluss findet ein Vortrag seitens der Lehrperson statt, der nun anders als in Einheit 1 eine Reihe an theoretischen Aspekten beinhaltet, die für den späteren Kompetenzerwerb relevant sind. Insbesondere die Unterschiede und Gemeinsamkeiten von Zurechenbarkeit und Verantwortung werden angesprochen und schüler*innengerecht veranschaulicht. Für die Lehrperson ergibt sich in diesem Kontext die Möglichkeit, frei aus den in früheren Teilen dieser Arbeit geschilderten Verantwortungskonzepten zu wählen. Es wurde vorsätzlich nicht definiert, welche Formen der Verantwortung genauer erläutert oder miteinander verglichen werden sollen. Je nach der Heterogenität der Klassengemeinschaft ist von der Lehrperson abzuwägen, ob die Möglichkeit besteht, Schüler*innen mit unterschiedlichen Verantwortungskonzepten vertraut zu machen, oder ob lediglich eine Art der Verantwortung in Kombination mit deren Gemeinsamkeiten und Differenzen zu Zurechnung beleuchtet werden soll. An dieser Stelle ist in jedem Fall bereits eine direktive Führung durch die Lehrperson erkennbar. Dies dient der gezielten Hinführung der Schüler*innen an die Komplexität der vorliegenden Problematik.

Basierend auf der veranschaulichten Theorie haben Schüler*innen im Anschluss die Möglichkeit, Fragen zu stellen und Unklarheiten dazulegen. Sind diese geklärt, werden die Jugendlichen in dieselben Gruppen wie in der vergangenen Einheit geteilt und dazu ersucht, die in dieser Konstellation ausgearbeiteten Definitionen von Verantwortung unter Bezugnahme auf die vermittelte Theorie zu überarbeiten. An diesem Punkt erhalten die Schüler*innen auch bereits ein neues Handout mit neuen Vignetten, die diesmal jedoch nicht nur auf menschliche Interaktion ausgerichtet sind, sondern insbesondere ethische Dilemmata in Bezug zur MMI enthalten. Auch dies dient dem erwünschten Kompetenzerwerb, umfasst

²²⁴ Anmerkung: Der in Stundenbild 2 erläuterte Ablauf wurde im schulischen Kontext nicht durchgeführt. Die Annahme, dass eine Kombination problemorientierten Lernens und kompetenzorientierten Unterrichts zu einer vertieften Auseinandersetzung mit dem zu behandelnden Themenbereich führt, basiert allerdings auf dem Feedback von Schüler*innen und Mentor*innen, die an anderen Unterrichtseinheiten zu Roboterethik und Robotik teilgenommen haben, in denen genau diese Verknüpfung [mit unterschiedlichen Inhalten] erprobt wurde. Die Idee, fachdidaktische Ansätze zu vereinen, entstammt demnach den eigenen Erfahrungen im schulischen Kontext.

allerdings wie in Stundenbild 1 gleichzeitig einen problemorientierten Zugang, da die Schüler*innen komplexe Inhalte durch eigenständiges Denken und tatsächliches Philosophieren einordnen, klassifizieren und bewerten sollen.²²⁵

Die Vignetten werden hierfür im Plenum gemeinsam gelesen und besprochen. Es soll sowohl festgestellt werden, ob Schüler*innen Verantwortung in zwischenmenschlichen Situationen anders zuschreiben würden als in jenen, die zwischen einer artifiziellen und einer menschlichen Partei stattfinden und wenn ja, womit dies begründet wird. Des Weiteren sollen die Schüler*innen durch die Reflexion eigener Denkprozesse sowie der Übertragung theoretischer Inhalte auf praktische Fallbeispiele eine Wertung hervorbringen. Mittels des Versuchs, die erarbeiteten Definitionen auf Fallbeispiele anzuwenden und zu diesem Zweck auch jene Definition einzubeziehen, welche von der Lehrperson vorgegeben wurde, kann bei den Schüler*innen Unsicherheit darüber entstehen, ob die Verantwortungszuschreibung in der MMI überhaupt realistisch möglich ist. Genau an dieser Stelle besteht die Möglichkeit der Lehrperson, den Begriff der Utopie in Verbindung mit den dargelegten Konzepten von Verantwortung und Zurechenbarkeit zu bringen und gemeinsam mit den Schüler*innen ausführlich zu diskutieren, welche Aspekte der Fallbeispiele eine Verantwortungszuschreibung erschweren oder gar utopisch scheinen lassen.

Im Sinne der Kombination von POL und Kompetenzorientierung wird nun der gewünschte Kompetenzerwerb unterstützt, indem die Lehrperson im Sinne des „Cognitive Apprenticeship“²²⁶ die Rolle des „Diagnostiker[s] und Analytiker[s] von Lern- und Arbeitsprozessen“ übernimmt, „der hilft, kritische Momente des Lernverhaltens, begangene Fehler, den Nutzen von Lernstrategien zu artikulieren und zu reflektieren.“²²⁷ Die Rolle der Lehrperson kann an diesem Punkt der Problembearbeitung jedoch auch die eines „Krisenmanager[s]“ entsprechen, „der einem Lernenden, einer Gruppe aus einer Krise hilft.“²²⁸ Schüler*innen erleben sich hierdurch im Idealfall einerseits selbstwirksam, weil sie zu eigenständiger Reflexionsarbeit angeleitet werden, sind jedoch keineswegs auf sich alleine gestellt, sondern erhalten durch die zielgerichtete (kompetenzorientierte) Beteiligung der Lehrperson gleichzeitig genügend Hilfestellung, um die bestehende Problematik so bearbeiten zu können, dass sie Neugier auslöst anstatt Frustration hervorzurufen.

²²⁵ Vgl. Roeger (2016), S. 193.

²²⁶ Reusser (2005), S. 170.

²²⁷ Ebd. S. 170.

²²⁸ Ebd. S. 170.

Genau an diesem Punkt der Unterrichtseinheit zeigt sich, ob eine Verknüpfung von POL und kompetenzorientiertem Lernen erfolgreich umgesetzt werden konnte oder ob sich hierbei Probleme ergeben. Liessmann hatte vor einigen Jahren postuliert, Kompetenzorientierung sei unter anderem dafür verantwortlich, Unterrichtsgegenstände ihrer individuellen fachspezifischen Besonderheiten zu berauben, da die Orientierung an Kompetenzen gewissermaßen die Inhalte des Unterrichtsfaches selbst verblendet.²²⁹ Überdies schätzt Liessmann die Chance, dass Kompetenzerwerb mit Neugierde seitens der Schüler*innen verbunden ist, als gering ein.²³⁰

Diese Einschätzung teile ich in Bezug auf Unterrichtssequenzen, die lediglich und ausschließlich auf den theoriebasierten Erwerb bestimmter Kompetenzen ohne Anwendung abwechslungsreicher Unterrichtsführung oder Methodenvielfalt aufbauen. In Anbetracht der Idee, theoretische Inhalte jedoch mit problemorientierten Phasen der selbstständigen Denkarbeit zu verknüpfen, erscheint es wiederum gut möglich, in jungen Menschen Neugier für das Erlernen bestimmter Fähigkeiten hervorzurufen und Kompetenzorientierung als hilfreiches Werkzeug für das spätere Erwachsenenleben anzusehen. Nach eigener These dürfte die Neugier genau dann zum Tragen kommen, wenn Schüler*innen mittels einer Verknüpfung von POL und Kompetenzorientierung begreifen, wie vielfältig Denkprozesse sein können und wie groß das Potenzial ebendieser für das Hervorbringen neuer Ideen und Lösungsansätze ist. Dieses Ziel wird im vorgelegten Stundenverlauf durch eine Phase des Transfers fokussiert. Im Rahmen einer Partner*innenarbeit können Schüler*innen die im gemeinsamen Gespräch aufgekommene Aspekte schließlich filtern und strukturieren. Gemeinsam sollen die Jugendlichen im Anschluss daran Ansätze zur Lösung der bestehenden Problematik suchen und diese im Plenum vorstellen. Hierdurch erlernen die Jugendlichen gleichzeitig die Relevanz einer kritischen Betrachtung gesellschaftlicher Probleme und verstehen im Idealfall, dass gerade im Bereich der Roboter- und Maschinenethik in Hinblick auf die Verantwortungsdiffusion Lücken bestehen, die zukünftig dringend einer Lösung bedürfen.

Zu beachten ist abschließend, dass das Suchen und Finden von Lösungsansätzen und das gemeinsame Diskutieren ebendieser im Plenum kein vorgefertigtes Ziel ist oder ein solches verfolgt. Das ursprüngliche Ziel der zweiten Doppelstunde ist es, die Utopie der Verantwortungszuschreibung in der MMI für junge Menschen greifbar zu machen und bei

²²⁹ Vgl. Liessmann, Konrad Paul: Geisterstunde: Die Praxis der Unbildung. Eine Streitschrift. Wien: Zsolnay 2014, S. 75, zitiert nach Roeger (2016), S. 190.

²³⁰ Vgl. Liessmann (2014), S. 76, zitiert nach Roeger (2016), S. 185.

den Schüler*innen Denkprozesse anzuregen, die im Idealfall zu der Konzipierung von Ideen und Lösungen führen können. Selbst wenn Schüler*innen jedoch keine realistisch umsetzbaren Konzepte hervorbringen, ist die Unterrichtseinheit nicht weniger erfolgreich. Die Auseinandersetzung mit unterschiedlichen Kompetenzen und deren Anforderungen sowie das Durchlaufen eines philosophisch-problemorientierten Denkprozesses finden durch das Verbinden zweier fachdidaktischer Konzepte zeitgleich oder parallel zueinander statt, wodurch sich zwangsläufig auch ohne das strikte Definieren eines ultimativen Lehr- oder Lernziels ein ertragreiches Ergebnis abzeichnet, nämlich das Verständnis junger Menschen für die Komplexität vieler Inhalte und damit korrelierend das Wissen um die Notwendigkeit einer multidimensionalen Betrachtung ebendieser Sachverhalte.

9. Fazit

Im Rahmen der vorliegenden Masterarbeit wurden vier komplexe Fragestellungen behandelt und hinsichtlich ihrer Relevanz für ausgewählte Arbeitsfelder der Roboter- Medien- und Technikethik untersucht. Der Schwerpunkt dieser Arbeit lag hierbei auf der Analyse bestehender Literatur zu künstlicher Intelligenz, Konzepten moralischer Akteursschaft und spezifischen Problemfeldern innerhalb unterschiedlicher Bereichsethiken. Theoretische Fachliteratur wurde als Grundlage für die eingehende Problematisierung der Verantwortungszuschreibung in der MMI genutzt. Die These, die Zuschreibung von Verantwortung sei in der Mensch-Maschine-Interaktion grundlegend utopisch, erforderte zunächst eine Reihe an begrifflichen Abgrenzungen. Sowohl der Terminus „Künstliche Intelligenz“ als auch dessen Implikationen wurden strukturiert erläutert, um aufbauend in zugehörige Bereichsethiken einführen zu können. In diesem Kontext wurde auch eine zentrale Unterscheidung zwischen „Robotern“ und „Maschinen“ getätigt. Diese Differenzierung diene der Rechtfertigung des vorliegenden Arbeitstitels, in welchem als Teil der Bezeichnung „Mensch-Maschine-Interaktion“ das Wort „Maschinen“ sowohl für Roboter als auch für Maschinen stellvertretend ist. Möglich ist diese terminologische Verschmelzung von Robotern und Maschinen wie erläutert jedoch nur in eine Richtung. Während alle Roboter Maschinen sind, sind im Umkehrschluss nicht alle Maschinen Roboter.²³¹ Die Mensch-Maschine-Interaktion schließt somit sowohl Maschinen ein, die Roboter sind, als auch jene, die keine Roboter sind. Aus diesem Grund stellte sich die Frage

²³¹ Vgl. Loh (2019), S. 10.

nach der Verantwortungszuschreibung im Bereich artifizierter Systeme, die zweifelsfrei in den Bereich der MMI fallen, als umso interessanter heraus.

Für die vorliegende Masterarbeit wurden daher Forschungsfragen gewählt, die eine möglichst relationale Beleuchtung der Verantwortungsproblematik erlauben. Ausschlaggebend waren hierfür die folgenden zentralen Punkte: Wie kann Verantwortung definiert werden? Welche Kriterien für die Verantwortungszuschreibung gibt es? Inwiefern ist der Diskurs um die Verantwortungszuschreibung in der MMI sowie in Bezug auf autonome artifizelle Systeme utopisch? Wie können Schüler*innen der Sekundarstufe II philosophisch und problemorientiert an die vorliegende Thematik herangeführt werden?

Während Verantwortung umgangssprachlich ein Konstrukt zu sein scheint, welches keine oder vergleichsweise wenig Klärung verlangt, ergibt sich bei einer philosophisch-ethischen Betrachtung eine Bandbreite an Perspektiven, die nicht nur, jedoch insbesondere im Kontext der MMI für mehr Schwierigkeiten als Lösungen sorgen. Wie sich herausgestellt hat, liegt das jedoch nicht daran, dass in der Fachliteratur der Diskurs um die Verantwortungszuschreibung im zwischenmenschlichen Bereich lückenhaft ist, sondern dass schlichtweg zu viele verschiedene Konzepte existieren, die – wenn überhaupt – nur in Teilen die Verantwortlichkeit im Rahmen der MMI oder artifizelle Systeme als solche thematisieren.

Die Frage nach einer Definition von Verantwortung ergab folgende Ergebnisse: Verantwortung hat im Sprachgebrauch bereits eine lange Tradition, die in ihrem Ursprung auf das 15. Jahrhundert zurückgeht.²³² Der Terminus „Verantwortung“ hat hierbei unterschiedliche Konnotationen und Bedeutungen. Häufig scheint in gängiger Literatur die Fähigkeit, für etwas Rede und Antwort zu stehen²³³ als zentrales Element für das Tragen von Verantwortung auf. Differenziert wird jedoch in diesem Zuge zwischen unterschiedlichen Formen der Verantwortung, die ihrerseits wiederum mit unterschiedlichen Kriterien der Verantwortungszuschreibung einhergehen. Ergebnis der Frage nach einer Definition von Verantwortung ist demnach, dass der Terminus ohne eine Kontextualisierung irreführend ist und nicht zielbringend angewandt werden kann. Betrachtet man die zweite Forschungsfrage, welche die eben erwähnten Kriterien der Verantwortungszuschreibung in den Blick nimmt, so konnte Folgendes ausgearbeitet werden:

²³² Vgl. Ritter, Joachim und Karlfried Gründer u.a. (Hrsg.): Historisches Wörterbuch der Philosophie. Onlineversion Gesamtwerk. Basel: Schwabe AG Verlag.

²³³ Vgl. Höffe (2021), S. 3, siehe auch: Heidbrink (2017), S. 4.

Einer Reihe an Wissenschaftler*innen, hierunter Höffe, Bendel, Misselhorn, Blöser und anderen, ist bereits eine Verknüpfung des Verantwortungsbegriffs mit jenen Voraussetzungen gelungen, die Verantwortung zu einem anwendbaren Konstrukt machen. Interessanterweise finden sich hierbei immer wieder Überschneidungen. Höffe, welcher drei Formen der Verantwortung unterscheidet, setzt als Kriterium für die Zuschreibung ebendieser voraus, dass es sich bei den Adressaten um Personen, also Subjekte moralischen Handelns handelt. Sein Konzept der Primär- Sekundär- und Tertiärverantwortung richtet sich daher in erster Linie an jene Instanzen, die rechtlich als Personen gelten. Auch Bendel vertritt diesen Standpunkt, während Eckert die Problematik des Personenstatus um einen Aspekt erweitert, der im Rahmen der MMI relevant ist. So geht Eckert davon aus, dass ein Personenbegriff immer Handlungs- und Schuldfähigkeit impliziert. Diese Handlungs- und Schuldfähigkeit kommt jedoch – wie in späteren Teilen der Arbeit aufgezeigt wurde und wie in späteren Teilen dieses Fazits erneut erläutert wird – nicht nur realen beziehungsweise natürlichen Personen zu, sondern kann tatsächlich auch juristischen Personen, zu denen nachweislich auch komplexe Maschinen zählen können, zugeschrieben werden.

Eine andere Betrachtung des Verantwortungsbegriffs und deren Kriterien ergibt sich durch Blösers Unterscheidung von Verantwortung und Zurechnungsfähigkeit, die eng mit der These konnotiert ist, dass Kriterien für Verantwortungszuschreibung auch von der Art der Verantwortung selbst abhängen. Retrospektive Verantwortung, die sich auf vergangene Ereignisse bezieht, erfordert hiermit andere Kriterien als jene prospektive Verantwortung, die auf zukünftig geschehende Ereignisse ausgerichtet ist. Prospektive Verantwortung erinnert in ihren Ausführungen, die im Rahmen dieser Arbeit besprochen wurden, dem Konstrukt der Primärverantwortung nach Höffe. Beide setzen voraus, dass eine Person grundsätzlich ein Verantwortungsgefühl für zukünftige und geplante Handlungen hat und dieses pflichtmäßig auf sich nimmt. Die Fähigkeit, Verantwortung übernehmen zu können, also für diese zurechnungsfähig zu sein, scheint hierbei immer wieder als zentrales Kriterium für die Zuschreibung ebendieser auf.

Als ertragreich erwies sich in diesem Kontext auch die Betrachtung von Kriterien der Verantwortungszuschreibung in der Wissenschaft, die dem Themengebiet dieser Arbeit sehr nahe liegt. Hier wird, je nachdem, welche Kriterien erfüllt sind, von interner beziehungsweise von Rollenverantwortung oder von externer beziehungsweise Kausalhandlungsverantwortung gesprochen. Während interne Verantwortung voraussetzt, dass eine oder mehrere Personen identifiziert werden kann/können, der/denen bestimmte Rollen und Aufgaben zugeschrieben werden, für die sie in weiterer Folge die Verantwortung

trägt/tragen, erstreckt sich die externe Verantwortung über die gesamte Gesellschaft bis hin zu Natur und den Folgen wissenschaftlicher Forschung für diese.

Für die Beantwortung der dritten Forschungsfrage, die darauf abzielte, verständlich zu machen, inwiefern bei der Zuschreibung von wie auch immer gearteter Verantwortung in der MMI eine Utopie vorliegt, werden sämtliche bereits im Verlauf dieser Arbeit immer wieder aufgegriffenen Thesen und Erklärungen erneut zusammengefasst. Die erste Problematik besteht darin, dass Verantwortung entgegen der Erwartung kein konkretes eindimensionales Konstrukt ist, sondern je nach Kriterien und Auffassungen unterschiedlicher Wissenschaftler*innen sehr unterschiedlich interpretiert werden kann. Hierdurch ergibt sich die Schwierigkeit, nicht einfach sagen zu können, dass eine Instanz A einer anderen Instanz B „Verantwortung“ zuschreibt, wenn nicht geklärt ist, ob jene Form der Verantwortung, die mit der Hyperonymie „Verantwortung“ gemeint ist, der Instanz B überhaupt zugeschrieben werden kann oder ob Instanz A ein Recht hat, Instanz B Verantwortung zuzuschreiben. Gerade im Kontext der MMI ist dies von Relevanz, da schließlich, wie bereits erläutert, nach ursprünglicher Auffassung nur eine Person das Recht hat, Verantwortung zuzuschreiben. Wenn jedoch in Betracht gezogen wird, dass gerade durch den Diskurs um juristische Personen der Personenbegriff nicht mehr so eindeutig ist, wie dies für eine klare Verantwortungszuschreibung nötig wäre, ergibt sich daraus die Problematik, dass all jene Verantwortungskonzepte, die das Kriterium der Personenschaft voraussetzen, nicht mehr einwandfrei umzusetzen oder anzuwenden sind.

Eine zweite Problematik schließt hieran an. Selbst wenn nun eine Instanz A als Person klassifiziert wurde, die das Recht hat, einer Instanz B Verantwortung zuzuschreiben, so stellt sich dennoch die Frage, ob Instanz B jene Verantwortung, die ihr zugeschrieben wird, überhaupt realistisch tragen kann, oder ob es sich hierbei um einen Fehlschluss handelt. Insbesondere in Bezug auf die debattierte Problematik der Zuschreibung von *moral agency* oder *moral patiency* verschärft sich diese Frage. Wie aufgezeigt, gibt es in einigen Bereichen der Roboter- und Maschinenethik artifizielle Systeme, die keinem klassischen Personenbegriff entsprechen und dennoch aufgrund ihrer Fähigkeiten den Kriterien für moralische Akteursschaft gerecht würden. Dies stellt menschliche Instanzen, die ethisch über die Frage der Verantwortungszuschreibung im Bereich von KI-Systemen entscheiden sollen, vor neue Herausforderungen. Ein autonomes Waffensystem, welches eindeutig als moralisches Handlungssubjekt klassifiziert werden kann, müsste demnach in der Lage sein,

Verantwortung für sein Tun und Handeln zu übernehmen.²³⁴ Es bestehen hierfür schließlich genügend im Rahmen dieser Arbeit diskutierte Konzepte, die eine Verantwortungszuschreibung theoretisch erlauben würden. Dies bedeutet jedoch nicht, dass diese Verantwortungszuschreibung auch praktisch funktioniert. Sie ist, wie bereits mehrfach aufgezeigt, utopisch:

Ein autonomes Waffensystem oder ein autonomes artifizielles Pflegesystem verfügen über Fähigkeiten, die sie zu *moral agents* machen. Damit all diese Systeme jedoch überhaupt autonom fungieren können, wurde ein enormer Kostenaufwand betrieben. Doch nicht nur die Kosten, die in die Entwicklung und Produktion solcher Systeme fließen, sind beachtlich. Insbesondere die unüberschaubare Menge an Menschen aus unterschiedlichen Arbeitsfeldern, die in die Entwicklung und Programmierung dieser Maschinen involviert sind, kann ein enormes Problem darstellen. Insbesondere wenn ein autonomes System, ein vermeintliches Handlungssubjekt, einen Fehler begeht, der gravierende Folgen für die Menschen hat, die in eine MMI involviert sind, jedoch auch dann, wenn einem moralischen Handlungsobjekt solche Fehler unterlaufen. Selbstverständlich wäre es möglich, die Nutzer*innen oder Interaktionspartner*innen autonomer Systeme im Sinne der gebrauchtorientierten Verantwortung alleinig für die Folgen verantwortlich zu machen. Genauso wäre es denkbar, die Schuld im Sinne herstellungsorientierter Verantwortung auf die Konstrukteur*innen der jeweiligen Systeme abzuwälzen. Beides wäre denkbar, beides wäre möglich, aber egal, auf welche Weise das Blatt gedreht und gewendet wird: Es würde nichts etwas an der Tatsache ändern, dass hier eine Lücke besteht. Ich verweise hier erneut auf Josifovic, der dieses Dilemma als *Accountability Gap* bezeichnet. Verantwortung kann im zwischenmenschlichen Bereich problemlos zugeschrieben werden. Eine Ummünzung auf artifizielle Interaktionspartner wie Maschinen und Roboter lässt sich jedoch nicht ohne Weiteres vollbringen. Es entsteht hierdurch eine Reihe an Problematiken, die selbst durch die Betrachtung mehrerer ausgereifter Verantwortungsmodelle nicht gelöst werden kann. Hierfür bräuchte es deutlich einheitlichere Meinungen hinsichtlich der Kriterien für Instanzen, die in der Lage sind, Verantwortung zu übernehmen. Artifizielle Systeme können bislang niemals tatsächlich für ihr Handeln verantwortlich gemacht werden, solange immer noch menschliche Instanzen für deren Konstruktion, Entwicklung und Programmierung zuständig sind.

²³⁴ Vgl. Parthemore und Whitby (2013), S. 2.

Die vierte Forschungsfrage richtete sich auf eine fachdidaktische Umsetzung der geschilderten Problematik im Unterricht der Sekundarstufe II. Nachdem geklärt wurde, welche Arten der Unterrichtsführung hierfür in Frage kommen, wurden sowohl die Vor- als auch die Nachteile erläutert. Während kompetenzorientierter Unterricht von Schüler*innen häufig als langweilig empfunden wird, erlaubt er andererseits auch eine klare zielgerichtete Auseinandersetzung mit jedweder Thematik. Problemorientierter Unterricht hingegen ermöglicht es Schüler*innen, sehr frei und eigenständig an Unterrichtsthemen heranzutreten, ist jedoch für die Lehrkraft nicht immer einfach zu lenken und beinhaltet das Risiko, dass jene Kompetenzen, die nachweislich erreicht werden müssen, um den Anforderungen der Lehrpläne gerecht zu werden, nicht erfüllt werden. Aus diesem Grund entschied ich mich zu einer Verknüpfung von kompetenzorientiertem Unterricht und problemorientiertem Lernen.

Insbesondere in Hinblick des sehr komplexen Themenfeldes ist es von Relevanz, Schüler*innen einen groben Fahrplan vorzugeben und sich bereits im Vorhinein darüber im Klaren zu sein, welche Kompetenzen Schüler*innen am Ende der Unterrichtsstunde(n) entwickelt haben sollen. Trotzdem ist es nicht weniger bedeutsam, auch das echte und eigenständige Philosophieren der Jugendlichen zu unterstützen. Dies gelingt insbesondere dann, wenn Schüler*innen durch flexible Methoden und die mit Bedacht gewählte und adaptierte Rolle der Lehrperson die Freiheit besitzen, eigene Gedanken so zu strukturieren, dass das Themengebiet verstanden und gezielt bearbeitet werden kann.

Hierfür wurde bei der Auswahl der Methoden und dem Ablauf der Unterrichtseinheiten darauf geachtet, eine Ausgewogenheit zwischen jenen Phasen, in denen zielgerichteter Input von Seiten der Lehrperson gegeben wird und jenen, in denen eigenständig oder in Gruppen gearbeitet werden kann, herzustellen. Auf diese Weise wird nach eigener These am ehesten ein philosophischer Diskurs ermöglicht, der im Idealfall Schüler*innen dazu anleitet, bestehende Konstrukte gezielt zu hinterfragen, Querverbindungen zu schaffen, diese zu verstehen und selbst Ideen und Konzepte auszuarbeiten, die die Zukunft der Forschung positiv beeinflussen können.

10. Literaturverzeichnis

- Aethon. TUG. www.aethon.com (8.7.2023)
- American Psychological Association. (o. D.) Intelligence. In: APA Dictionary of Psychology. <https://dictionary.apa.org/intelligence> (24.2.2023)
- Bekey, George A: Current trends in robotics: Technology and ethics. In: Lin, Patrick und Keith Abney u.a. (Hrsg.): Robot ethics. The ethical and social implications of robotics. Cambridge: MIT Press 2012, S. 17–34.
- Bendel, Oliver: 300 Keywords Informationsethik. Grundwissen aus Computer-, Netz- und Neue Medien-Ethik sowie Maschinenethik. Wiesbaden: Springer 2016.
- Bendel, Oliver: Roboter im Gesundheitsbereich. In: Bendel, Oliver (Hrsg.): Pflegeroboter. Windisch: Springer Gabler 2018.
- Bien, Günther (Hrsg.): Aristoteles. Nikomachische Ethik. Hamburg: Felix Meiner⁴ 1985.
- Beranek, Angelika: Auch eine Frage der Menschenrechte: Was ist Künstliche Intelligenz? In: Politikum. Frankfurt am Main: Wochenschau Verlag 2021.
- Black, Jessica E.: An Introduction to the Moral Agency Scale. Individual Differences in Moral Agency and Their Relationship to Related Moral Constructs, Free Will, and Blame Attribution. In: Social psychology 47(6) 2016, S. 295–310.
- Blöser, Claudia: Zurechnung bei Kant. Zum Zusammenhang von Person und Handlung in Kants praktischer Philosophie. Boston: De Gruyter 2014.
- Bostrom, Nick: How long before superintelligence? In: International Journal of Futures Studies 2 (1998). <https://nickbostrom.com/superintelligence> (18.7.2023)
- Canguilhem, Georges: Maschine und Organismus. In: Die Erkenntnis des Lebens. Berlin 2012, S. 213–227.
- Collins, Allan und John Seely Brown u.a.: Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In: Resnick, Lauren B. (Hrsg.): Knowing, learning, and instruction: Essays in the honour of Robert Glaser. Hillsdale, New Jersey: Erlbaum 1989, S. 453–494.
- Christaller, Thomas und Michael Decker u.a.: Robotik: Perspektiven für menschliches Handeln in der zukünftigen Gesellschaft 14. Berlin/Heidelberg: Springer 2013.
- Danaher, John: The rise of the robots and the crisis of moral patiency. In: AI & society. London: Springer 2017, S. 33–55.
- Dobler, Ralph-Miklas: „Künstliche Intelligenz“ als Simulation und als Simulakrum. In: Brandstetter, Nicole, und Ralph-Miklas Dobler u.a. (Hrsg.): Mensch und Künstliche Intelligenz. Herausforderungen für Kultur, Wirtschaft und Gesellschaft. München: UVK Verlag 2021.
- Dobler, Ralph-Niklas: Mensch und Maschine. In: Brandstetter, Nicole und Ralph-Niklas Dobler u.a. (Hrsg.): Künstliche Intelligenz Interdisziplinär. München: UVK Verlag 2020.
- Dudenredaktion. (o. D.). Intelligenz. In: Duden online. <https://www.duden.de/rechtschreibung/Intelligenz> (24.2.2023)
- Eckert, Julia: Tugend, Recht und Moral: Tendenzen der Verantwortungszuschreibung. In: DZPhil 64(2) 2016.

- Europäische Kommission: High-Level Expert Group on Artificial Intelligence. A definition of AI: Main capabilities and scientific disciplines. Brüssel 2018. https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_deember_1.pdf (23.7.2023)
- Fenner, Dagmar: Einführung in die Angewandte Ethik. Tübingen: Narr Francke Attempto Verlag² 2022.
- Floridi, Luciano und J.W. Sanders: On the Morality of Artificial Agents. In: Wallach, Wendel und Peter Asaro (Hrsg.): Machine Ethics and Robot Ethics. London: Routledge 2017, S. 317–347.
- Fuchs, Michael: Gute wissenschaftliche Praxis. In: Heinemann, Thomas und Bert Heinrichs u. a. (Hrsg.): Forschungsethik. Eine Einführung. Heidelberg: J.B. Metzler 2010.
- Funk, Michael: Computer und Gesellschaft: Roboter und KI als soziale Herausforderung – Grundlagen der Technikethik, Band 3. Wiesbaden: Springer Fachmedien 2022.
- Fraunhofer-Institut für Produktionstechnik und Automatisierung. Care-O-bot 4. www.care-o-bot.de (8.7.2023)
- Gandhi, Sharad und Christian Ehl: AI&U – Translating Artificial Intelligence into Business. SV 2017.
- Gardner, Howard: Frames of Mind – The Theory of Multiple Intelligences. New York: Basic Books 1983 [Dt.: Abschied vom IQ – Die Rahmentheorie der vielfachen Intelligenzen. Stuttgart: Klett-Cotta 1991]
- Gausling, Tina: KI und DS-GVO im Spannungsverhältnis. In: Ballestrem, Johannes Graf und Ulrike Bär u.a. (Hrsg.): Künstliche Intelligenz. Wiesbaden: Springer Fachmedien GmbH, 2020.
- GoogleWatchBlog: Peinlicher Fehler: Google Photos erkennt/erkannte dunkelhäutige Personen als Gorillas. <https://www.googlewatchblog.de/2015/06/peinlicher-fehler-google-photos/> (23.7.2023)
- Grunwald, Armin: Technikethik. In: Düwell, Marcus und Christoph Hübenthal u.a. (Hrsg.): Handbuch Ethik. Stuttgart/Weimar 2011³.
- Grunwald, Armin: Überblick über die Technikethik. In: Grunwald, Armin und Rafaela Hillerbrand (Hrsg.): Handbuch Technikethik. Stuttgart/Weimar 2021².
- Heidbrink, Luger: Definitionen und Voraussetzungen der Verantwortung. In: Heidbrink, Ludger und Claus Langbehn u.a. (Hrsg.): Handbuch Verantwortung. Wiesbaden: Springer Fachmedien GmbH 2017.
- Himma, Kenneth Einar: 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 (2009).
- Hornscheidt, Antje Lann und Adibeli Nduka-Agwu: Der Zusammenhang zwischen Rassismus und Sprache. In: Nduka-Agwu, Adibeli u. Antje Lann Hornscheidt (Hrsg.): Rassismus auf gut Deutsch. Ein kritisches Nachschlagewerk zu rassistischen Sprachhandlungen. Frankfurt am Main: Brandes & Apsel 2013.
- Höffe, Otfried: Verantwortung: Begriff und Prinzip. In: Kahle, Reinhard und Niels Weidtmann: Verantwortung. Ein Begriff in seiner Aktualität. Boston: BRILL 2021.
- Jonas, Hans und Robert Habeck: Das Prinzip Verantwortung: Versuch einer Ethik für die technologische Zivilisation. Berlin: Suhrkamp 2020.

- Josifović, Saša: Denn wir wissen nicht, was sie tun. Das Problem der Verantwortungszuschreibung beim Einsatz selbstlernender Algorithmen. In: Philosophisches Jahrbuch (Freiburg) 127(1) 2020, S. 48–62.
- Julliard, Yannick: Ethische Technikgestaltung. Technikethik aus Sicht eines Ingenieurs. Frankfurt am Main: Peter Lang 2003.
- Karpov, Valery E.: Can a robot be a moral patient? In: Kuznetsov, Sergei O. und Aleksandr I. Panov u.a. (Hrsg.): Artificial Intelligence. Springer Nature Switzerland AG 2020.
- Klafki, Wolfgang: Allgemeinbildung in der Grundschule und der Bildungsauftrag des Sachunterrichts. In: Widerstreit Sachunterricht 4, 2005, S. 4.
- Klieme, Eckhard und Hermann Avenarius u.a.: Zur Entwicklung nationaler Bildungsstandards: eine Expertise. Bonn, Berlin: Bundesministerium für Bildung und Forschung 2009.
- Lehner, Michael und Inga Gryl u.a.: Problemorientierter Unterricht ≠ Problemorientierter Unterricht. In: GW-Unterricht 1 (2021).
- Liessmann, Konrad Paul: Geisterstunde: Die Praxis der Unbildung. Eine Streitschrift. Wien: Zsolnay 2014.
- Loh, Janina: Roboterethik. Eine Einführung. Stuttgart: Suhrkamp 2019.
- Malle, Bertram F. und Guglielmo, Steve u.a.: Moral, cognitive, and social: The nature of blame. In: Forgas, Joseph und Klaus Fiedler u.a. (Hrsg.): Social thinking and interpersonal behaviour. Philadelphia, PA: Psychology Press 2012, S. 311–329.
- Mayer-Steflic, Karin und Verena Toschner u.a.: Wortwechsel 4. Sprachbuch für die 8. Klasse AHS. Linz: Veritas 2019.
- Metzler Lexikon Philosophie: <https://www.spektrum.de/lexikon/philosophie/> (8.7. 2023)
- Misselhorn, Catrin: Robots as Moral Agents? In: Rövekamp, Frank und Friederike Bosse (Hrsg.): Ethics in Science and Society: German and Japanese Views. München: IUDICIUM Verlag GmbH 2013.
- Misselhorn, Catrin: Grundfragen der Maschinenethik. Stuttgart: Reclam⁴ 2019.
- Misselhorn, Catrin: Künstliche Intelligenz und Empathie. Vom Leben mit Emotionserkennung, Sexrobotern & Co. Ditzingen: Reclam 2021.
- Moor, James H.: The Nature, Importance, and Difficulty of Machine Ethics. In: IEEE Intelligent Systems, 21(4) 2006, S. 18–21.
- Neumann-Endres, Meike-Lesley: Die Wirkung der Emotionalen Intelligenz auf die interkulturelle Sensibilität – ein wissenschaftliches Essay. In: Valeva, Milena (Hrsg.): Emotionale Intelligenz. München: Rainer Hampp Verlag, 2019.
- Nevejans, Nathalie: European Civil Law Rules in Robotics. European Union: October 2016. [http://www.europarl.europa.eu/RegData/etudes/STUD/2016/571379/IPOL_STU\(2016\)571379_EN.pdf](http://www.europarl.europa.eu/RegData/etudes/STUD/2016/571379/IPOL_STU(2016)571379_EN.pdf) (8.7. 2023)
- Nida-Rümelin, Julian: Wissenschaft. In: Rippe, Klaus Peter (Hrsg.): Angewandte Ethik. Die Bereichsethiken und ihre theoretische Fundierung. Ein Handbuch. Stuttgart 2005².
- Objective or biased. On the questionable use of Artificial Intelligence for job applications. <https://web.br.de/interaktiv/ki-bewerbung/en/> (18.7.2023)
- Ott, Konrad: Verantwortung. In: Grupe, Ommo und Dietmar Mieth (Hrsg.): Lexikon der Ethik im Sport. Schorndorf: Karl Hofmann 1998, S. 578–587.

- Parthemore, Joel und Blay Whitby: What makes any agent a moral agent? Reflections on machine consciousness and moral agency. In: International Journal of Machine Consciousness 5(2) 2013.
- Pasupathi, Monisha und Cecilia Wainryb: Developing moral agency through narrative. In: Human Development 53, 2010, S. 55–80.
- Peter, Fabienne: Pure epistemic proceduralism. In: Episteme 5(1) 2008.
- Rechtsinformation des Bundes: Lehrplan der allgemeinbildenden höheren Schulen. <https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=10008568&FassungVom=2017-08-31> (13.3.2023)
- Reusser, Kurt: Problemorientiertes Lernen.- Tiefenstruktur, Gestaltungsformen, Wirkung. In: Beiträge zur Lehrerbildung 23(2) 2005, S. 159-182.
- Ritter, Joachim und Karlfried Gründer u.a. (Hrsg.): Historisches Wörterbuch der Philosophie. Onlineversion Gesamtwerk. Basel: Schwabe AG Verlag. <https://www.schwabeonline.ch> (11.4.2023)
- Roeger, Carsten: Philosophieunterricht zwischen Kompetenzorientierung und philosophischer Bildung 1(2). Opladen: Verlag Barbara Budrich, 2016.
- Ropohl, Günter: Neue Wege, die Technik zu verantworten. In: Lenk, Hans und Günter Ropohl (Hrsg.): Technik und Ethik. Stuttgart 1993².
- Rösch, Anita: Kompetenzorientierung im Philosophie- und Ethikunterricht: Entwicklung eines Kompetenzmodells für die Fächergruppe Philosophie, Praktische Philosophie, Ethik, Werte und Normen, LER. Berlin: LIT 2009.
- Schiedermeier, Gudrun: Diskriminierende Systeme – Rassismus und Frauenfeindlichkeit in KI-Systemen. In: Brandstetter, Nicole und Ralph-Miklas Dobler u.a. (Hrsg.): Mensch und Künstliche Intelligenz. Herausforderungen für Kultur, Wirtschaft und Gesellschaft. München: UVK Verlag 2021.
- Skorupinski, Barbara: Technikethik – Technikfolgenabschätzung. In: Maring, Matthias (Hrsg.): Ethisch-Philosophisches Grundlagenstudium 2. Münster: LIT 2005.
- Spiecker gen. Döhmman, Indra: Zur Zukunft systemischer Digitalisierung – Erste Gedanken zur Haftungs- und Verantwortungszuschreibung In: Computer und Recht 32(10) 2016, S. 698–704.
- Stahl, Bernd Carsten und Mark Coeckelbergh: Ethics of healthcare robotics: Towards responsible research and innovation. In: Robotics and Autonomous Systems 86, 2016, S. 152–161.
- Strandh, Sigvard: Die Maschine. Geschichte – Elemente – Funktion. Freiburg im Breisgau: Herder 1980.
- Sukopp, Thomas: Grenzen des Einsatzes von Künstlicher Intelligenz. In: Neue Technologien – neue Kindheiten? Stuttgart: J.B. Metzler, 2020. S. 197–225.
- Sullins, John P.: When Is a Robot a Moral Agent? In: International Review of Information Ethics 6(12) 2006.
- Schwartländer, Johannes: Verantwortung. In: Hermann Krings und Hans Michael Baumgarten u.a. (Hrsg.): Handbuch philosophischer Grundbegriffe Band 6. München: Kösel 1974, S. 1577–1588.

Wolff, Christian: Vernünftige Gedanken von Gott, der Welt und der Seele des Menschen, und auch allen Dingen überhaupt: den Liebhabern der Wahrheit mitgetheilet [1720]. Hildesheim 2009.

Wong, Henry: Siri, Alexa and unconscious bias: the case for designing fairer AI assistants. In: Design Week 2020.