



universität
wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Roboter als moral agents und das Kriterium der
moralischen Verantwortung“

verfasst von / submitted by

Mag. Arno Windisch

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Arts (MA)

Wien, 2019 / Vienna 2019

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

A 066 641

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

Interdisziplinäres Masterstudium Ethik für Schule
und Beruf

Betreut von / Supervisor:

Univ.-Prof. Mark Coeckelbergh, PhD

Mitbetreut von / Co-supervisor:

Dr. Janina Loh, MA

Die Wahrheiten der Menschen sind die unwiderlegbaren Irrtümer.

– Friedrich Nietzsche

Those who can imagine anything, can create the impossible.

– Alan Turing

Danksagung und persönliches Vorwort

Vor etwa einem Jahr besuchte ich das Seminar *Roboterethik* (bei Dr. Janina Loh, MA) und war immer wieder überrascht von Inhalten und Fragestellungen, die für mich zum Großteil neuartig waren. Besonders hat mich damals schon die Debatte um moral agency und Verantwortung interessiert, weshalb ich in diesem Gebiet auch meine Seminararbeit verfasste. Im Herbst 2018 war – nach Gesprächen mit Dr. Janina Loh, MA und Univ.-Prof. Mark Coeckelbergh, PhD – klar, dass ich auch meine MA-Arbeit in diesem Bereich schreiben möchte.

Auch wenn das Verfassen der Arbeit letztlich von viel Interesse getragen wurde, war der Weg von den ersten Lektüren bis zur Fertigstellung – das Recherchieren, Lesen, Vergleichen, Schreiben, Umschreiben etc. – ein mühevoller, aber lohnenswerter. Deshalb möchte ich diese Gelegenheit nun dazu nutzen, jenen Personen *DANKE* zu sagen, die mich hierbei tatkräftig unterstützt haben:

Besonderer Dank gilt meiner Mitbetreuerin Dr. Janina Loh, MA, die mich – immer wenn ich Rat und Hilfe brauchte – durch Gespräche und elektronische Nachrichten (hilfreich und immer sehr rasch) unterstützen konnte. Da Univ.-Prof. Mark Coeckelbergh, PhD – dem ich ebenfalls Dank für die Betreuung meiner Masterarbeit aussprechen möchte – zeitlich sehr begrenzt anzutreffen ist, entfällt die eigentliche Hauptlast der Betreuung/ Beurteilung der Arbeit auf Dr. Janina Loh, MA, wofür ich nochmals ganz herzlich *DANKE* sagen möchte.

Ein Dank gilt zudem meiner Familie und Freunden für jegliche Unterstützung sowie im Speziellen meinem Vater Gerhard Paul, der mir durch seine akribische Hilfe beim Korrekturlesen zur Seite stand.

Widmen möchte ich diese Arbeit meiner treuesten Begleiterin – Sally, denn ewig bleibt es unvergessen, was das Herz dem Herzen gab!

Inhaltsverzeichnis

1. EINLEITUNG	1
1.1. Ausgangspunkt, Relevanz und Erkenntnisinteresse der Arbeit.....	1
1.2. Eingrenzung, Zielsetzung und konkrete Fragestellung der Arbeit..	3
1.2.1. Eingrenzung der Arbeit.....	3
1.2.2. Zielsetzung der Arbeit.....	3
1.2.3. Festlegung der konkreten Fragestellung der Arbeit.....	4
1.3. Vorbemerkungen zur Methode und Materialauswahl.....	4
1.4. Darstellung der Gliederung der Arbeit.....	5
2. THEORETISCHE GRUNDLAGEN	7
2.1. Grundbegriffe im philosophischen Diskurs.....	7
2.1.1. Roboter.....	7
2.1.2. Moral, moralisch und Moralfähigkeit.....	9
2.1.3. Agency, moral agency bzw. moral agents.....	10
2.1.4. (Moralische) Verantwortung.....	11
2.1.4.1. Kurzer philosophiegeschichtlicher Abriss.....	12
2.1.4.2. Versuch einer Definition von Verantwortung.....	14
2.1.4.3. Verantwortung in Abgrenzung zu ähnlichen Begriffen/ Konzepten.	16
2.1.4.4. Verantwortung als Relationsbegriff.....	17
2.1.4.5. Zuschreibung von Verantwortung.....	20
2.2. Grundbegriffe im spezifisch roboterethischen Diskurs.....	23
2.2.1. Roboter.....	23
2.2.2. Moral, moralisch und Moralfähigkeit.....	24
2.2.3. Agency, moral agency bzw. moral agents.....	25
2.2.3.1. Moral agency, moral patiency und relationale Sichtweisen in der Roboterethik.....	26
2.2.3.2. (Moral) agency-Konzepte in der Roboterethik.....	27

2.2.3.3.	Übersicht über philosophische Positionen in der roboterethischen moral agency-Debatte.....	31
2.2.3.4.	Kriterien für moral agency.....	34
2.2.4.	(Moralische) Verantwortung.....	35
3.	METHODE UND MATERIALAUSWAHL	40
3.1.	Entwicklung und theoretische Hintergründe der Analysemethode...	40
3.1.1.	Philosophische Textanalyse.....	40
3.1.2.	Ansatz einer philosophisch-methodologischen Triangulation.....	41
3.1.3.	Philosophische Methoden.....	42
3.2.	Analyseschema: Die Anwendung der Methode im Kontext der Fragestellung.....	44
3.3.	Auswahl der zu analysierenden Texte.....	46
3.4.	Kritische Reflexion des methodischen Ansatzes.....	48
4.	DURCHFÜHRUNG DER TEXTANALYSE	51
4.1.	Analyse von: <i>Ethical Acts in Robotics</i> (Irrgang, 2006).....	51
4.1.1.	Kontextualisierung.....	51
4.1.2.	Argumentative Analyse.....	52
4.1.3.	Formale Analyse.....	62
4.2.	Analyse von: <i>When is a Robot a Moral Agent?</i> (Sullins, 2006)	71
4.2.1.	Kontextualisierung.....	71
4.2.2.	Argumentative Analyse.....	72
4.2.3.	Formale Analyse.....	84
5.	VERGLEICHENDE DISKUSSION	96
5.1.	Vergleichender Überblick wesentlicher Analyseaspekte der Texte....	97

5.2. Die Bedeutung des Kontexts/ des Diskurses für die Bestimmung/ Bedeutung von (moralischer) Verantwortung.....	99
5.2.1. Moral und (moral) agency als Prämissen für die Bestimmung/ Bedeutung von (moralischer) Verantwortung.....	99
5.2.2. Zusammenhang zwischen Roboter bzw. Mensch-Roboter-Verhältnis und (moralischer) Verantwortung.....	103
5.2.3. Kontextualisierender Blick auf die Autoren der Analysetexte als Prämisse für die Bedeutung/ Bestimmung von (moralischer) Verantwortung.....	104
5.3. Konkrete Bestimmung und Bedeutung von (moralischer) Verantwortung.....	105
5.3.1. Kriteriologisch-argumentative Bestimmung/ Bedeutung.....	106
5.3.2. Konzeptuelle Bestimmung/ Bedeutung.....	106
5.3.3. Bestimmung/ Bedeutung hinsichtlich formaler Aspekte.....	109
 6. SCHLUSSTEIL	 114
6.1. Schlussbemerkung zur Vorgehensweise.....	114
6.2. Wesentliche Erkenntnisse der Arbeit.....	114
6.3. Abschließende Reflexion, offene Fragen und Ausblick.....	116
 7. LITERATURVERZEICHNIS	 118
 Abstract (English).....	 123
Abstract (Deutsch).....	124

Tabellenverzeichnis

Tab.1: Grundlegende Bereiche der Roboterethik bzw. Sichtweisen auf Roboter in Bezug auf deren moralischen Status.....	27
Tab.2: Mögliche Modelle/ Formen von moral agency im Kontext der Roboterethik.....	30
Tab.3: Überblick über zentrale Positionen in der philosophischen Auseinandersetzung mit moral agency bei Robotern.....	33
Tab.4: Überblick über wesentliche Positionen zur Bedeutung des Kriteriums der (moralischen) Verantwortung für die roboterethische moral agency-Debatte.....	39
Tab.5: Schema zur (textspezifischen) Analyse der Bedeutung des Kriteriums der (moralischen) Verantwortung im moral agency-Diskurs der Roboterethik.....	45
Tab.6: (Vergleichende) Betrachtung bzw. Charakterisierung wesentlicher Aspekte der Textanalysen.....	97

Abkürzungsverzeichnis

Abb.	Abbildung
AI	Artificial Intelligence
AMA	Artificial Moral Agent
Anm.	Anmerkung
bzw.	beziehungsweise
engl.	englisch
etc.	et cetera
IT	Informationstechnologie
KI	Künstliche Intelligenz
lat.	lateinisch
m.a.	moral agency bzw. moral agent
mor.	moralisch
S.	Seite
Tab.	Tabelle
tw.	teilweise
u.a.	unter anderem
v.a.	vor allem
vgl.	vergleiche
z. B.	zum Beispiel
z. T.	zum Teil

Vorbemerkung zur Schreibweise und zum Gendern

Die in dieser Arbeit mehrfach erwähnten und aus dem Englischen stammenden Begriffe *agent*, *agency*, *moral agency*, *moral agent(s)*, *patient*, *patiency*, *moral patiency*, *moral patient(s)* wurden – auch wenn hauptwörtlich gebraucht – klein geschrieben, um deren Ursprungscharakter beizubehalten.

Diese Arbeit wurde nach bestem Wissen und Gewissen gegendert. Um keine Missverständnisse in diesem Punkt aufkommen zu lassen, soll hier eine ergänzende Vorbemerkung beigelegt werden: Bei der Verwendung der Begriffe *Akteur/ Akteurin* und *Adressat/ Adressatin* wurde – wenn allgemein gebraucht – jeweils die männliche und die weibliche Form angeführt. Wenn die Begriffe jedoch ausschließlich in Zusammenhang mit Robotern (bzw. artifiziellen Systemen) gebraucht wurden, wurde (da hier von keinen Geschlechtern auszugehen ist) die allgemeine Form, d.h. *Akteur* und *Adressat* verwendet.

1. EINLEITUNG

1.1. *Ausgangspunkt, Relevanz und Erkenntnisinteresse der Arbeit*

Im schnelllebigen Zeitalter der Informationstechnologie spielen technische Systeme (wie etwa Roboter) eine immer größer werdende Rolle (Kornwachs, 2013). Dies gilt nicht nur für den industriellen Sektor, in dem bereits eine Vielzahl von mechanischen Arbeiten artifiziellen Systemen mehr oder weniger überlassen werden, denn „*robotic systems are no longer confined to controlled industrial environments*“ (Allen, Wallach & Smit, 2011, S. 56), sondern zunehmend auch für den menschlichen Alltag, man denke an Bereiche wie Haushalt und Arbeit (z. B. Staubsaugerroboter), Sicherheit (z. B. Home-Security Roboter), das Gesundheitswesen (z. B. Pflegeroboter), Verkehr (z. B. selbstständig fahrende Autos) etc., um nur einige relevante Gebiete zu nennen (Lin, 2012). Es ist davon auszugehen, dass dieser Einfluss sowie die Komplexität und Möglichkeiten technischer Systeme weiter steigen werden (Floridi & Sanders, 2004; Lin, 2012).

Lin weist auf folgenden Umstand hin: „*The robotics industry is rapidly advancing, and robots in society today are already raising of these [Anm.: moral] questions.*“ (2012, S.12). Moralische Fragen bzw. Bedenken gehen also mit der Entwicklung von Robotern einher. In Bezug auf die eingangs erwähnten Beispiele könnten diese etwa lauten: Nach welchen (moralischen) Prinzipien sollen Pflegeroboter verfahren oder wer trägt die Schuld, wenn ein selbstständig fahrendes Auto in einen Unfall verwickelt ist? Obwohl artifizielle Systeme selbst keine natürlichen Entitäten darstellen, scheinen Beispiele wie diese klar zu machen bzw. gilt es als unumstritten, dass (zumindest einige Roboter) von moralischer Signifikanz sind (Asaro, 2006; Bendel, 2016; Johnson, 2011).

Entwicklungen bringen Wandel: Kornwachs (2013) stellt die zweideutige Frage: „*Was verändert Technik?*“ (S.25) und spielt damit einerseits darauf an, dass Gesellschaften das Verständnis und Wesen von Technik abwandeln, andererseits auf Veränderungen, die durch Technik in Gesellschaften ausgelöst werden. Lin fragt: „*How might society change with the Robotics Revolution?*“ (2012, S.10). Entwicklung und Wandel bzw. neuartige Situationen können zu (moralischen) Schwierigkeiten bzw. Orientierungsproblemen führen: Roboter ermöglichen neue Handlungen und damit können sich z. T. transformierende (moralische, gesellschaftliche etc.) Wertvorstellungen einhergehen: Auf die konkreten Beispiele (siehe: oben) bezogen, könnte etwa die Benutzung eines selbstständig fahrenden Autos (Anm.: neue Handlungsmöglichkeit) zur Vorstellung führen,

dass im Falle eines Unfalls dieser Roboter für schuldig erklärt werden sollte (Anm.: transformierende Wertvorstellung). Einen Roboter für etwas schuldig oder gar verantwortlich zu machen, mag im ersten Moment befremdlich wirken, da traditionellerweise Menschen (v.a. EntwicklerInnen einer Maschine) als alleinige moralische Handlungssubjekte gelten (Gunkel, 2012; siehe u.a.: 2.1.4.1.) – daher die Rede von transformierenden Wertvorstellungen. Ist die Vorstellung, einem Roboter (zumindest einen kleinen Teil an [moralischer]) Handlungs- oder Verantwortungsfähigkeit zuzuschreiben wirklich so abwegig? Und wovon – von welchen Fähigkeiten bzw. Kriterien – hängt dies ab? (siehe u.a.: 2.2.)

Stellt man Fragen, wie in dem eben angeschnittenen Beispiel, so befindet man sich mitten im Bereich der roboterethischen Debatte: In der Roboterethik, die als noch junge Bereichsethik gilt, geht es einerseits darum, ob/ inwiefern Roboter als moralische Handlungsobjekte – *moral patients* – gelten, d.h. passive Träger moralischer Rechte, andererseits beschäftigt sie sich damit, ob/ inwiefern Roboter selbst moralische Akteure/ Handlungssubjekte – *moral agents* – sein können, d.h. aktive Träger moralischer Pflichten (Loh 2017a). Im Laufe der näheren Beschäftigung mit dem moral agency-Diskurs trifft man auf verschiedene ethische Grundbegriffe bzw. Kriterien, die z. T. als notwendig erachtet werden, wenn es darum geht, auszumachen, wer/ was in einem moralischen Sinne als handlungsfähiges Subjekt gelten kann. Dazu zählen u. a. (moralische) Verantwortung, Autonomie, Rationalität, Personalität, Moralfähigkeit, Intentionalität etc. (siehe u.a.: 2.2.3.4.). Um dies auf das Beispiel von Vorhin zu beziehen, könnte etwa diskutiert werden, ob bzw. inwiefern das selbstfahrende Auto über Kompetenzen, wie (moralische) Verantwortung, Autonomie, Rationalität etc. verfügt und ob bzw. inwiefern es dadurch nicht bloß passiver Träger moralischer Rechte ist, sondern selbst den Status eines moralischen Akteurs einnehmen könnte.

Auf Grund der oben angesprochenen Entwicklungen und damit in Zusammenhang stehender moralischer Fragen erscheint es als erstrebenswert, den moral agency-Diskurs in der Roboterethik genauer zu betrachten. Im nächsten Abschnitt (1.2.) geht es darum, dieses Erkenntnisinteresse präziser einzugrenzen, um daraus letztlich eine sinnvoll zu bearbeitende, konkrete Fragestellung ableiten zu können.

1.2. Eingrenzung, Zielsetzung und konkrete Fragestellung der Arbeit

1.2.1. Eingrenzung der Arbeit

Im Bereich der roboterethischen moral agency-Debatte stößt man immer wieder auf die ethische (Schlüssel-)Kategorie bzw. das Kriterium der (moralischen) Verantwortung (siehe: 2.2.3., 2.2.4. bzw. z. B. Asaro, 2006; Gunkel, 2012; Loh, 2017a; Wallach & Allen, 2009).

Eine eingehendere Auseinandersetzung mit dem Begriff bzw. Kriterium der (*moralischen*) *Verantwortung* in der Roboterethik führt zu einer komplexen Debatte (siehe: 2.2.4.). Dem zu Grunde liegen unterschiedliche Konzepte/ Vorstellungen bzw. Sichtweisen von/ auf Roboter, Moral, (moral) agency und (moralische[r]) Verantwortung (siehe: 2.1.4., 2.2.4.). Welche Bedeutung kann (moralischer) Verantwortung als ein Kriterium für moral agency tatsächlich zukommen? Wie es für ethische Fragestellungen typisch ist, gibt es auch hier keine einfachen oder eindeutigen Lösungen bzw. Antworten, und so meint etwa Gunkel (2012): „[...] *that question* [Anm.: are machines moral agents?] *remains unanswered*.“ (2012, S.91). Umso bedeutungsvoller scheint daher eine reflektierte ethische Auseinandersetzung mit dem Themenkomplex *Roboter(ethik) – Moral/ (moral) agency – (moralische) Verantwortung* zu sein, was in der vorliegenden Arbeit geschehen soll. Die konkrete Zielsetzung sowie die genaue Festlegung der Fragestellung werden in den nächsten beiden Unterpunkten dieses Kapitels näher erläutert bzw. ausformuliert (siehe: 1.2.2., 1.2.3.).

1.2.2. Zielsetzung der Arbeit

Die vorliegende Arbeit macht sich – unter Berücksichtigung des aktuellen Forschungsstandes – zum Ziel, die Bedeutung des Kriteriums der (moralischen) Verantwortung im Kontext der roboterethischen moral agency-Debatte kritisch zu beleuchten (Anm.: Inwiefern moral agency-Konzepte und das Kriterium der [moralischen] Verantwortung zusammenhängen können, thematisiert z. B. 2.2.4.). Gunkel (2012) betont in seiner Auseinandersetzung mit moralischer Akteurschaft, dass dieses Konzept voraussetzungsgeladen, uneinheitlich und z. T. ambivalent ist und kommt daher zu folgendem Schluss: „*What has been discovered is that the concept of moral agency is already so thoroughly confused and messy that it is now unclear whether we [...] are in fact moral agents*.“ (S.91).

Vor diesem Hintergrund scheint die Annahme naheliegend, dass ebenso jene Kriterien, die häufig in (konstitutivem) Zusammenhang mit moral agency gebracht werden (siehe: 2.2.3.3., 2.2.4.), alles andere als einheitlich und voraussetzungsfrei behandelt bzw. beurteilt werden. Mit dem besonderen Interesse für das Kriterium der (moralischen) Verantwortung, dem häufig große Bedeutung in der Debatte zukommt (siehe z. B. Asaro, 2006; Gunkel, 2012; Misselhorn, 2013; Wallach & Allen, 2009), lässt sich folgende Vermutung aufstellen, die es im Laufe dieser Arbeit zu analysieren gilt:

Das Kriterium der (moralischen) Verantwortung spielt häufig eine (mit)entscheidende Rolle im roboterethischen moral agency-Diskurs (siehe v.a.: 2.2.3., 2.2.4.), ist zugleich jedoch voraussetzungsgeladen und wird uneinheitlich diskutiert. Die nähere Auseinandersetzung mit dieser Annahme stellt die Herausforderung der Arbeit dar. Die konkrete Fragestellung hierzu wird im nächsten Abschnitt beschrieben (siehe: 1.2.3.).

1.2.3. Festlegung der konkreten Fragestellung der Arbeit

Die bisher vorgenommene Erläuterung und Eingrenzung des Erkenntnisinteresses (siehe: 1.1., 1.2.1.) sowie die Konkretisierung der Zielsetzung (siehe: 1.2.2.) sollen nun in die folgende Fragestellung übergeführt werden, die es im Rahmen der Arbeit zu behandeln gilt:

Welche (einheitlichen oder unterschiedlichen) Bestimmungen (v.a. hinsichtlich: Prämissen, begrifflichen/ konzeptuellen Implikationen, [formaler] Charakteristika etc.) bzw. Bedeutungen (v.a. argumentativer, kriteriologischer, konzeptueller Natur) erfährt das Kriterium der (moralischen) Verantwortung im moral agency-Diskurs der Roboterethik?

1.3. Vorbemerkungen zur Methode und Materialauswahl

Das methodische Vorgehen der Arbeit verlangt vorab eine theoretische Auseinandersetzung mit Begriffen, die im Kontext der Fragestellung (siehe: 1.2.3.) als wichtig erachtet werden (siehe: 1.4.; 2.). Eine detaillierte Ausführung bzw. Begründung der Methode kann daher erst sinnvoll nach dem Theorieteil erfolgen (siehe: 3.). An dieser Stelle sollen lediglich Vorbemerkungen (inklusive relevanter Verweise) hierzu gegeben werden:

Im Rahmen einer Analyse von zwei Texten, die zur Beantwortung der Fragestellung als geeignet erachtet werden (Anm.: Kriterien für deren Auswahl, siehe: 3.3.), soll die Bedeutung des Kriteriums der (moralischen) Verantwortung im roboterethischen moral agency-Diskurs ermittelt werden. Da anzunehmen ist, dass es sich hierbei um ein voraussetzungsgeladenes und komplexes Thema handelt (siehe: 1.2.2.), wird das Verfahren einer philosophisch-methodologischen Triangulation herangezogen (Flick, 2004), das es erlaubt, mehrere Methoden zu kombinieren, um eine adäquate Herangehensweise an das Analysematerial zu entwickeln (siehe: 3.1.2.; Anm.: Welche philosophischen Methoden in diesem Kontext als sinnvoll bzw. zielführend erachtet werden, wird ausführlich in 3.1.3. beschrieben.).

Das praktisch ausgerichtete methodische Ziel besteht nun darin, ein speziell an die Fragestellung adaptiertes Schema zur Textanalyse zu entwickeln (siehe: 3.2.). Um dabei sicherzustellen, dass dies den philosophischen bzw. spezifischen Diskurs ausreichend mitberücksichtigt, scheint ein vorausgehender Theorieteil hierfür unabdingbar (siehe: 2.).

Dieser Ansatz, der in 3.4. ausführlich hinsichtlich seiner Vorzüge und möglicher Einschränkungen bzw. Nachteile reflektiert wird, soll es schließlich ermöglichen, in einem methodisch sowie theoretisch begründeten bzw. fundierten Vorgehen Fragen an Texte der Roboterethik zu richten, um ein besseres bzw. tiefergehendes Verständnis für die Bestimmung/ Bedeutung des Kriteriums der (moralischen) Verantwortung im moral agency-Diskurs zu bekommen.

1.4. Darstellung der Gliederung der Arbeit

In diesem Abschnitt geht es darum, einen groben Überblick über die Darstellung/ Gliederung der weiteren Arbeit zu vermitteln, die der Beantwortung der Fragestellung (siehe: 1.2.3.) dienen soll. Zu diesem Zweck soll folgendermaßen systematisch vorgegangen werden:

Da (moralische) Verantwortung im Sinne eines Kriteriums für moral agency bei Robotern im Zentrum der Arbeit steht, ist zu Beginn eine theoretische Auseinandersetzung mit (*moralischer*) Verantwortung bzw. damit in Zusammenhang stehenden Begriffen nötig (siehe: 2.). Um dabei eine ausreichende philosophische Grundlage bzw. Tiefe für das weitere Vorhaben zu erlangen, erscheint es als sinnvoll, dies zuerst in einem philosophischen (siehe: 2.1.) und erst anschließend im spezifisch roboterethischen Kontext (siehe: 2.2.) zu tun.

Der darauf folgende Abschnitt (siehe: 3.) setzt sich zum Ziel, die für die Beantwortung der Fragestellung adaptierte Textanalysemethode nachvollziehbar zu beschreiben bzw. zu erläutern (siehe: 3.1.). In Bezugnahme auf die bis hierhin geleistete methodische und theoretische Vorarbeit soll außerdem ein konkretes Analyseschema abgeleitet werden, das sich als praktikabel für eine spätere Textanalyse erweisen soll (siehe: 3.2.; Tab.5). Außerdem werden die für die Analyse ausgewählten Texte (sowie die Kriterien für deren Auswahl) bekanntgegeben (siehe: 3.3.) und das methodische Vorgehen hinsichtlich seiner Stärken und Schwächen kritisch beleuchtet (siehe: 3.4.).

Im nächsten Abschnitt (siehe: 4.) werden zwei Texte aus dem Bereich der Roboterethik nach einem speziell hierfür adaptierten sowie theoretisch und methodisch begründeten Verfahren analysiert (siehe: oben).

Um elaboriertere Einsichten in die Bedeutung/ Bestimmung des Kriteriums der (moralischen) Verantwortung im Diskurs zu bekommen, werden die aus der Analyse gewonnenen Erkenntnisse danach vergleichend diskutiert (siehe: 5.).

Der Schlussteil betrachtet die Arbeit kurz bzw. zusammenfassend und geht dabei v.a. darauf ein, inwiefern das Ziel der Arbeit erreicht werden konnte und welche Fragen (noch) offen bleiben (siehe: 6.).

2. THEORETISCHE GRUNDLAGEN

In diesem Kapitel geht es darum, die theoretischen Grundlagen zum Thema zu erarbeiten. Damit soll eine Basis für die Herausarbeitung einer Analysemethode zur adäquaten Auseinandersetzung mit der Fragestellung der Arbeit geschaffen werden (siehe: 3.), die letztlich eine Voraussetzung für die konkrete Textanalyse (siehe: 4.) darstellt.

Dieses Kapitel gliedert sich dabei in zwei inhaltlich/ logisch getrennte Abschnitte, die insgesamt einen möglichst ganzheitlichen, theoretischen Unterbau liefern sollen:

In einem ersten Teil (siehe: 2.1.) sollen – möglichst unabhängig vom spezifisch roboterethischen Diskurs – jene Begriffe näher beleuchtet werden, die für die Arbeit eine fundamentale Rolle spielen, und zwar vor allem jene, die sich direkt aus der Fragestellung (siehe: 1.2.3.) ablesen lassen, mit besonderem Schwerpunkt auf den Begriff/ das Konzept der (moralischen) Verantwortung.

Der zweite Teil dieses Kapitels (siehe: 2.2.) versteht sich hingegen als eine Auseinandersetzung mit zentralen Begriffen (die sich aus der Fragestellung ergeben [siehe: 1.2.3.]) im spezifisch roboterethischen Kontext.

Diese Herangehensweise – Betrachtung außerhalb bzw. innerhalb des speziellen Diskurses – soll dabei helfen, die umfassende Bedeutung/ Bestimmung wichtiger Begriffe herauszuarbeiten und Zusatzannahmen, Prämissen, Einschränkungen, besondere Charakteristika etc. der spezifischen Auseinandersetzung besser zu erkennen. In beiden Abschnitten dieses Kapitels gibt es zahlreiche Verweise darauf, inwiefern bzw. wo diese theoretischen Inhalte (v.a. das theoretisch herausgearbeitete [moralische] Verantwortungskonzept) in die Analysemethode miteinfließen (siehe: 3.2., Tab.5).

2.1. *Grundbegriffe im philosophischen Diskurs*

2.1.1. *Roboter*

Der Begriff *Roboter* entstammt dem Tschechischen, bedeutet ursprünglich *Fronarbeit* oder *Knechtschaft* und geht auf Josef und Karel Čapek (1920/1921) zurück, wobei letztgenannter damit

„humanoide Apparaturen, die Serviceleistungen und Arbeit an des Menschen Stelle übernehmen“ (Loh, 2017b, S.220) bezeichnet.

Auf Grund der Kürze der Begriffsgeschichte (siehe: oben) und wohl auch auf Grund der bei Weitem beschränkteren technischen Möglichkeiten der Vergangenheit, treten Roboter expressis verbis erst spät in philosophischen Betrachtungen auf. So gelten auch moralische Fragen, die damit in Zusammenhang stehen bzw. die Roboterethik selbst als eine junge Disziplin (Loh, 2017a). Auch wenn es vor Čapek praktisch keine direkte Philosophiegeschichte der Roboter gibt, wurde bzw. wird die heutige philosophische Roboterdebatte dennoch maßgeblich von philosophisch bedeutend älteren Ideen bzw. Denkansätzen geprägt, auf die hier (auf Grund des begrenzten Rahmens der Arbeit) nur kurz eingegangen werden kann (Anm.: Im Laufe der weiteren Beschäftigung werden diese immer wieder auftauchen bzw. thematisiert werden [siehe z. B. 2.1.3., 2.2.].):

Mechanistische Sichtweisen bzw. Argumente sind auch im heutigen Denken nach wie vor tief verwurzelt (Gunkel, 2012): So gelten etwa der Kartesischen Idee zufolge Maschinen (Anm.: Roboter als spezielle Form der Maschine, siehe: 2.2.1.) und Tiere als etwas durch natürliche Kräfte Determiniertes bzw. ihr Verhalten als das Resultat von Notwendigkeit (Gunkel, 2012). Dies ist eng mit einer dualistischen, anthropozentrischen Tradition (Coeckelbergh 2014; Gunkel, 2012) und instrumentalisierenden Denkansätzen (Feenberg, 1991; siehe: unten) verknüpft. Wenn es in philosophischen Auseinandersetzungen mit Robotern um Fragen rund um geistesphilosophische Aspekte bzw. *Philosophy of Mind* (Dennett, 1998) bzw. Fragen moralischer Akteurschaft geht (siehe: 2.1.3., 2.2.3.), spielen Roboter unter diesen Prämissen häufig eine nur untergeordnete Rolle, die v.a. in der Roboterethik zur Diskussion steht (siehe: 2.2.).

Die bekannteste Sichtweise auf Roboter (bzw. technische Systeme allgemein) kann nach Feenberg (1991) folgendermaßen beschrieben werden: „[...] *the instrumentalist theory offers the most widely accepted view of technology. It is based on the common sense idea that technologies are 'tools' standing ready to serve the purposes of users.*“ (S.5). In einem instrumentalisierenden Sinne gilt es auch heute noch vielfach, dass Roboter als vom Menschen hergestellte, künstliche Artefakte angesehen werden und unter einer solchen Perspektive höchstens als Werkzeuge ihrer ErzeugerInnen gelten: „*Computers and related systems are typically understood to be and conceptualized as a tool.*“ (Gunkel, 2012, S.24). Diese Sichtweise sowie die oben beschriebenen Einflüsse bzw. Denkansätze prägen auch rezent zahlreiche philosophische Auseinandersetzungen mit Robotern, stehen aber speziell auf Grund neuer technischer Entwicklungen und Möglichkeiten

vermehrt zur Diskussion (siehe: 1.1.). Damit in Zusammenhang stehende ethische Fragen behandelt die Roboterethik (siehe: 1.1., 2.2.). (Anm.: Die Frage, inwiefern Roboter im hier beschriebenen Sinne verstanden werden, wird im Rahmen der Textanalyse mitberücksichtigt [siehe: 3.2., Tab.5, B II, a].)

2.1.2. *Moral, moralisch und Moralfähigkeit*

Ethik und Moral haben mit Handlungen zu tun, sind jedoch voneinander abzugrenzen (Pieper, 2017). Während unter Ethik meist die Reflexion der Moral verstanden wird, beschäftigt sich Moral (aus dem lat. *mores*: Sitte, Brauch) mit der Sittlichkeit von Handelnden/ einer Handlung (Pieper, 2017).

Als moralisch gilt nach Pieper (2017) ein/e Handelnde/r bzw. eine Handlung, wenn diese/r einer Regel, einem ethischen System oder einer – in einer jeweiligen Gesellschaft – geltenden Moral(vorstellung) folgt bzw. entspricht, was wiederum eng mit Normen, Werturteilen und Institutionen zusammenhängt. Moral stellt nach Höffe und Forschner (2008) den normativen Grundrahmen für das Verhalten im gelebten, sozialen, politischen, kulturellen etc. Kontext dar. Üblicherweise geht es häufig um die Moral von Menschen, Gruppen, Individuen oder Organisationen (Pieper, 2017).

Während Menschen (in den meisten Fällen) als prinzipiell moralfähig gelten (häufig auf Grund unterschiedlicher Vermögen, wie etwa: Vernunftfähigkeit, Sprachlichkeit etc. [Gunkel, 2012]), ist die Frage, ob bzw. inwiefern andere Entitäten (wie etwa Tiere, Pflanzen, artifizielle Systeme) moralfähig sind bzw. (zukünftig) sein könnten, ein umstrittenes Thema, das eng in Zusammenhang mit einer Kritik an anthropozentristischen Exklusionsargumenten steht (z. B. thematisiert bei: Gunkel, 2012; Singer, 1976). (Anm.: Die Frage, ob/ inwiefern etwa Roboter selbst moralfähig sind/ sein könn(t)en, lädt in den spezifisch roboterethischen Diskurs ein und wird daher in 2.2.2.-2.2.4. theoretisch beleuchtet).

Konzeptuell gesehen kann Moral(fähigkeit) binär (im Sinne von: moralfähig versus nicht moralfähig, z. B. bei Kant [Gunkel, 2012]) oder graduell (im Sinne eines Kontinuums zwischen nicht moralfähig und voll moralfähig, z. B. bei Singer [1976]) aufgefasst werden. (Anm.: Die Frage, welches prinzipielle Moralverständnis einem Text zu Grunde liegt, wird – in Anlehnung an dieses Kapitel – im Rahmen der Textanalyse mitberücksichtigt [siehe: 3.2., Tab.5, B I, a; B II, b].)

2.1.3. Agency, moral agency bzw. moral agents

Agency hängt mit der „*idea of being capable of doing something that counts as an act or action.*“ (Himma, 2009, S.19) – also einer Form von Handlungsfähigkeit – zusammen. Jemand, der über agency verfügt, gilt als AkteurIn bzw. Handlungssubjekt. In der philosophischen Debatte geht es u.a. darum, damit in Verbindung stehende Kriterien bzw. Prämissen zu klären.

Moral agency bezeichnet hingegen eine spezielle Form der Handlungsfähigkeit: Diese bezieht eine moralische Perspektive mit ein, die einen individuellen Status relativ zu verschiedenen Vermögen (physischer, sensorischer, kognitiver etc. Natur) – die jeweils als entscheidender Ausdruck für moralische Handlungsfähigkeit gelten – beurteilt (Johnson, 2001). Im Vergleich zu agency ist moral agency somit exklusiver. Eine Entität, die über (moral) agency verfügt, wird demnach als (moral) agent bzw. (moralische[r]/ moralisches) AkteurIn/ Handlungssubjekt bezeichnet (Gunkel, 2012).

Während ein moral agent prinzipiell moralisch handeln kann, gibt es in diesem Zusammenhang zumeist auch ein Gegenüber (Entität, Individuum, Gruppe etc.), das diese Handlung erfährt/ erduldet bzw. auf das sie einwirkt, ohne dass dieses dabei selbst moralisch handeln kann. Letzteres wird als moralisches Handlungsobjekt bzw. moral patient bezeichnet (Loh, 2017a). Moral patiency stellt im Gegensatz zu moral agency ein inklusiveres Konzept dar (Gunkel, 2012).

Die Feststellung, ob bzw. inwieweit eine Entität als (moral) agent gilt oder nicht, dient nicht bloß einer Charakterisierung, sondern steht in wesentlichem Zusammenhang mit der Zuschreibung von (moralischen) Rechten und Pflichten: So gelten moralische AkteurInnen als aktive TrägerInnen moralischer Pflichten, während moral patients bloß passive TrägerInnen moralischer Rechte darstellen (Loh, 2017a).

Welche konkreten Entitäten gelten aus philosophischer Sicht als moral agents? Während Menschen (in den meisten Fällen) als moralisch handlungsfähig und somit als moral agents betrachtet werden, ist das für andere Wesen umstritten, denn die Frage, ob jemand/ etwas als moral agent gelten kann oder nicht, hängt letztlich vom vorherrschenden Moralverständnis (siehe: 2.1.2.) bzw. Handlungsverständnis ab und ist bisweilen häufig eng an individualistische Prinzipien bzw. ein Konzept von Persönlichkeit geknüpft (Loh, 2017b). Gunkel bezeichnet moral agency-Konzepte nicht zuletzt auf Grund dessen als voraussetzungsgeladen, uneinheitlich und z. T. ambivalent (2012; siehe: 1.2.2.). In diesem Zusammenhang sind v.a. tierethische Bemühungen der letzten

Jahrzehnte zu nennen, die häufig für eine Weitung eines zu stark anthropozentrisch geprägten – und daher zu exklusiven – Moralbegriffs bzw. moral agency-Konzepts eintreten (z. B. Singer, 1976).

Ähnlich, wie im letzten Abschnitt zu Moralfähigkeit beschrieben (siehe: 2.1.2.), kann auch moral agency als ein binäres (im Sinne von: eine Entität ist entweder moral agent oder kein moral agent) oder graduelles (im Sinne eines Kontinuums zwischen keiner und voller moral agency) Konzept aufgefasst werden (siehe z. B. Wallach & Allen, 2009).

(Anm.I: Die Frage, ob/ inwiefern Roboter moral agents sind/ sein könn[t]en, wird im spezifisch roboterethischen Theorieteil in 2.2.3. beleuchtet. Anm.II: Die Frage, welches agency-/ bzw. moral agency-Verständnis einem Text zu Grunde liegt, wird – in Anlehnung an dieses Kapitel – im Rahmen der Textanalyse mitberücksichtigt [siehe: 3.2., Tab.5, B I; B II, b].)

2.1.4. *(Moralische) Verantwortung*

Methodische Vorbemerkung: In der Analyse der Bedeutung des Kriteriums der (moralischen) Verantwortung in der roboterethischen moral agency-Debatte wird in vielen Punkten von einem etymologisch erarbeiteten, vorwiegend formalen Begriff/ Konzept von Verantwortung ausgegangen (siehe: dieses Kapitel bzw.: Sombetzki, 2014; siehe auch: 3.2., Tab.5, C I-IV). Ein solcher Zugang kann in diesem Kontext Aufschlüsse geben, z. B. indem betrachtet wird, über welche Bedingungen für die Zuschreibung von Verantwortung diskutiert, wie das Kriterium allgemein definiert und inwiefern es graduell verstanden wird etc. (siehe auch: 2.2.4.). Um diesem Anspruch gerecht zu werden, wird im Zuge dieses Kapitels versucht, vorwiegend fundamentale etymologische Aspekte bzw. formale Charakteristika eines Verantwortungsbegriffs herauszuarbeiten, die später an entsprechendes Analysematerial herangetragen werden können, um damit einen Beitrag zur Klärung der Bedeutung des zu untersuchenden Kriteriums zu leisten [siehe: 3., 4.].

Um ein Verständnis für (moralische) Verantwortung (als Begriff, Konzept, Modell) zu erlangen, gibt es mehrere Möglichkeiten. So kann z. B. versucht werden, dem Begriff in philosophiehistorischen Analysen nachzugehen. Da dieses Unterfangen im Rahmen dieser Arbeit jedoch zu umfangreich wäre (und nur ansatzweise berücksichtigt werden kann), besteht das primäre Ziel dieses Kapitels v.a. darin, formale und strukturelle Aspekte von Verantwortung genauer herauszuarbei-

ten. Warum dieser Zugang im Sinne der Beantwortung der Fragestellung dieser Arbeit als besonders hilfreich und praktikabel erachtet wird, ist in 2.2.4. bzw. im Rahmen der Erläuterung der Methode (siehe: 3.) genauer erklärt. Daher werden – nach einem kurzen philosophiegeschichtlichen Abriss (siehe: 2.1.4.1.) – die Schwerpunkte auf eine vorwiegend formale Betrachtung von Verantwortung gelegt, indem eine möglichst fundamentale, etymologisch herausgearbeitete Definition von Verantwortung (Sombetzki, 2014; siehe: 2.1.4.2.) herangezogen und auf die Begriffsstruktur, Begriffselemente (siehe: 2.1.4.4.) sowie Möglichkeiten der Zuschreibung von Verantwortung (2.1.4.5.) eingegangen wird. Außerdem wird der Begriff *Verantwortung* von ähnlichen, z. T. synonym verwendeten Begriffen – wie *Zurechnung*, *Pflicht*, *Schuld* – unterschieden (siehe: 2.1.4.3.). (Anm.I: Eine derartige Auseinandersetzung mit dem Begriff/ Konzept von Verantwortung soll eine theoretische Basis liefern, die zu späteren Analysezielen im Rahmen der Fragestellung dieser Arbeit dienen kann [siehe: 3.2.; vgl. oben]; Anm.II: Dieses Kapitel beschäftigt sich im Unterschied zu 2.2.4. jedoch weitgehend unabhängig vom roboterethischen Diskurs, versteht sich somit als eine Art neutrale Vorarbeit für die spätere Analyse der spezifischen Fragestellung. Anm.III: Verweise darauf, inwiefern die Erkenntnisse der einzelnen Abschnitte dieses Kapitels in der späteren Analyse roboterethisch relevanter Texte mitberücksichtigt werden, finden sich in jedem der Unterkapitel 2.1.4.1. – 2.1.4.5. und sind auch konkret dem Analyseschema [siehe: 3.2., Tab.5, C I-IV] zu entnehmen.)

2.1.4.1. Kurzer philosophiegeschichtlicher Abriss

Der Begriff *Verantwortung* hat eine vergleichsweise kurze Philosophiegeschichte (Precht & Burkard, 1999). Es finden sich zwar bereits in der Antike unterschiedliche Termini, die sich implizit mit (Aspekten) der menschlichen Verantwortung auseinandersetzten, häufig jedoch – etwa bei Aristoteles und Platon – rein kausaler Natur sind, z. B. Ursache, Erklärung, Begründung, Schuld zum Thema haben (Meyer & Hause, 2017; Anm.: zur Klärung begrifflicher Unterschiede, siehe: 2.1.4.3.). Viele DenkerInnen des Mittelalters, die Moral thematisieren, befassen sich ebenfalls mehrfach – wenn auch nicht *expressis verbis* – mit Fragen rund um Verantwortung, nicht selten in einem christlich-theologischen Kontext (Meyer & Hause, 2017).

Im Deutschen tritt der Begriff zwar schon im 15.Jahrhundert auf, wird dort – so wie in der Folge – jedoch überwiegend in einem juristischen Kontext verstanden (Fonnesu, 2017). Von der Neuzeit an rücken Fragen jenseits theologischer Prämissen vermehrt in den Vordergrund, die noch heute

zu zentralen Themen der (Zuschreibung von) Verantwortung(sfähigkeit) zählen und etwa die Problematik der (Willens-)Freiheit – die mehrheitlich als spezifisch menschliche Bedingung aller Moralität gilt – betreffen (Fonnesu, 2017). (Moralische) Verantwortung wird von DenkerInnen dieser Zeit eng in Verbindung mit personaler Identität, dem (Selbst-)Bewusstsein, Zurechnungsfähigkeit, Handlungs- und Willensfreiheit etc. gebracht und somit mit Kategorien, die auch heute noch eine wesentliche Rolle im Diskurs spielen (Fonnesu, 2017). Außerdem wird diese dabei – meist ganz stillschweigend – (fast) ausschließlich im Bezug auf menschliche Verantwortung(sfähigkeit), d.h. unter anthropozentrischem Blickwinkel verstanden (Fonnesu, 2017).

Zur philosophischen (Schlüssel-)Kategorie avancierte Verantwortung jedoch erst ab dem 18. Jahrhundert, also vergleichsweise spät im Gegensatz zu anderen, zentralen Kategorien der Philosophie (Bayertz & Beck, 2017). Im Laufe des 19. Jahrhunderts gewinnt diese an maßgeblicher philosophischer Bedeutung, und nicht früher als im 20. Jahrhundert spielt Verantwortung die Rolle einer ethischen Schlüsselkategorie, die (weiterhin und weitgehend) dem Menschen vorbehalten bleibt (Bayertz & Beck, 2017).

In jüngerer Zeit (v.a. in den letzten Jahrzehnten) spielen philosophische Diskurse, die – v.a. auf Grund des gesellschaftlichen/ technischen Wandels bzw. Fortschritts – eine Reflexion gängiger Konzepte und Vorstellungen von Verantwortung einfordern, zunehmend eine Rolle (siehe: 1.1., 2.2.4.): Gespeist werden solche Anliegen v.a. aus Bereichen der Tier- und Technikethik, die vielfach Kritik an einer zu anthropozentrischen Sichtweise moralisch relevanter Konzepte (u. a. Verantwortung) ausüben (z. B. Singer, 1976). Um bisher wenig bis überhaupt nicht berücksichtigte potenzielle Verantwortungssubjekte in den Diskurs mit aufzunehmen, wird hier u.a. für eine Weitung eines stark anthropozentrischen Moral- und Verantwortungsbegriffs argumentiert (Singer, 1976; siehe: 2.2.4.).

In der rezenten Verantwortungsforschung besteht in einigen Feldern weitgehender Konsens, beispielsweise darüber, dass Verantwortung von dialogisch-intersubjektiver Natur und als ein Relationsbegriff aufzufassen ist (Sombetzki, 2014; siehe: 2.1.4.2, 2.1.4.4.). Bayertz und Beck (2017) bezeichnen das relationale auch als das klassische Verantwortungsmodell und erwähnen, dass sich ein solches Verständnis seit der Neuzeit durchgesetzt hat und im Grunde auch heute von Gültigkeit ist. Weniger Einigkeit besteht in spezifischeren Angelegenheiten, wie etwa darin, wie viele oder welche Relationselemente der Begriff umfassen sollte (Loh, 2017b).

(Anm.: Die Frage, inwiefern [moralische] Verantwortung in roboterethischen Auseinandersetzungen hinsichtlich unterschiedlicher Aspekte der hier kurz angerissenen historischen Betrachtung diskutiert wird, wird im Rahmen der Textanalyse mitberücksichtigt [siehe: 3.2., Tab.5, B III].)

2.1.4.2. Versuch einer Definition von Verantwortung

Obwohl Verantwortung selbst im Alltag vielfach Verwendung findet, ist der Begriff alles andere als selbstverständlich oder einheitlich definiert (Heidbrink, 2017). Ganz im Gegenteil stößt man bei einer intensiveren Auseinandersetzung auf zahlreiche Definitionen und Voraussetzungen, die jeweils maßgeblich für ein Verständnis von Verantwortung sind, denn: *„Definitionen beeinflussen die Verwendung und den Gebrauch des Verantwortungsbegriffs und richten den Blick auf die erforderlichen Bedingungen verantwortlichen Handelns.“* (Heidbrink, 2017, S.4). Es sollte daher – bzw. in Rekurs auf die Zielsetzung der Arbeit (siehe: 1.2.2.) – große Aufmerksamkeit auf diesen Umstand gelegt werden.

Um keine wesentlichen Charakteristika des Verantwortungsbegriffs auszuklammern, könnte die Frage gestellt werden, ob es so etwas wie begriffliche Mindestanforderungen gibt, die in jedem Fall erfüllt sein müssen, wenn von Verantwortung die Rede ist. Sombetzki (2014) beschreibt in diesem Zusammenhang eine Minimaldefinition, die sich nur auf fundamentale etymologische Aspekte als Ausgangspunkt einer solchen Betrachtung bezieht. In diesem Sinne wird Verantwortung als ein dialogisches, intersubjektives Prinzip (im Unterschied zu z. B. Zurechnung, siehe: 2.1.4.3.) verstanden, das in minimalster Form auf nur drei Komponenten basiert (Sombetzki, 2014): (Anm.I: Diese Betrachtungsweise soll als Ausgangspunkt der weiteren Beschäftigung dienen, wobei anzumerken ist, dass auch andere Zugangsweisen denkbar wären. Auf die Frage, warum die gewählte Auffassung im Rahmen der Arbeit als sinnvoll erachtet wird, wird in 2.2.4. und 3. näher eingegangen. Anm.II: Die Frage, inwiefern [moralische] Verantwortung in roboterethischen Auseinandersetzungen im Sinne der folgenden Definition verstanden wird, wird im Rahmen der Textanalyse mitberücksichtigt [siehe: 3.2., Tab.5, C I].)

i) Verantwortung heißt, jemand steht Rede und Antwort (Sombetzki, 2014): Oben wurde bereits angemerkt, dass Verantwortung (etymologisch betrachtet) weitgehend als etwas Dialogisches bzw. Intersubjektives angesehen wird. Im Unterschied zum bloßen Antworten besteht in der Verantwortung somit ein gewisser Anspruch, ein gewisses Vermögen, der/ das als eine *„zweckgebundene Form des Antwortgebens“* (Sombetzki, 2014, S.37) bzw. als ein Rede-und-Antwort-Stehen

aufgefasst werden kann (Sombetzki, 2014). Immer dann, wenn es um Verantwortung bzw. Verantwortungszuschreibung geht, muss es auch notwendig ein Subjekt bzw. eine/n TrägerIn der Verantwortung geben (vgl. Verantwortung als Relationsbegriff, siehe: 2.1.4.4.). Nach Loh (2017b) ist es dieses Subjekt, das Rede-und-Antwort-stehen können muss. Im Verantwortungsdiskurs besteht keine Einigkeit darüber, wann eine Entität hierzu fähig ist, jedoch knüpfen klassische Verantwortungskonzepte in diesen Belangen eng an individualistische Prinzipien bzw. ein Konzept von Personalität (Loh, 2017b).

ii) Verantwortung hat mit einem normativen – nicht rein kausalen bzw. deskriptiven – Geschehen zu tun (Sombetzki, 2014): Der englische Begriff *responsibility* wird zwar bis heute häufig in einem sehr allgemein-kausalen Sinn verstanden (Fonnesu, 2017), dennoch kommt Verantwortung (v.a. in ethischen Debatten) meist ein speziellerer Gehalt zu, der ohne Normen und Regeln nicht auskommt: „Die normativen Kriterien definieren Verantwortungsbereiche, in denen jemand – begrenzt durch Normen – Rede und Antwort steht.“ (Loh, 2017b, S.224). (Anm.I: Im weiteren Verlauf dieser Arbeit wird [sofern nicht anders angegeben] Verantwortung in einem normativen [häufig auch kausalen, aber keinesfalls rein kausalen] Sinn behandelt und verstanden. Anm.II: Wichtig scheint speziell für die vorliegende Arbeit die Bemerkung zu sein, dass sich Verantwortung als ein normatives Geschehen aufgefasst, dadurch nicht automatisch im Bereich der Moral bewegen muss bzw. mit moralischer Verantwortung gleichzusetzen ist: Im ausschließlich moralischen Sinne versteht sich Verantwortung bzw. Verantwortlichkeit als etwas in der Sphäre des Moralischen [Wallace, 1994]. Anm.III: Auf [moralische] Verantwortung im roboterethischen Kontext wird im Kapitel 2.2.4 näher eingegangen).

iii) Verantwortung hat mit gewissen Kompetenzen der angesprochenen Entität zu tun (Loh, 2017b) bzw. setzt Verantwortung „eine spezifische psychomotivationale Verfasstheit des Verantwortlichen [voraus]“ (Sombetzki, 2014, S.39): In der Verantwortungsforschung besteht kein Konsens darüber, welche Fähigkeiten bzw. welche konkrete Verfasstheit einer verantwortlichen Entität zu Grunde liegen müssen, allerdings gibt es Tendenzen (Sombetzki, 2014). In einer etymologischen Interpretation des Verantwortungsbegriffs macht Sombetzki (2014) einige jener Kompetenzen aus (bzw. weist auf eine Verfasstheit hin), die auch im Diskurs weitgehend als wichtig erachtet werden (wird) und sich am einfachsten mit dem Wort *Erklärungsfähigkeit* translatieren lassen (lässt): In Bezug auf das Verb *stehen* beschreibt die Autorin den „Charakter der Bewusstheit“ (2014, S.40) und „ein allgemeineres Betroffen-Sein oder Sich-Angesprochen-Fühlen des Verantwortlichen“ (2014, S.41), in Hinblick auf das Rede-Stehen (im Sinne von Rechnung als

einem „*Ergebnis eines reflektierten Nachdenkens*“ [2014, S.40]) und das Antwort-Geben (auch in seiner prospektiven/ präventiven Bedeutung als Erklärungsfähigkeit) verweist sie auf kognitive Vermögen, wie Reflexionsfähigkeit und Urteilskraft (Fähigkeit zur Urteilsbildung) und letztlich stellt „*der eigentliche Akt des Rede-und-Antwort-Stehens eine bewusste und bedeutsame Handlung dar*“ (2014, S.41). (Anm.: Auf Aspekte der Zuschreibung von Verantwortung wird in 2.1.4.5. näher eingegangen.)

2.1.4.3. Verantwortung in Abgrenzung zu ähnlichen Begriffen/ Konzepten

Um (die Bedeutung von) Verantwortung klarer verstehen zu können, kann es nützlich sein, den Begriff von ähnlichen – z. T. sogar (funktionell) synonym verwendeten – abzugrenzen: Daher sollen in diesem Abschnitt einige dieser Verwechslungsmöglichkeiten bzw. verwandten Begriffe näher betrachtet bzw. von Verantwortung unterschieden werden: (Anm.: Diese begrifflichen Klärungen bzw. Abgrenzungen sollen außerdem darin hilfreich sein, ein möglicherweise abweichendes Verständnis von [moralischer] Verantwortung in Textanalysen aufzudecken und werden daher im Analyseschema mitberücksichtigt [siehe: 3.2., Tab.5., C II].)

Zurechnung: Fonnesu (2017) sieht die Wörter *Zurechnung* bzw. *imputatio* (lat. *imputare*: rechnen) als zentrale Vokabel einer philosophischen Auseinandersetzung mit Verantwortung bzw. verantwortlichem Handeln. Der Begriff findet auch funktionell häufig eine ähnliche Verwendung wie Verantwortung, weist jedoch einige Unterschiede (formaler Natur) auf: Zurechnung ist für gewöhnlich etwas Deskriptives/ Hierarchisches – in Form einer Subjekt-Objekt-Zuschreibung –, das (im Gegensatz zur weit verbreiteten Auffassung des Verantwortungsbegriffs) nicht unbedingt ein Bewusstsein für ein Gegenüber haben muss (Sombetzki, 2014). Als Akt der Zuschreibung ist Zurechnung konzeptuell außerdem weniger komplex als ein (meist) mehrstellig aufgefasstes, relationales Verantwortungskonzept (siehe: 2.1.4.4.). Außerdem bemerkt Günther (2006), dass sich Zurechnung bzw. Zurechenbarkeit stärker auf eine retrospektive Form von Verantwortlichkeit (vgl. engl. *accountability*) bezieht, wodurch der prospektive Charakter (siehe: 2.1.4.2., iii) von Verantwortung vernachlässigt wird.

Pflicht: Der Pflichtbegriff wurde (u.a. bei Crusius, Wolff) und wird vielfach nach wie vor eng in Zusammenhang mit Verbindlichkeit(en) oder Notwendigkeit aufgefasst (Höffe & Forschner, 2008). In diesem Sinne ist eine Pflicht nicht an langfristigen oder übergeordneten Zielen orientiert

(Höffe & Forschner, 2008) und untergräbt somit gewissermaßen die Notwendigkeit eines Folgenbewusstseins, das etwa bei Sombetzki (2014) als eine Bedingung für Verantwortungszuschreibung eine Rolle spielt (siehe: 2.1.4.5.). Auch Heidbrink (2017) bemerkt, dass Verantwortung nur unzureichend im Sinne eines klassischen Pflichtbegriffs aufgefasst werden kann. Letzterem zufolge spielt im Zusammenhang mit Verantwortung ebenso die Beachtung von Handlungsfolgen bei der Umsetzung von Regeln und Normen eine entscheidende Rolle. Pflicht hingegen – als eine bloße Einhaltung einer Regel bzw. Norm aufgefasst – wird dem Begriff der Verantwortung „*als das Entstehen eines Akteurs für die Folgen seiner Handlungen in Relation zu einer geltenden Norm.*“ (Heidbrink, 2017, S.5) nicht gerecht.

Schuld: Im alltäglichen Sprachgebrauch wird *schuldig sein für* häufig synonym zu *verantwortlich sein für* aufgefasst, und wenn es darum geht, jemandem Verantwortung für etwas (eine Handlung bzw. Handlungsfolgen) zuzuschreiben, dann geht es in der Tat auch häufig um Schuldzuschreibungen (engl. *[to] blame*) bzw. um die Frage – wie z. B. von Dennett (1997) formuliert – „*Who’s to blame (or praise)?*“ (S.351). Schuld haben bzw. schuldig zu sein ist eng an Konzepte von Verursachung bzw. Kausalität angelehnt. Während der englische Begriff *responsibility* bis heute z. T. sehr allgemein verstanden in einem schwachen Sinne als kausal aufgefasst werden kann (Fonnesu, 2017) – und somit beispielsweise auch etwa ein Stein als kausal verantwortlich für eine Verletzung gelten kann –, wird bei der Rede von Verantwortung meist stillschweigend von einem normativen Begriff (siehe: 2.1.4.2) bzw. einem klassischen Verantwortungsbegriff ausgegangen (Bayertz & Beck, 2017).

2.1.4.4. Verantwortung als Relationsbegriff

Das *Historische Wörterbuch der Philosophie* beschreibt Verantwortung u.a. als einen relationalen Begriff (Ritter, Gründer & Gabriel 2011). Auch in der Forschung besteht weitgehender Konsens darüber, dass Verantwortung ein relationales Konzept darstellt, während weniger Übereinstimmung darin besteht, welche die Relata (Relationselemente) sind bzw. über deren Stelligkeit (d.h. die Anzahl der Relata) sowie über den Status der Elemente dieses Begriffs (Sombetzki, 2014).

Für die weitere Arbeit soll das Konzept eines fünfstelligen Verantwortungsbegriffs in Anlehnung an Ritter, Gründer und Gabriel (2011) bzw. Sombetzki (2014) herangezogen werden. Demnach

kann Verantwortung in folgendem Sinne als ein formaler Relationsbegriff aufgefasst werden: Jemand (ein Verantwortungs-Subjekt als TrägerIn von Verantwortung; siehe unten: i) ist für etwas (Objekt/ Gegenstand der Verantwortung; Handlung, Handlungsfolgen; siehe unten: ii) vor einer Urteilsinstanz (vor der man sich verantwortlich zeigt; siehe unten: iii) gegenüber einem Adressaten bzw. einer Adressatin (als Gegenüber der Verantwortung; siehe unten: iv) in Bezug auf ein normatives Kriterium/ normative Kriterien (als Maßstab der Verantwortungszuschreibung; siehe unten: v) verantwortlich (Ritter, Gründer & Gabriel, 2011; Sombetzki, 2014). Verantwortung als relationaler Begriff aufgefasst bedeutet, dass diesen mehrere Relationselemente ausmachen, die in einem strukturierten (aber nicht hierarchischen) Sinne miteinander in (wechselseitiger) Beziehung stehen (Sombetzki, 2014). Auf die einzelnen Relata wird im Folgenden näher eingegangen: (Anm.: Die Frage, inwiefern [moralische] Verantwortung in roboterethischen Auseinandersetzungen als relationaler Begriff verstanden wird, wird im Rahmen der Textanalyse mitberücksichtigt [siehe: 3.2., Tab.5, C III].)

i) *Subjekt der Verantwortung*: Das Subjekt bzw. der/ die TrägerIn der Verantwortung gilt als gesichertes Element und wird beispielsweise schon bei Bayertz (1995) erwähnt. Weitere Bestimmungen in diesem Zusammenhang (wie z. B. die Frage, ob es sich in einem konkreten Fall beim Subjekt um ein Individuum oder ein Kollektiv handelt etc.) werden von Sombetzki (2014) als Subrelata dieses Elements bezeichnet. Komplex und uneinheitlich scheint jedoch die Frage, über welche Fähigkeiten bzw. Kompetenzen ein Subjekt verfügen muss, um tatsächlich als ein solches zu gelten und demnach Verantwortung übernehmen zu können (siehe: 2.1.4.5., 2.2.4.).

ii) *Objekt der Verantwortung*: Das Objekt bzw. der Gegenstand der Verantwortung beschreibt dasjenige, wofür Verantwortung übernommen bzw. zugeschrieben wird und lässt sich nach Loh (2017b) bzw. Sombetzki (2014) meist in (Form von) Handlungen und Handlungsfolgen auffassen. Über gewisse Rollen(vorstellungen) stehen Subjekt und Objekt der Verantwortung verschiedenartig (und nicht immer klar abgrenzbar) miteinander in Beziehung (Loh, 2017b). Der Gegenstand der Verantwortung steht in Zusammenhang mit vergangenen (retrospektiven) oder zukünftigen (prospektiven) Verantwortungsaspekten (Sombetzki, 2014).

iii) *Inстанz der Verantwortung*: Wie bereits erwähnt, kommt Verantwortung – betrachtet man die Begriffsgeschichte – schon relativ früh in einem juristischen Kontext vor (siehe: 2.1.4.1.), wobei das Gericht die Instanz darstellt, vor der sich ein Subjekt zu verantworten hat (Fonnesu, 2017). Man kann im Bereich der Verantwortungsinstanz unterscheiden zwischen: einer externen (z. B. der/ die RichterIn) oder internen (z. B. das gute Gewissen), einer absoluten/ letzten (Anm.: im

Sinne von unhinterfragbaren) oder vorletzten (Anm.: im Sinne von begrenzt in der Kompetenz) Instanz (Loh, 2017b). Nach Loh (2017b) handelt es sich meist um eine externe, vorletzte Instanz, die entweder öffentlich (d.h. anerkannt, akzeptiert, intersubjektiv geteilt, mit Sanktionskraft in Zusammenhang stehend) oder privat (d.h. intersubjektiv nicht geteilt, konventionell, von wenig/keiner Sanktionskraft) sein kann.

iv) *AdressatIn der Verantwortung*: Loh (2017b) bezeichnet den/ die AdressatIn als „*Warum der Verantwortung beziehungsweise als Gegenüber des Verantwortlichen*“ (S.224) und leitet dieses Relationselement aus dem Rede-und-Antwort-Stehen – das als ein normatives Konzept einer Begründung bedarf – ab (siehe: 2.1.4.2.). Während die Relata i – iii (siehe: oben) als (relativ) gesichert gelten, besteht im Verantwortungsdiskurs bezüglich des Adressaten bzw. der Adressatin größerer Dissens: So wird dieses Element z. T. synonym bzw. zu wenig differenziert zur Instanz der Verantwortung angesehen (Bayertz, 1995). Im Unterschied zur Verantwortungsinstanz muss der/ die AdressatIn „*nicht selbst potenzieller Verantwortungsträger sein können [...], nicht selbst die Kompetenzen für die Möglichkeit einer Verantwortungszuschreibung [mitbringen]*“ (Loh, 2017b, S.224).

v) *Normative Kriterien der Verantwortung*: In Bezug auf die herausgearbeitete Definition von Verantwortung (siehe: 2.1.4.2.) stellen normative Kriterien den Maßstab der Verantwortungszuschreibung dar (Ritter, Gründer & Gabriel, 2011). Als solche begründen Normen Verantwortlichkeit, „*definieren Verantwortungsbereiche, in denen jemand – begrenzt durch Normen – Rede und Antwort steht*“ (Loh, 2017, S.224) und das in unterschiedlichen Weisen (z. B. als Gesetze, Regeln, Gebote etc.), wobei sie häufig als Pflichten eine Rolle spielen (Loh, 2017b; Anm.: zur begrifflichen Abgrenzung: Pflicht – Verantwortung, siehe: 2.1.4.3.). Loh (2017b) bemerkt außerdem, dass der Kontext bzw. das Set der Kriterien festlegen, ob es sich letztlich um eine moralische Verantwortung (oder andere Formen, wie etwa wirtschaftliche/ rechtliche Verantwortung etc.) handelt. (Anm.: Es wurde bereits in 2.1.4.2. angemerkt und soll hier auf Grund der Wichtigkeit wiederholt werden, dass Verantwortung als ein normatives Geschehen aufgefasst, dadurch nicht automatisch mit moralischer Verantwortung gleichzusetzen ist.)

2.1.4.5. Zuschreibung von Verantwortung

Unter welchen Bedingungen sind Subjekte in der Lage, Verantwortung zu übernehmen, (ab) wann kann einem Subjekt Verantwortung zugeschrieben werden? Beschäftigt man sich mit Fragen dieser Art (d.h. im Kontext von Bedingungen für Verantwortungsfähigkeit), so gerät man häufig in eine schnell ausufernde Diskussion weiterer Begriffe und Konzepte, wie etwa Bewusstsein, Denkfähigkeit oder Autonomie (vgl. *Philosophy of Mind* [Dennett, 1998]), um nur einige wenige zu nennen. Laut Heidbrink (2007) habe man sich dazu passend mit einem ganzen „*Bündel an Vermögen und Fertigkeiten, die vom persönlichen Moralbewusstsein über praktisches Handlungswissen bis zum Umgang mit Unsicherheit reichen.*“ (S.10) auseinanderzusetzen. Die Eingangsfrage zu beantworten, wäre also eine komplexe, umfangreiche, gewissermaßen ungeordnete (im Sinne von nicht-hierarchische, vgl. *Bündel* [Zitat oben]) Angelegenheit, der im Rahmen dieser Arbeit nicht im Detail nachgegangen werden kann. Das gezwungenermaßen bescheidene Ziel dieses Abschnitts soll daher – v.a. in Anlehnung an Sombetzki (2014) – sein, jene Aspekte knapp herauszuarbeiten, die sich in einer „*Auseinandersetzung mit dem Wort selbst, seiner Genese und Etymologie*“ (Sombetzki, 2014, S.44) ergeben, wobei manche Vermögen des Fertigkeitenbündels nicht (näher) behandelt werden können. Sombetzki (2014) beschreibt drei Aspekte von Verantwortungsfähigkeit bzw. drei Verantwortungsbedingungen. Je nachdem, in welchem Ausmaß diese Bedingungen ausgeprägt bzw. verwirklicht sind, kann man von einem Mehr oder Weniger – und in diesem Sinne von einer Gradualität – an Verantwortung(sfähigkeit) sprechen (Loh, 2017b): (Anm.: Die Frage, inwiefern [moralische] Verantwortung in roboterethischen Auseinandersetzungen im Sinne dieser Aspekte verstanden und inwiefern in diesem Zusammenhang von einem graduellen Verantwortungsverständnis ausgegangen wird, wird im Rahmen der Textanalyse mitberücksichtigt [siehe: 3.2., Tab.5, C IV].)

i) *Kommunikationsfähigkeit*: Auf das Rede-und-Antwort-Stehen bzw. den sich etymologisch erschließen lassenden dialogischen Charakter von Verantwortung verweisend (siehe: 2.1.4.2.), macht Sombetzki (2014) die Fähigkeit der Kommunikation zu einer wesentlichen Verantwortungsbedingung. Derart aufgefasst, stellt jede Verantwortlichkeit auch einen Kommunikationsakt dar (Sombetzki, 2014). (Anm.: Inwiefern diese Fähigkeit mit Sprachlichkeit zusammenhängt bzw. sich davon unterscheidet, wird hier nicht näher behandelt und kann z. B. bei Sombetzki [2014] nachgelesen werden.)

ii) *Handlungsfähigkeit*: Ohne die Fähigkeit, handeln zu können, kann ein Rede-und-Antwort-Stehen gar nicht zum Ausdruck gebracht werden, denn erst „*Mit der Beschreibung eines Verhaltens*

als Handlung beginnt die Möglichkeit der Verantwortungszuschreibung“ (Nida-Rümelin, 2007, S.60). Was ist jedoch eine Handlung und was grenzt eine solche von (z. B. instinktivem) Verhalten ab? Handlungen sind von intentionaler (d.h. zielgerichteter) Natur (Nida-Rümelin, 2007). Was bedeutet es wiederum, dass jemand aus einer Intention heraus handelt? Antworten auf diese Frage sind bisweilen uneinheitlich und rekurren – mehr oder weniger – auf geistesphilosophische Begriffe/ Konzepte die alles andere als klar bestimmt, voraussetzungsfrei oder unmetaphysisch sind. Jedoch werden sie häufig in Zusammenhang mit *Philosophy of Mind* (Dennett, 1998) gebracht und ganz unabhängig hiervon wäre Verantwortung schwerlich verstehbar. Im Folgenden sollen die in diesem Kontext bedeutsamsten Kompetenzen (siehe unten: a-e) (v.a. in Anlehnung an: Loh, 2017b; Sombetzki, 2014) – die hier wiederum nur knapp und pragmatistisch, vorwiegend deskriptiv behandelt werden können – beschrieben bzw. thematisiert werden: (Anm.: Da im Rahmen der Arbeit keine ausführliche Beschäftigung mit nachstehenden Begriffen/ Konzepten möglich ist, wird mehrfach versucht, Verweise auf damit in Zusammenhang stehende Themen/ Problematiken sowie sich kritisch dazu äussernde AutorInnen zu geben.)

a) *Folgenbewusstsein*: Dieses kann als ein vorausblickendes Handeln, das Wissen bedingt, aufgefasst werden (Sombetzki, 2014). Wie dieses Wissen bestimmt ist, ist nicht einfach zu klären und wird beispielsweise von Birnbacher (2003) diskutiert.

b) *Kontextwahrnehmung (als Geschichtlichkeit)*: Die möglichen Folgen einer Handlung können nur dann ermittelt, eingeordnet bzw. eingeschätzt werden, wenn die jeweilige Situation (der Handlungskontext) vom Handlungssubjekt wahrgenommen werden kann (Sombetzki, 2014).

c) *Personalität*: Verantwortung stellt (häufig) ein individualistisches Prinzip dar (Loh, 2017b) und auch wenn das Konzept von Personalität (philosophiegeschichtlich) schwierig und komplex bzw. metaphysisch geladen ist, scheint es dennoch untrennbar an Verantwortungsfähigkeit geknüpft zu sein (Sombetzki, 2014). Das Problem der anthropozentrischen bzw. spezieisistischen Genese bzw. Gleichsetzung des ursprünglich neutralen Begriffs *Person* mit *Mensch* kritisieren bzw. behandeln beispielsweise Singer (1999) oder Gunkel (2012).

d) *Einflussmöglichkeit*: Das Unterlassen einer Handlung – d.h. nicht zu handeln, wo man hätte handeln können – wird weitläufig als Handlung angesehen (Pieper, 2017). Anders ist es jedoch in Situationen, in denen für ein Subjekt überhaupt keine Möglichkeit zur Einflussnahme gegeben ist. Ein/e VerantwortungsträgerIn muss also in der Lage sein, auf eine Situation/ einen Sachverhalt

mittels einer Handlung einwirken zu können. Für Handlungsfähigkeit – verstanden als ein intentionales Verhalten (siehe: oben) – scheint die Möglichkeit des absichtsvollen Einflussnehmens Voraussetzung zu sein, was aber nicht unbedingt bedeutet, dass Kontrolle in einem absoluten Sinne vorliegen muss (Sombetzki, 2014).

e) *Freiheit*: Ähnlich wie Personalität ist auch Freiheit ein uneinheitlich definierter, komplexer und metaphysisch aufgeladener Begriff und dennoch – da so fundamental im Denken bzw. sozialen Leben verankert (Kutz, 2002) – kaum als Bedingung für Handlungsfähigkeit auszuklammern, womit auch der Verantwortungsdiskurs weitgehend konform geht (Sombetzki, 2014). Der Begriff ist eng mit Debatten um Willensfreiheit, Handlungsfreiheit, Determinismus etc. verbunden, worauf hier nicht näher eingegangen werden kann (beispielsweise nachzulesen bei: Frankfurt, 1969; Pauen, 2001; Sombetzki, 2014). Im Kontext der Arbeit, im Sinne der Praktikabilität und in Rekurs auf Sombetzki (2014) wird Freiheit im Folgenden vorwiegend im Sinne von Autonomie und kognitiven Kompetenzen (Rationalität und Urteilkraft; siehe: 2.1.4.2. iii) verstanden, wobei letztgenannte Autorin Autonomie als *„Selbstbestimmung, in der die drei Verantwortungsbedingungen, Kommunikations- und Handlungsfähigkeit sowie Urteilkraft zusammenfallen und Personalität konstituieren.“* (2014, S.57) beschreibt. (Anm.: Auch im Analyseschema [siehe: 3.2., Tab.5] bzw. in der konkreten Textanalyse [siehe: 4.] wird vorwiegend mit einem Freiheitsverständnis in diesem Sinne gearbeitet.)

iii) *Urteilkraft*: Verantwortung in einem nicht rein kausalen Sinne bezieht sich auf ein System bestimmter Werte oder Normen (siehe: 2.1.4.2.) und hat mit Werturteilen zu tun (Loh, 2017b). Damit so etwas wie ein Normensystem entstehen bzw. sich jemand auf ein solches festlegen kann, bedarf es einer Urteilkraft, *„die dem Verantwortlichen die Gründe und Kriterien an die Hand reicht, mit deren Hilfe er sein Handeln als Rede-und-Antwort-Stehen beurteilen kann.“* (Sombetzki, 2014, S.58). Diese hängt mit weiteren kognitiven Kompetenzen (wie Reflexionsfähigkeit und Rationalität) zusammen. Sombetzki erwähnt zudem, dass Urteilkraft als eine Bedingung für Verantwortung außerdem *„auch mit zwischenmenschlichen Institutionen wie Versprechen, Verlässlichkeit und Vertrauen als Grundlage zwischenmenschlicher Beziehungsverhältnisse zusammen[hängt]“* (2014, S.60) und zwar insofern, als erst durch die Urteilkraft (bzw. sich daraus ergebende, handlungsleitende Gründe) ein Ausgangspunkt entsteht, von dem ausgehend die vorhin genannten Institutionen in der sozialen Praxis folgen können (Sombetzki, 2014).

2.2. *Grundbegriffe im spezifisch roboterethischen Diskurs*

2.2.1. *Roboter*

In Texten mit direkter roboterethischer Relevanz (Anm.: in denen Roboter ausdrücklich thematisiert werden) trifft man häufig auf ähnliche, z. T. synonym zu *Roboter* verwendete Begriffe, wie *Maschine* oder *Computer* (z. B. bei: Gunkel, 2012; Wallach & Allen, 2009), was zu Verwirrungen führen kann. Dies mag daran liegen, dass sich die Roboterethik aus der Computerethik entwickelt hat bzw. ein Spezialgebiet der Maschinenethik darstellt (Bendel, 2016). Es ist schwierig, Roboter im Sinne von notwendigen oder hinreichenden Bedingungen zu definieren, wenn diese jedoch in moralischen Belangen zum Thema werden, werden sie meist als Maschinen bezeichnet, die in irgendeiner Form auf die Umgebung einwirken können (Ichbiah, 2005). Dadurch heben sie sich häufig von einfacheren Computern ab, wobei anzumerken ist, dass es praktisch nicht immer eindeutig oder einfach festzustellen ist, ab wann etwas ein bzw. ab wann etwas kein Roboter (sondern z. B. bloß ein Computer) ist. (Anm.: Inwiefern Einflussmöglichkeit unabhängig von der Roboterethik bzw. im Sinne eines etymologisch hergeleiteten Verantwortungskonzepts als Bedingung für Verantwortung eine Rolle spielt, wurde bereits behandelt [siehe: 2.1.4.5. ii, d]. Die Frage, inwieweit rezente Roboter – auf Grund ihrer Möglichkeiten des Einwirkens bzw. der Einflussnahme auf die Umwelt – zumindest in einem rudimentären Sinn als autonom bezeichnet werden können, ist jedoch umstritten [Gunkel, 2012].)

In der Roboterethik werden außerdem teilweise die Bezeichnungen *artifizielles System* (z. B. bei: Loh, 2017b) oder *AMA* (engl. *artificial moral agent*, z. B. bei: Floridi & Sanders, 2004) verwendet, womit häufig komplexe Maschinen, Computer(systeme), aber auch Roboter angesprochen werden. Dennoch steht die Bezeichnung *artifizient* nicht der Tatsache (bzw. dem Trend) entgegen, dass – im Zuge der zunehmenden technischen Entwicklung und Komplexität – einige hochkomplexe Roboter zunehmend menschenähnlicher bzw. natürlicher (in Aussehen, bestimmten Kompetenzen, Verhalten etc.) wirken bzw. zumindest so erscheinen (Sparrow, 2012; siehe: 2.2.3.), weshalb in diesem Zusammenhang häufig auch die Rede von *humanoiden Robotern* ist (Bekey, 2012).

Auf die traditionelle bzw. weit verbreitete instrumentalisierende Auffassung von Robotern als Werkzeuge wurde bereits in 2.1.1. eingegangen. In der Roboterethik besteht hingegen – und nicht zuletzt auf Grund zunehmender Komplexität artifizienter Systeme (siehe: 1.1.) – vielfache Kritik

an derartigen Sichtweisen, was häufig in Argumentationen rund um graduelle, alternative, inklusive und relationale Ansätze/ Konzepte deutlich wird, worauf in den nächsten Abschnitten ausführlicher eingegangen wird (siehe: 2.2.2.-2.2.4.).

(Anm.: Die Frage, inwiefern Roboter in roboterethischen Auseinandersetzungen im Sinne einer hier thematisierten Charakterisierung aufgefasst werden, wird im Rahmen der Textanalyse mitberücksichtigt [siehe: 3.2., Tab.5, B II].)

2.2.2. Moral, moralisch und Moralfähigkeit

Die Tatsache, dass Roboter selbst unbelebt und möglicherweise rezent nicht (oder nur z. T.) moralfähig sind, schließt diese jedoch nicht kategorisch aus dem moralischen Universum aus (Loh, 2017b). Dass Roboter (zunehmend) – unabhängig von der Frage, ob/ inwieweit sie selbst moralisch handeln können – Entitäten von moralischer Signifikanz sind, scheint innerhalb und außerhalb des roboterethischen Diskurses unumstritten zu sein (Asaro, 2006; Bendel, 2016; Floridi & Sanders, 2004; Johnson, 2011; siehe: 1.1.) und wird an Beispielen des Alltags (z. B. autonome Fahrassistenzsysteme, Pflegeroboter etc.) besonders deutlich. Roboter haben also mit Moral bzw. moralischen Problematiken zu tun, doch Johnson (2011) betont, dass eine moralische Entität nicht automatisch moral agent ist. Moral selbst wird von unterschiedlichen DenkerInnen verschieden betrachtet: So wird diese traditionell als etwas spezifisch Menschliches (z. B. bei: Bringsjord, 2007), aber beispielsweise auch im kompletten Kontrast dazu von Nadeau (2006) als etwas über den Menschen Hinausgehendes bezeichnet (siehe: 2.2.3.3.).

Vor diesem Hintergrund stellt sich in roboterethischer Hinsicht häufig die Frage, ob Roboter selbst moralische/ moralfähige Wesen oder sogar moralische Akteure sind (bzw. zukünftig sein könnten). Auf letzteren Punkt wird im nächsten Abschnitt (siehe: 2.2.3.) näher eingegangen. (Anm.: Auf damit in Zusammenhang stehende, weiterführende Debatten [wie etwa die AI-Diskussion, Top-Down- sowie Bottom-Up-Ansätze zur Implizierung von Moral bei/ in Robotern etc.] kann hier aus Platzgründen nicht ausführlich eingegangen werden. Diese Themen werden beispielsweise von Gunkel [2012], Lin, Abney und Bekey [2012] oder Wallach und Allen [2009] diskutiert.)

Johnson (2011) betont, dass Begriffe bzw. Konzepte wie Moral (aber auch agency- oder moral agency-Konzepte) tief in der westlichen Tradition wurzeln und Menschen somit historisch bedingt als freier, selbstbestimmter etc. angesehen werden (siehe: 2.1.2.). Ausgehend von einer daran anschließenden Kritik an stark anthropozentrischen und daraus resultierenden, exklusiven Moralvorstellungen wird in der Roboterethik vielfach für eine Weitung/ Öffnung/ Änderung bzw. für alternative/ relationale moralische Konzepte/ Ansätze argumentiert, die es ermöglichen sollen, Robotern – auf Grund ihrer Wichtigkeit im (sozialen) Alltag (siehe: 1.1.) – in der ethischen Diskussion einen Raum ethischer Berücksichtigung zu geben (Gunkel, 2012). Loh (2017a) erwähnt mit einem Mathenozentrismus „eine weitere Sicht zur Lokalisierung von Phänomenen im moralischen Universum [...], die alle Wesen mit einem Eigenwert bemisst, die lernfähig sind.“ (S.33) (Anm.: Da die ausführlichere Beschäftigung mit diesen Strategien bzw. Argumentationen schnell in die moral agency-Debatte führt, werden diese aus Gründen der logischen Strukturierung der Arbeit hier vorerst nur genannt und im Folgenden [siehe v.a.: 2.2.3.2] genauer behandelt): Dies kann durch graduelle Sichtweisen auf Moral (bzw. moral agency-Konzepte) geschehen (z. B. bei: Floridi & Sanders, 2004), durch Herabsetzung des Abstraktionsniveaus für Moral (bzw. moral agency) (z. B. bei: Floridi & Sanders, 2004), alternative Moralkonzepte (z. B. bei: Wallach & Allen, 2009) oder ein relationales Moralverständnis (z. B. bei: Coeckelbergh, 2014). (Anm.: Die Frage, ob bzw. inwiefern Roboter als moralische Entitäten – in Rekurs auf dieses Kapitel – verstanden werden, wird im Rahmen der Textanalyse mitberücksichtigt [siehe: 3.2., Tab.5, B I-II].)

2.2.3. Agency, moral agency bzw. moral agents

Floridi (2011) definiert in seiner Auseinandersetzung mit artifiziellen Handlungssubjekten moral agents sehr allgemein als „the class of all entities that can in principle qualify as sources or senders of moral action.“ (S.184). Die Frage, welche Entitäten als Quellen moralischer (im Gegensatz zu bloß kausal verursachter) Handlungen gelten können, hängt häufig davon ab, „über welche Kompetenzen sie [Anm.: Roboter] hierfür in welchem Maße verfügen müssen.“ (Loh, 2017a, S.21). Im Folgenden wird versucht, moral agency, damit in Zusammenhang stehende Grundpositionen und Kompetenzen näher im Bereich der Roboterethik zu beleuchten, bevor es im nächsten Abschnitt (siehe: 2.2.4.) im Speziellen um das Kriterium der (moralischen) Verantwortung geht. (Anm.: Diese Auseinandersetzung dient v.a. der späteren Einordnung bzw. Beurteilung des moral agency-Verständnisses, das den zu analysierenden Texten bzw. einem Verantwortungs-Konzept zu Grunde liegt [siehe: 3.2., Tab.5, B I; B II, b; B III, d].)

2.2.3.1. *Moral agency, moral patiency und relationale Sichtweisen in der Roboterethik*

Wenn man die Frage stellt, inwiefern Roboter nicht nur von moralischer Signifikanz sind, sondern ob bzw. über welchen moralischen Status sie selbst verfügen (könnten), so betritt man wichtige Arbeitsbereiche der Roboterethik: Das Gegenüber eines moralischen Aktes, das eine gewisse Handlung erfährt/ erduldet, ohne dass dieses dabei selbst moralisch handeln kann, wird als moralisches Handlungsobjekt bzw. moral patient bezeichnet (Gunkel, 2012; siehe: 2.1.3.). Während die Debatte um Roboter als moral patients inklusiver geführt wird (Gunkel, 2012), geht es im Bereich der moral agency um eine exklusivere Angelegenheit, nämlich um die Frage, ob/ inwiefern Roboter (rezent bzw. zukünftig) selbst moralische Akteure/ Handlungssubjekte – moral agents – sein können, d.h. aktive Träger moralischer Pflichten (Loh, 2017a).

Wenn es um den moralischen Status von artifiziellen Systemen geht, dann greift eine bloße Unterscheidung zwischen Roboter als moral agents oder moral patients jedoch zu kurz: Von einer Kritik an einem traditionellen/ kartesischen Moralverständnis (siehe: 2.1.1.) ausgehend, beschreibt Coeckelbergh (2014) einen relationalen, nicht-kartesischen moralischen Ansatz im Kontext von Robotern, „*which sees moral status as something that emerges through relations between entities.*“ (S.64). Entitäten können demzufolge nicht ohne Referenz auf (soziale, natürliche) Beziehungen moralisch definiert bzw. beurteilt werden. In direktem Bezug auf Roboter bedeutet das, dass deren moralischer Status nicht durch bloße Betrachtung isolierter Fähigkeiten (z. B. kognitiver Natur) verstanden werden kann, sondern nur in einem kontextualisierten Sinne, d.h. im Sinne von Relationen zu/ mit Menschen/ anderen Maschinen bzw. der (natürlichen, materiellen, sozialen, kulturellen etc.) Situiertheit des Roboters. Da die Beurteilung der Relationen (und somit des moralischen Status eines Roboters) letztlich auf der Subjektivität der Menschen (als epistemologische Subjekte) beruht, ist der moralische Status als ein subjektives Konstrukt zu betrachten (Coeckelbergh, 2014). Ein solcher Ansatz sollte jedoch nicht dogmatisch verstanden werden, kann aber dabei helfen, der (moralischen) Betrachtung von Mensch-Roboter-Beziehungen nötigen Raum zu geben (Coeckelbergh, 2014).

Die folgende Tabelle (siehe: Tab.1) nennt die oben beschriebenen Bereiche/ Sichtweisen im Kontext des moralischen Status von Robotern (siehe: Tab.1, Spalte 1) und charakterisiert diese in wesentlichen Zügen (siehe: Tab.1, Spalte 2).

Tab.1: Grundlegende Bereiche der Roboterethik bzw. Sichtweisen auf Roboter in Bezug auf deren moralischen Status: moralischer Status des Roboters (Spalte 1); der dem jeweiligen Status typischerweise zugehörige Arbeitsbereich bzw. dessen Charakterisierung (Spalte 2); (Anm.: erstellt unter Bezugnahme auf: Coeckelbergh, 2014; Gunkel, 2012; Loh, 2017a; Loh, 2017b;)

<i>mor. Status des Roboters</i>	<i>Arbeitsbereich bzw. Charakterisierung</i>
moral patient	Roboter als mor. Handlungsobjekt; Roboter besitzt höchstens gewisse (indirekte) mor. Rechte, aber keine mor. Pflichten;
moral agent	Roboter als (zumindest basales) moralisches Handlungssubjekt; Roboter besitzt gewisse mor. Rechte und Pflichten;
relational	mor. Status des Roboters ist nur in einem kontextualisierenden Sinne verstehbar; Beziehungen (u.a. zwischen Mensch-Roboter) konstituieren mor. Status des Roboters als ein (epistemologisch bedingtes) subjektives Konstrukt;

2.2.3.2. (Moral) agency-Konzepte in der Roboterethik

Während traditionelle moral agency-Debatten v.a. historisch bedingt (fast ausschließlich) für Menschen reserviert sind und hier wenig bis gar keinen Raum für die ethische Diskussion artifizierlicher Systeme bieten (siehe: 2.1.3.), wird eine solche Auffassung von zahlreichen RoboterethikerInnen skeptisch betrachtet (z. B. Asaro, 2006; Coeckelbergh, 2014; Floridi & Sanders, 2012; Gunkel, 2012): Roboter sollten auf Grund ihrer (zunehmenden) Komplexität, ihres Vermögens und v.a. ihrer großen (sozialen) Relevanz im Alltag nicht kategorisch aus der moral agency-Debatte ausgeschlossen werden (Gunkel, 2012). Außerdem bestehen erkenntnistheoretische/ epistemologische Probleme in der Begründung traditioneller bzw. anthropozentrischer moral agency-Konzepte (Coeckelbergh, 2014; Gunkel, 2012). Vor diesem Hintergrund begegnet man in roboterethischen Texten vielfach Überlegungen bzw. Strategien (siehe: unten, i-iii), die im Folgenden geschildert werden und die viel (moralisches) Diskussionspotenzial mit sich bringen, denn moral agency ist eine philosophische Kategorie, die traditionellerweise Kriterien für die Zuschreibung von menschlicher Verantwortung mit sich bringt (Gunkel, 2012). Moral agency-Konzepte auf

artifizielle Entitäten auszuweiten bringt demnach neue Herausforderungen mit sich (Allen, Wallach & Smit, 2011; Gunkel, 2012).

Einige der zentralen Strategien bzw. Charakteristika der roboterethischen moral agency-Debatte werden im Folgenden dargestellt. Aus Platzgründen ist es nur ansatzweise möglich, auf damit in Zusammenhang stehende Probleme bzw. Kritikpunkte einzugehen bzw. zu verweisen (was an entsprechenden Stellen getan wird; siehe: unten). Die folgende Darstellung soll v.a. der späteren Textanalyse bzw. Diskussion (siehe: 4.-5.) – im Sinne einer Einordnung der einem Verantwortungsverständnis zu Grunde liegenden moral agency-Auffassung – dienen und kann insofern einen wertvollen Beitrag zur Beantwortung der Fragestellung leisten:

i) *Das (moral) agency-Konzept als Kontinuum*: Während bereits darauf eingegangen wurde, dass (moral) agency-Konzepte binär oder graduell aufgefasst werden können (siehe: 2.1.3.), besteht bei roboterethischen DenkerInnen vielfach die Bevorzugung bzw. Argumentation (für) letztere(r) Sichtweise (z. B. bei: Asaro, 2006; Wallach & Allen, 2009). Asaro (2006) arbeitet beispielsweise mit einem Kontinuum, das von nicht-moralischen bis hin zu voll-autonomen moralischen AkteurInnen reicht und meint: „*robots might move gradually along this continuum as they acquire greater capabilities and ethical sophistication.*“ (2006, S.9). Ein mehrstufiges Konzept bringt den Vorteil mit sich, dass nicht nur Menschen (Anm.: Diese werden weithin als genuine moral agents anerkannt [Gunkel, 2012]) auf der Grundlage eines hohen moral agency-Grades ethisch betrachtet werden können, sondern auch über abgestufte bzw. rudimentäre Formen moralischer Akteurschaft – wie diese vermeintlich bei einigen Robotern anzutreffen sind – nachgedacht werden kann (Gunkel, 2012). Damit in Zusammenhang stehen u.a. Einwände, die zu bedenken geben, ob/ inwiefern bei solchen rudimentären Formen überhaupt noch die Rede von (moralischer) Handlungsfähigkeit sein kann oder ob die Debatte damit nicht in eine andere Richtung gelenkt wird (vgl. Moral/ [moral] agency im philosophischen Kontext, siehe: 2.1.2., 2.1.3.).

ii) *Herabsetzung des Abstraktionsniveaus*: Methoden der Abstraktion, d.h. der Herabsetzung eines vorherrschenden Verständnisses – in diesem Fall von Moral bzw. moral agency – finden sich bei mehreren roboterethisch relevanten DenkerInnen (z. B. bei: Floridi & Sanders, 2004; Wallach & Allen, 2009). Dabei wird häufig – wie beispielsweise bei Floridi und Sanders (2004) – ein zu stark an einer *Philosophy of Mind* (Dennett, 1998) ausgerichtetes Moral- bzw. moral agency-Verständnis kritisiert und ausschlaggebende Kriterien für Moral bzw. moralische Akteurschaft im Sinne einer „*mind-less morality*“ (Floridi & Sanders, 2004, S.1) heruntergebrochen. Dies bringt – ähnlich wie in einer graduellen Sichtweise des moral agency-Konzepts (siehe: oben) – den Vorteil

mit sich, Roboter als (potenziell) moralische Handlungssubjekte (mehr oder weniger) unabhängig von anthropozentrisch überladenen Begriffen bzw. Konstrukten betrachten zu können. In der roboterethischen Debatte bestehen diesbezüglich jedoch geteilte Meinungen: Während Floridi und Sanders (2004) im Prinzip jedes Abstraktionslevel als sinnvoll erachten, das sich für gewisse Zwecke als nützlich erweist, betont etwa Johnson (2011) *„that certain levels of abstraction are not relevant to debate about the moral agency of computers, in particular, those levels of abstraction that separate machine behavior from the social practices of which it is a part and the humans who design and use it.“* (S.172). KritikerInnen beschäftigen sich außerdem u.a. mit Fragen, inwiefern die Abstraktion oder der Transfer von Konzepten philosophisch haltbar sein können und ob eine heruntergebrochene Form von Moral oder moral agency die Debatte nicht (terminologisch oder kontextuell) verfälscht (z. B. Johnson & Miller, 2008) oder ob/ inwieweit die Nachahmung (vgl. *„just behaving-as-if“* [Gunkel, 2012, S.55]) bzw. Imitation (vgl. Turing-Test [Gunkel, 2012]) von menschlichem Verhalten moralische Relevanz haben könnte.

iii) *Alternative Moral-/ moral agency-Konzepte*: Traditionelle Konzepte geben Robotern kaum bis gar keine Möglichkeit zur moralischen Beachtung (siehe: oben). Wenn es rezent auch fraglich ist, ob bzw. inwiefern artifizielle Systeme über Moralfähigkeit oder moral agency verfügen, so betonen dennoch zahlreiche RoboterethikerInnen (z. B. Gunkel, 2012; Lin, 2012), dass es problematisch wäre, den zunehmenden Einfluss, den Roboter bereits in der realen Welt haben (bzw. zukünftig wahrscheinlich in noch höherem Maße haben werden [siehe: 1.1.]), moralisch komplett außer Acht zu lassen und sich stattdessen etwa (ausschließlich) auf metaphysische Konstrukte (z. B. Kants kritische Vernunft) zu konzentrieren. Im Gegensatz dazu stehen alternative Konzepte:

So klammern beispielsweise funktionale Konzepte von moral agency Fragen rund um maschinelles Bewusstsein, Personenstatus etc. aus und haben einen stärker pragmatisch-praktischen bzw. utilitaristisch ausgerichteten Zugang. Beispielsweise schreiben Lin, Bekey und Abney: *„Sophisticated robots will require a kind of functional morality, such that the machines themselves have the capacity for assessing and responding to moral considerations.“* (2008, S.26). Sehr ähnlich charakterisieren auch Wallach und Allen (2009) in diesem Zusammenhang funktionale moral agents, die – im Sinne einer funktionalen Äquivalenz (Loh, 2017a) – zumindest über eine gewisse *„capacity for assessing and responding to moral challenges“* (Wallach & Allen, 2009, S.9) verfügen müssen.

Außerdem gibt es Konzepte von operationaler Moralität (Wallach & Allen, 2009) bzw. operationalen moral agents, in denen Roboter im Sinne einer *„capacity for making moral judgments“* (Lin,

Bekey & Abney, 2008, S.26) diskutiert werden, hierbei jedoch unter direkter Fremdkontrolle stehen. Ob bzw. inwiefern agency in solchen Konzepten gerechtfertigterweise als moral agency gelten kann (bzw. könnte) oder die Debatte in eine ganz andere Richtung abgeleitet, bleibt dabei kritisch zu betrachten. (Anm.: Damit beschäftigen sich u.a. Gunkel [2012] oder Johnson [2011].)

Zur Übersicht sind in der folgenden Tabelle unterschiedliche Formen/ Modelle/ Konzepte von moral agency (die in der Roboterethik eine zentrale Rolle spielen) tabellarisch zusammengefasst (siehe: Tab.2): Dabei geht es darum, die jeweilige Auffassung in Zusammenhang mit der typischen Charakterisierung (siehe: Tab.2, Spalte 2) sowie mit Beispielen dazu (siehe: Tab.2, Spalte 3) zu bringen. (Anm.: Diese tabellarische Aufarbeitung ermöglicht außerdem eine übersichtlichere Analyse des [einem Verantwortungsverständnis zu Grunde liegenden] moral agency-Konzepts im roboterethischen Diskurs, das im Rahmen der Arbeit eine Rolle spielt [siehe: oben].)

Tab.2: Mögliche Modelle/ Formen von moral agency im Kontext der Roboterethik: Form von moral agency (Spalte 1); Charakterisierung der jeweiligen moral agency-Form (Spalte 2); Beispiele, die der jeweiligen Form von moral agency zugeordnet werden können (Spalte 3); (Anm.I: erstellt unter Bezugnahme auf: Floridi & Sanders, 2004; Gunkel, 2012; Lin, Bekey & Abney, 2008; Wallach & Allen, 2009; Anm.II: ergänzt durch Verweise auf relevante Kapitel der Arbeit;)

<i>Form von m.a.</i>	<i>Charakterisierung</i>	<i>Beispiele</i>
genuin	m.a. im traditionellen Sinne (siehe: 2.1.3., 2.2.3.)	Mensch als traditioneller/ genuiner m.a. (siehe: 2.1.3., 2.2.3.)
funktional	m.a. als Kapazität/ Vermögen, moralische Herausforderungen zu beurteilen bzw. zu bewerten; Argument funktionaler Äquivalenz (siehe: 2.2.3.2.); v.a. praktisch/ pragmatisch ausgerichtet (siehe: 2.2.3.2.);	einige komplexe/ hochentwickelte Roboter als funktionale m.a.
operational	m.a. als Kapazität/ Vermögen, moralische Urteile zu fällen (siehe: 2.2.3.2.); m.a. bleibt unter Fremdkontrolle (siehe: 2.2.3.2.);	verschiedene Roboter bzw. artifizielle Systeme, wie z. B. autonomes Fahrassistenzsystem als operationaler m.a.

2.2.3.3. *Übersicht über philosophische Positionen in der roboterethischen moral agency-Debatte*

Konkret können nach Sullins (2006) vier verschiedene philosophische Positionen in Bezug auf Roboter und moral agency unterschieden werden, die hier im Folgenden im Sinne eines vergleichenden Überblicks dargestellt werden sollen: (Anm.I: geordnet nach dem Grad an moral agency, der Robotern in der Debatte zugesprochen wird; Anm.II: Die Darstellung dieser Positionen soll v.a. der späteren Textanalyse [siehe: 3., 4.] – im Sinne einer Einordnung des einem Verantwortungsverständnis zu Grunde liegenden moral agency-Verständnisses dienen [siehe: 3.2., Tab.5, B II, b; III, d].)

Roboter sind weder jetzt noch in Zukunft in der Lage, über moral agency zu verfügen (Johnson, 2011; siehe: Tab.3, Zeile 2). Argumente, die in diese Richtung gehen, führen häufig an, dass artifizielle Systeme letztlich nichts tun können, was ihnen nicht zuvor einprogrammiert wurde, weshalb sie letztlich keinen autonomen Willen besitzen (Bringsjord, 2007). Konzepte, wie Personalität und Subjektivität werden in solchen Positionen auch häufig angeführt, um aufzuzeigen, dass Roboter – v.a. im Gegensatz zu Menschen – keine moral agents sind bzw. sein können (Bringsjord, 2007). Insgesamt spielen anthropozentrische Elemente in derartigen Argumentationen häufig eine zentrale Rolle. Der Umstand, dass Roboter selbst keine moral agents sein können, bedeutet jedoch nach Johnson (2011) nicht automatisch auch, dass diese gar keine zu berücksichtigende Komponente von moralischer Akteurschaft wären (siehe: 2.2.2.).

Eine weitere Position (siehe: Tab.3, Zeile 3) besteht darin, dass Roboter zum derzeitigen Stand (noch) nicht als moral agents betrachtet werden, dies aber potenziell in Zukunft (zumindest in einem rudimentären Sinne) sein könnten (Dennett, 1998). Autoren, wie Dennett (1998) sprechen in diesem Kontext von einem gewissen Geisteszustand (engl. *state of mind*), der u.a. motivationale und kognitive Aspekte und höhergradige Intentionen (etwa Metaüberzeugungen) umfasst, die Roboter aktuell (noch) nicht besitzen und daher derzeit keine moral agents darstellen. Die Berücksichtigung der zunehmenden Komplexität und Vermögen artifizieller Systeme (siehe: 1.1.) unterstützen Sichtweisen, die die potenzielle moralische Handlungsfähigkeit zukünftiger Roboter (mehr oder weniger) optimistisch beurteilen.

Freier Wille, Intentionalität und andere geistige Konzepte sind ungelöste Themen in der Geistesphilosophie und daher nicht geeignet für die Anwendung bei Robotern (Floridi & Sanders, 2004). Setzt man hingegen die Abstraktions- und Toleranzlevel angemessen herab (siehe: 2.2.3.2.), so

können auch artifizielle Entitäten – wie etwa Roboter – als agents angesehen werden (siehe: Tab.3, Zeile 4): Beispielsweise verweisen Floridi und Sanders (2004) darauf, dass man davon ausgehen könne, dass agency vorliege, wenn Handlungen einer Entität interaktiv, autonom und anpassungsfähig seien. Liegt zudem ein bestimmter Grad an moralischer Entscheidungskraft vor, dann sind auch Handlungen dieser Entitäten im Sinne moralischer Kategorien zu berücksichtigen (Floridi & Sanders, 2004). Diese und ähnliche Positionen stehen außerdem eng in Zusammenhang mit graduellen und alternativen Konzepten rund um Roboter und moral agency (siehe: 2.2.3.2.).

Den Gegenpol zur zuerst geschilderten Position stellen Sichtweisen dar, die Roboter als moral agents ansehen, Menschen hingegen nicht (Nadeau, 2006; siehe: Tab.3, Zeile 5). Dem zu Grunde liegt die Überlegung, dass Handlungen nur dann als wirklich frei zu bezeichnen seien, wenn die damit in Zusammenhang stehenden Gründe das Ergebnis eines streng logisch operierenden Akteurs/ einer streng logisch operierenden Akteurin sind (Nadeau, 2006). Setzt man den freien Willen als Prämisse für moral agency an und fasst man Freiheit im Sinne Nadeaus auf, dann wird klar, warum Roboter hier als moral agents gelten, Menschen jedoch nicht (oder zumindest viel eingeschränkter).

Die folgende Tabelle (siehe: Tab.3) versteht sich als ein Überblick über die oben geschilderten philosophischen Positionen im Bezug auf das Verhältnis zwischen Roboter und moral agency (siehe: Tab.3, Spalte 1), damit in Zusammenhang stehende Charakteristika und Implikationen (siehe: Tab.3, Spalte 2) sowie eine Auswahl an (roboterethisch relevanten/ bedeutsamen) VertreterInnen, die der jeweiligen Position zugeordnet werden können (siehe: Tab.3, Spalte 3).

Tab.3: Überblick über zentrale Positionen in der philosophischen Auseinandersetzung mit moral agency bei Robotern: Verhältnis – Roboter zu moral agency (Spalte 1); der jeweiligen Position zu Grunde liegende Charakteristika und Implikationen (Spalte 2); Auswahl von VertreterInnen, die der jeweiligen Position zugeordnet werden können (Spalte 3); (Anm.I: erstellt unter Bezugnahme auf: Floridi & Sanders, 2004; Gunkel, 2012; Johnson, 2011; Sullins, 2006; Wallach & Allen, 2009; Anm.II: ergänzt durch Verweise auf relevante Kapitel der Arbeit;)

<i>Verhältnis: Roboter – moral agency</i>	<i>Charakteristika und Implikationen</i>	<i>VertreterInnen (Auswahl)</i>
Roboter sind keine moral agents (jetzt und zukünftig)	anthropozentristische Argumente; autonomer Wille, Personalität, Subjektivität etc. als Kriterien für m.a. (siehe: 2.1.1.-2.1.3.; 2.2.3.);	Bringsjord, 2007; Johnson, 2011
Roboter sind momentan keine moral agents, aber potenziell in der Zukunft	motivationale, kognitive Aspekte, höhergradige Intentionen als Kriterien für m.a.; besondere Berücksichtigung (der Potenzialität) zukünftiger Entwicklungen (siehe: 1.1., 2.2.3.);	Dennett, 1998
Roboter sind (zu einem gewissen Grad) moral agents	Kritik an geistesphilosophischen Konzepten von m.a., Herabsetzung des Abstraktionslevels von m.a., graduelle, alternative Konzepte von m.a. (siehe: 2.2.1.-2.2.3.);	Floridi & Sanders, 2004; Wallach & Allen, 2010
Roboter sind moral agents, Menschen nicht	streng logisch operierender Akteur als m.a.; Freiheit als Ergebnis eines solchen Akteurs (siehe: 2.2.3.);	Nadeau, 2006

2.2.3.4. Kriterien für moral agency

Sind Roboter moral agents? Wenn diese Frage nicht von vornherein negiert wird, sondern Raum für Diskussion besteht (siehe: 2.2.3.1-2.2.3.3.), so „[...] wird danach gefragt, inwiefern Roboter zu moralischem Handeln fähig sind und über welche Kompetenzen sie hierfür in welchem Maße verfügen müssen.“ (Loh, 2017a, S.21). Da das Konzept von moral agency sowie dafür als nötig erachtete Kriterien graduell aufgefasst werden können (Loh, 2017b; siehe: 2.2.3.2.) und da zukünftige Entwicklungen nicht außer Acht gelassen werden sollten (siehe: 2.2.3.3.), könnte die eingangs erwähnte Frage etwas präziser formuliert lauten: Können bzw. inwieweit könn(t)en einige Roboter rezent oder zukünftig als moral agents gelten? Diese Ausdrucksweise berücksichtigt die vielfach innerhalb (z. B. Wallach & Allen, 2009; siehe: 2.2.3.2.) und außerhalb des roboterethischen Diskurses (z. B. Singer, 1976; siehe: 2.1.2.) gestellte Forderung eines graduellen moral agency-Konzepts (vgl. oben: inwieweit) und verweist auf eine Miteinbeziehung der Reflexion potenzieller Entwicklungen (vgl. oben: könn[t]en bzw. zukünftig).

Während in traditionellen philosophischen Zugängen Kriterien für moral agency häufig anthropozentrischer oder ratiozentrischer Natur sind (Gunkel, 2012; siehe: 2.1.3.), wurde bereits erwähnt, dass in der Roboterethik mehrfach Anstrengungen für inklusivere Sichtweisen unternommen werden (siehe: 2.2.3.2.). Da moral agency-Konzepte jedoch auch innerhalb des roboterethischen Diskurses nicht einheitlich geführt werden (siehe: 2.2.3.3., Tab.3), ist es nicht möglich, von eindeutigen Kriterien in diesem Zusammenhang zu sprechen. Es können hier jedoch gewisse Kompetenzen immer wieder angetroffen werden, z. B. (moralische) Verantwortung (Asaro, 2006; Gunkel, 2012; Lin, Abney & Bekey, 2012), Autonomie (Asaro, 2006; Floridi & Sanders, 2004; Lin, Abney, Bekey, 2012), Empfindlichkeit/ Empfänglichkeit für Moral (Wallach & Allen, 2009), Interaktivität (Floridi & Sanders, 2004), Anpassungsfähigkeit (Floridi & Sanders, 2004) etc. Dabei ist anzumerken, dass diese Kriterien im Diskurs nicht (immer) einheitlich/ eindeutig/ in gleicher Anzahl/ von selber hierarchischer Bedeutung etc. argumentativ zum Einsatz kommen und außerdem (mehr oder weniger) stark von den traditionell philosophischen Bedeutungen abweichen können.

Da die nähere Auseinandersetzung mit all diesen Kompetenzen (siehe: oben) den Rahmen sprengen würde, wird im Folgenden – bzw. im Sinne des Ziels der vorliegenden Arbeit – auf die Bedeutung des Kriteriums der (moralischen) Verantwortung im moral agency-Diskurs der Roboterethik näher eingegangen (siehe: 2.2.4.). Trotzdem soll dies nicht den Eindruck vermitteln, dass es sich bei den einzelnen Kriterien um jeweils völlig voneinander isolierte Vermögen handelt: So

können (je nach begrifflicher Auffassung) beispielsweise Verantwortung und Autonomie zu einem gewissen Grad überlappen und sind nicht unbedingt/ zur Gänze unabhängig voneinander zu denken (Langbehn, 2017). Das zeigt sich unter anderem auch daran, dass beispielsweise Sombetzki (2014) Handlungsfähigkeit als eine wichtige Bedingung für die Zuschreibung von Verantwortung(sfähigkeit) auffasst, wobei die Fähigkeit zu handeln u.a. von Freiheit – welche eng mit dem Autonomiebegriff verbunden ist – abhängt (siehe: 2.1.4.5. ii, e).

2.2.4. (moralische) Verantwortung

Im Kontext artifizieller Systeme gelten in einer traditionellen Perspektive menschliche EntwicklerInnen als verantwortliche Parteien, während Roboter häufig in einem instrumentalisierenden Sinne als bloße Werkzeuge bzw. technologische Artefakte aufgefasst werden (Matthias, 2004; siehe: 2.1.1., 2.2.1.) und in diesem Sinne höchstens von kausaler, nicht aber von normativer Verantwortung die Rede sein kann (siehe: 2.1.4.2. ii). Im letzten Abschnitt wurde jedoch bereits erwähnt, dass (moralische) Verantwortung dennoch häufig ein zentrales Kriterium darstellt, wenn es darum geht, Roboter im Zusammenhang mit moral agency-Konzepten zu betrachten. Mit der Frage, welche Bedeutung (moralische) Verantwortung im philosophischen Kontext einnimmt, hat sich bereits das Kapitel 2.1.4. ausführlicher beschäftigt. Im Folgenden geht es nun darum, zu erläutern, warum bzw. in welcher Form diese Kompetenz entgegen traditioneller Sichtweisen häufig von großem Stellenwert in roboterethischen Debatten ist: (Anm.I: Das Ziel besteht hierbei nicht darin, die Fragestellung der Arbeit [siehe: 1.2.3.] vorab zu beantworten, sondern – ausgehend von Tendenzen der Roboterethikdebatte und unabhängig von den späteren Analysetexten [siehe: 3.3.] – einen Überblick über die Bedeutung des Kriteriums der (moralischen) Verantwortung im moral agency-Diskurs zu geben, um u.a. darauf basierend spätere Texte analysieren und einordnen zu können. Anm.II: Die Frage, inwiefern [moralische] Verantwortung in Textanalysen – im Sinne der folgenden Inhalte – eine Rolle spielt, wird im Analyseschema mitberücksichtigt [siehe: 3.2., Tab.5, B III].)

Dafür, dass (moralische) Verantwortung in roboterethischen Debatten häufig ein Kriterium von zentraler Bedeutung für moralische Akteurschaft darstellt, gibt es zahlreiche Hinweise (z. B. bei: Asaro, 2006; Gunkel, 2012; Lin, Abney, Bekey, 2012; Loh, 2017b). Warum ist es aber so, dass diese in den Fokus vieler moralischer Überlegungen gerückt wird, wie z. B. Johnson betont: „*The idea that an individual is primarily responsible for his or her intended voluntary behavior is at*

the core of most accounts of moral agency.“ (2011, S.173). Das liegt vermutlich v.a. am Umstand, dass Roboter im Zuge technischer Entwicklungen zunehmend größer werdenden Einfluss auf verschiedenste Lebensbereiche haben, dies von traditionellen Konzepten jedoch nicht bzw. nur unzureichend berücksichtigt wird bzw. wurde (siehe: 2.2.3.). Matthias (2004) schreibt passend dazu: *„there is an increasing class of machine actions, where the traditional ways of responsibility ascription are not compatible with our sense of justice and the moral framework of society because nobody has enough control over the machine’s actions to be able to assume responsibility for them.“* (S.177). Roboter werden komplexer, fähiger sowie unabhängiger (bzw. vielleicht auch autonomer) von EntwicklerInnen, und so kann es zu Situationen kommen, in denen nicht ohne Weiteres festzustellen ist, wer (in welchem Ausmaß moralisch) verantwortlich (zu machen) ist. Diese Schwierigkeit bzw. Lücke, die sich in der Zuschreibung von (moralischer) Verantwortung auftun kann, bezeichnet Matthias (2004) als *„responsibility gap“* (S.175). Vor diesem Hintergrund scheint es wenig überraschend, dass (moralische) Verantwortung häufig zum Thema im Zusammenhang mit roboterethischen moral agency-Diskussionen wird. Doch inwiefern und in welcher Bestimmung ist dieses Kriterium hier von konkreter Bedeutung? Im Folgenden wird ein grober Einblick in wesentliche Aspekte in diesem Zusammenhang gegeben: (Anm.I: Eine umfassendere, elaboriertere Analyse der Bedeutung bzw. Bestimmung des Kriteriums der [moralischen] Verantwortung im roboterethischen moral agency-Diskurs scheint nötig, um tiefere Einblicke zu bekommen, was sich die vorliegende Arbeit letztlich zum Ziel macht [siehe: 1.2.2., 3.–6.].)

i) *Zusammenhänge zwischen dem moral agency-Konzept/ Verständnis und (moralischer) Verantwortung als Kriterium für (moralische) Akteurschaft:* (Moralische) Verantwortung wird in roboterethischen Diskussionen häufig in konstitutiven Zusammenhang mit moralischer Akteurschaft gebracht, beispielsweise wenn Bekey moralische Akteure als *„beings capable of moral responsibility“* bezeichnet (2012, S. 25). Während in diesem Zitat explizit und daher klar ersichtlich ist, dass moralische Verantwortung (als ein Kriterium) in bedeutender Beziehung zur moralischen Handlungsfähigkeit steht, ist dies nicht in allen Fällen so klar: Häufig beschäftigen sich roboterethisch relevante Texte auch implizit/ indirekt mit (moralischer) Verantwortung oder zumindest mit (formalen) Teilaspekten dieses Konzepts (siehe: 2.1.4.) im Kontext von moralischer Akteurschaft, ohne dies immer explizit zu machen: Beispielsweise sehen Floridi und Sanders (2004) u.a. Formen von Interaktivität und Autonomie als Bedingungen für moral agency, sprechen hiermit nicht ausdrücklich Verantwortung an, sehr wohl aber Aspekte von Handlungsfähigkeit (siehe:

2.1.4.4. ii), die auch zur Charakterisierung des dieser Arbeit zu Grunde liegenden Verantwortungskonzepts aufgenommen wurden (siehe: 2.1.4.4.).

Außerdem ist es stark vom zu Grunde liegenden moral agency-Konzept abhängig, ob bzw. inwiefern (moralische) Verantwortung eine Rolle als Kriterium hierfür spielt. Dem Kriterium der (moralischen) Verantwortung liegt also eine Auffassung von moralischer Akteurschaft als eine Art Prämisse zu Grunde, die es nicht zu übersehen gilt (z. B. hinsichtlich: Anthropozentrismus, Exklusivität, Inklusivität [siehe: 2.1.3.]; Unterschiede in Gradualität, Abstraktionslevel, traditionellem/ alternativem Modell [siehe: 2.2.3.]). Beispielsweise besteht bei Vorliegen einer stark instrumentalistischen Auffassung von moralischer Akteurschaft (siehe: 2.1.1., 2.1.3.) auch ein starkes Bias bezüglich Verantwortung als Kriterium, wie Gunkel (2012) deutlich macht: *„According to the instrumentalist definition [Anm.: of moral agency], therefore, any action undertaken by a computer system is ultimately the responsibility of some human agent“* (S.28). (Anm.: Die Berücksichtigung bzw. Analyse der eben angesprochenen Punkte [moral agency-Konzept als Prämisse; explizite und implizite Verweise bei der Charakterisierung/ Anwendung des Kriteriums der {moralischen} Verantwortung/ des zu Grunde liegenden moral agency-Konzepts] spielt eine wesentliche Rolle im Zuge der Textanalyse [siehe: 3.2., Tab.5, B I, II b, III d].)

ii) *Konkrete Bedeutung/ Bestimmung des Kriteriums der (moralischen) Verantwortung – Gradualität und Abstraktionslevel:* (Moralische) Verantwortung erfährt in roboterethischen Debatten unterschiedliche Priorität und Spezifizierung: So wird diese – ähnlich zum moral agency-Konzept (siehe: 2.2.3.) – von vielen RoboterethikerInnen als stufenweise bzw. graduell aufgefasst (z. B. bei: Asaro, 2006; Loh, 2017b; Loh & Loh, 2017): *„Assigning responsibility is not a binary question of ‚all or nothing‘ but one of degree.“* (Loh & Loh, 2017, S.38). Dabei ist nicht nur (moralische) Verantwortung selbst, sondern sind ebenso auch *„Alle Kompetenzen als Voraussetzung für die etwaige Zuschreibung von Verantwortung [...] graduell bestimmbar“* (Loh, 2017b, S.221). Der Grad der Zuschreibung von (moralischer) Verantwortung sollte demgemäß je nach Vorliegen bestimmter – als notwendig erachteter Bedingungen – erfolgen (Loh & Loh, 2017; siehe: 2.1.4.5.). Parallel zum moral agency-Konzept ist davon auszugehen, dass auch die notwendigen Kompetenzen für (moralische) Verantwortung(sfähigkeit) in einem mehr oder weniger stark abstrahierten Sinne verstanden werden können (Floridi & Sanders, 2004; siehe: 2.2.3.2).

iii) *Konkrete Bedeutung/ Bestimmung des Kriteriums der (moralischen) Verantwortung – Zuschreibung von Verantwortung:* Welche Bedingungen bzw. Kompetenzen als Voraussetzung für

die Zuschreibung von (moralischer) Verantwortung im Roboterethikdiskurs als notwendig erachtet werden, ist nicht eindeutig bzw. einfach zu beantworten/ zu beurteilen und kann beispielsweise in Bezug auf etymologische Aspekte des Verantwortungsbegriffs erfolgen (siehe: 2.1.4.5.). (Anm.: Für die Analyse roboterethischer Texte sollen zu diesem Zweck die von Sombetzki [2014] etymologisch herausgearbeiteten Bedingungen für die Zuschreibung von Verantwortung herangezogen werden [siehe: 2.1.4.5].)

iv) *Verantwortungs-Netzwerk*: Wie bereits erwähnt, werden in roboterethischen Debatten traditionelle Sichtweisen – in denen EntwicklerInnen der Roboter meist als die einzig verantwortliche Partei angesehen werden – zur Diskussion gestellt. Das heißt aber nicht, dass das primäre Ziel der Roboterethik darin besteht, die Lage umzukehren und Robotern alleinige bzw. (moralische) Verantwortung im vollen Sinne zuzusprechen. Vielmehr geht es darum, (moralische) Verantwortung im Sinne eines Verantwortungs-Netzwerks zu betrachten: Dieses umfasst in einigen Situationen – man denke etwa an das Beispiel des selbstfahrenden Autos (siehe: 1.1.) – „*programmers, manufacturers, engineers, operators, artificial systems, etc.*“ (Loh & Loh, 2017, S.41-42). In einer solchen Sichtweise kann niemandem die alleinige (moralische) Verantwortung zugeschrieben werden, man kann hier viel eher von einer Aufteilung von (Einzel-)Verantwortungen sprechen (Sombetzki, 2014). Jede involvierte Entität, die in einer gewissen Situation als mehr oder weniger verantwortlich gelten kann, ist auch als mehr oder weniger verantwortlich zu bezeichnen. Im Fall von hochkomplexen Robotern (die zumindest über funktionale oder operationale Formen von Moralität verfügen [siehe: 2.2.3.2., Tab.2]) kann im Rahmen eines Verantwortungs-Netzwerks über Verantwortung in diesem distributiven Sinne nachgedacht werden.

Um einen Überblick über die bisherige Betrachtung der Bedeutung des Kriteriums der (moralischen) Verantwortung im Kontext des roboterethischen moral agency-Diskurses zu bekommen (der auch für weitere Analysezwecke genutzt wird; siehe: 3., 4.), folgt zum Abschluss dieses Kapitels eine Übersichtstabelle (siehe: Tab.4): Diese unterscheidet sehr grob, ob (moralische) Verantwortung eine Rolle im Zusammenhang mit moral agency spielt (siehe: Tab.4, Spalte 1) sowie damit in Verbindung stehende Konzepte und häufige Implikationen der jeweiligen Sichtweise (siehe: Tab.4, Spalten 2-3). Um feinere Bestimmungen in diesem Kontext ermitteln zu können, wird eine eingehendere Analyse und Diskussion als notwendig erachtet (siehe: 3.-5.).

Tab.4: Überblick über wesentliche Positionen zur Bedeutung des Kriteriums der (moralischen) Verantwortung für die roboterethische moral agency-Debatte: Bedeutung von (moralischer) Verantwortung als Kriterium für moral agency (Spalte 1); zu Grunde liegendes Konzept von moral agency (Spalte 2); typische/ häufige Implikationen dieser Positionierung (inklusive Verweise auf entsprechende Kapitel der Arbeit) (Spalte 3); (Anm.I: erstellt unter Bezugnahme auf: Floridi & Sanders, 2004; Gunkel, 2012; Lin, Bekey & Abney, 2008; Loh, 2017b; Loh & Loh, 2017; Anm.II: ergänzt durch Verweise auf relevante Kapitel der Arbeit;)

Bedeutung (mor.) Verantwortung im m.a.-Diskurs	zu Grunde liegendes m.a.-Konzept	typische/ häufige Implikationen
keine Bedeutung	traditionelles m.a.-Konzept (siehe: 2.1.3.)	Anthropozentrismus (siehe v.a.: 2.1.4.1., 2.2.3.2-2.2.3.4.; 2.2.4.), Instrumentalisierungsargumente (siehe: 2.1.1., 2.2.2., 2.2.4.), binäres m.a.- bzw. Verantwortungsverständnis (siehe v.a.: 2.1.3., 2.2.3.2., 2.2.4.), Anlehnung an <i>Philosophy of Mind</i> (Dennett, 1998) (siehe: 2.2.3.3.);
von Bedeutung	nicht-traditionelles m.a.-Konzept (siehe: 2.2.3.)	Kritik an traditionellen m.a.- bzw. Verantwortungskonzepten (siehe v.a.: 2.2.3., 2.2.4.), Gradualität, Herabsetzung des Abstraktionsniveaus, alternative Konzepte in Bezug auf m.a. bzw. (moralische) Verantwortung (siehe v.a.: 2.2.3., 2.2.4.); Verantwortung im Sinne eines Verantwortungs-Netzwerks (siehe: 2.2.4.);

3. METHODE

Zu Beginn dieses Kapitels soll die für das weitere Vorgehen erarbeitete bzw. adaptierte Analyse-methode nachvollziehbar beschrieben, theoretisch begründet sowie Bezug darauf genommen werden, warum diese als geeignet – d.h. der Beantwortung der zu Grunde liegenden Fragestellung dienlich – erachtet wird (siehe: 3.1.). Danach besteht ein praktisches Ziel darin, unter Bezugnahme auf die bis hierhin geleistete Theoriearbeit (siehe: 2.) ein konkretes Analyseschema abzuleiten (siehe: 3.2.), das sich als praktikabel für eine spätere Textanalyse erweisen soll. Im nächsten Abschnitt werden die zu analysierenden Texte sowie die Kriterien für deren Auswahl bekanntgegeben (siehe: 3.3.). Zum Abschluss des Kapitels gilt es, das methodische Vorgehen hinsichtlich seiner Stärken und Schwächen kritisch zu beleuchten (siehe: 3.4.).

3.1. Entwicklung und theoretische Hintergründe der Analyse-methode

3.1.1. Philosophische Textanalyse

Die der Arbeit zu Grunde liegende Fragestellung (siehe: 1.2.3.) soll mittels einer methodisch speziell adaptierten, philosophischen Textanalyse sowie einer daran anschließenden, vergleichenden Diskussion (siehe: 5.) bearbeitet bzw. beantwortet werden. Wie diese im Detail theoretisch begründet wird, ist Thema der nächsten Abschnitte (siehe: 3.1.2., 3.1.3.). An dieser Stelle soll vorerst nur kurz darauf eingegangen werden, warum der Zugang einer philosophischen Textanalyse bzw. eines Vergleichs der Texte in dieser Angelegenheit generell als angemessen erachtet wird:

Verschriftlichte Texte sind auch in medialen Zeiten wie der heutigen das Medium philosophischer Auseinandersetzungen (Flatscher, Posselt & Weiberg, 2011). Da sich die Arbeit in einem spezifisch ethischen Diskurs (Anm.: moral agency-Diskurs der Roboterethik) bewegt und sich innerhalb dessen mit philosophischen Kriterien und Begriffen auseinanderzusetzen hat, scheint der prinzipielle Zugang, nämlich der einer Analyse ausgewählter Texte (Anm.: [Kriterien zur] Auswahl der zu analysierenden Texte, siehe: 3.3.) gerechtfertigt zu sein. Einer derartigen Abhandlung anschließend eine komparative Auseinandersetzung folgen zu lassen (im Sinne einer vergleichenden Diskussion der Textanalysen, siehe: 5.) kann zusätzliche Erkenntnisse liefern.

3.1.2. Ansatz einer philosophisch-methodologischen Triangulation

Da der Arbeit eine sehr spezifische Fragestellung zu Grunde liegt und diese mit zahlreichen theoretischen Begriffen bzw. Konzepten zusammenhängt (siehe: 2.), bedarf es einer adäquaten Methode zur Analyse des Gegenstandes. Aus mehreren Gründen, auf die im Zuge dieses Abschnitts näher eingegangen wird, wurde eine anwendungsorientierte Textanalyse-Methode entwickelt, die auf der Idee einer philosophisch-methodologischen Triangulation (Flick, 2004) beruht.

Triangulation ist ein auf Norman Denzin zurückgehender und aus der qualitativen Forschung stammender Begriff, der vereinfacht gesagt beschreibt, „*dass ein Forschungsgegenstand von (mindestens) zwei Punkten aus betrachtet [...] wird. In der Regel wird die Betrachtung von zwei und mehr Punkten aus durch die Verwendung verschiedener methodischer Zugänge realisiert*“ (Flick, 2004, S.11). Auch wenn Triangulation nicht unbedingt unterschiedliche Methoden an eine Fragestellung herantragen muss (Anm.: Es ist beispielsweise auch die Kombination von Theorien denkbar [Theorien-Triangulation].), so ist dennoch die Integration mehrerer Methoden zur Auseinandersetzung mit einer komplexen bzw. vielschichtigen Problemstellung die prominenteste Form dieser Art (Flick, 2004). Im Sinne einer philosophisch-methodologischen Triangulation lassen sich mehrere spezifisch philosophisch begründete Methoden kombinieren, um daraus ein für die jeweilige Fragestellung angemessenes Verfahren zu generieren.

Der Vorteil einer solchen methodischen Herangehensweise liegt darin, dass aus mehreren Perspektiven (gleichzeitig) auf einen Analysegegenstand geblickt werden kann (Brüsemeister, 2008). Da der Arbeit eine vielschichtig zu bearbeitende Fragestellung zu Grunde liegt und man dabei an einer Auseinandersetzung mit teilweise widersprüchlichen, voraussetzungsgeladenen, historisch gewachsenen, philosophischen, kontextspezifischen etc. Begriffen und Konzepten (siehe: 2.) wohl nicht vorbeikommt, scheint ein multimethodischer Ansatz sinnvoll, um ein breites und tiefgehendes Verständnis für den Diskurs (bzw. für die Bedeutung/ Bestimmung von [moralischer] Verantwortung im Diskurs) zu bekommen. Flick (2004) betont dabei, dass die Auswahl von (spezifischen) Methoden stets kritisch in Hinblick auf die Angemessenheit zu befragen ist. Genau dies soll im Folgenden (siehe: 3.1.3.) – im Zuge der Charakterisierung jener philosophischer Methoden, deren Anwendung als geeignet angesehen wird – geschehen.

3.1.3. Philosophische Methoden

Im Folgenden geht es darum, jene philosophischen Methoden, die im Sinne einer methodologischen Triangulation (siehe: 3.1.2.) in die Arbeit mitaufgenommen werden, zuerst theoretisch zu beschreiben. Im Anschluss wird an jede Einzelmethode je eine begründete Stellungnahme dazu abgegeben, warum und in welcher Form der Einsatz der jeweiligen Methode im Kontext der konkreten Fragestellung (siehe: 1.2.3.) als sinnvoll erachtet wird: (Anm.: Die Methoden werden aus Übersichtsgründen zwar separat beschrieben, fließen aber allesamt in die Arbeit bzw. ins Analyseschema ein [siehe: 3.2., Tab.5], sodass letztlich eine umfassende Auseinandersetzung im Sinne einer methodologischen Triangulation stattfinden kann. Im Folgenden findet sich daher am Ende jeder Methodenbeschreibung auch ein Verweis auf jenen Bereich des Analyseschemas, in dem das jeweilige Verfahren zum Einsatz kommt.)

i) *Hermeneutik bzw. hermeneutisches Verstehen*: Hermeneutische Methoden verfolgen als Auslegungskunst das Ziel, den genuinen Sinn von Texten über die Zeit weg verstehbar zu machen (Flatscher, Posselt & Weiberg, 2011). Ein wichtiges Moment ist hierbei die Kontextualisierung, d.h. die Miteinbeziehung von Umfeld, Entstehungssituation, Zeit etc. einer Analysequelle (z. B. philosophischer Text), die „*immer in einer schon gewordenen, im weitesten Sinne sozial geordneten, d.h. kulturell und ökonomisch geprägten Welt* [vorliegt]“ (Schönwälder-Kuntze, 2015, S.91). Das Ziel der Methode besteht in diesem Zusammenhang darin, das „*Verstehen eines Textes durch Offenlegung seiner gleichsam kategorialen Voraussetzungen*“ (Schönwälder-Kuntze, 2015, S.93) zu ermöglichen.

Hermeneutisches Verstehen kann als Methode für die vorliegende Arbeit im Sinne einer ersten – oben beschriebenen – Kontextualisierung der zu analysierenden Texte nützlich sein: Die Berücksichtigung der Entstehungssituation bzw. des Umfeldes (z. B. Forschungsbereich bzw. Schwerpunkt des Autors/ der Autorin/ der AutorInnen etc.) sowie der Zeit/ Aktualität/ Relevanz (Erscheinungsjahr, Aktualität des Textes, andere Quellen, die sich auf den Text beziehen) können dabei helfen (oder vorsichtiger ausgedrückt: Indizien dafür liefern), nachzuvollziehen, warum gewisse Haltungen, Sichtweisen, argumentative Tendenzen etc. (verstärkt) in Bezug auf die Begründung, Charakterisierung, Anwendung etc. eines (moralischen) Verantwortungsverständnisses eingenommen werden. Die Offenlegung dieser kategorialen Voraussetzungen (inklusive ihrer vergleichenden Diskussion: siehe: 5.2.3.) könnte somit Indizien dafür liefern, inwiefern die Bedeutung/ Bestimmung des Kriteriums der (moralischen) Verantwortung dadurch (mit)bestimmt wird.

Die konkrete Berücksichtigung einer hermeneutischen Methode im Kontext der Fragestellung erfolgt unter diesem kontextualisierenden Schwerpunkt im Analyseschema unter: 3.2., Tab.5, A I-II.

ii) *Dekonstruktives Fragen*: Schönwälder-Kuntze (2015) beschreibt diese Methode als eine In-Frage-Stellung von Konstrukten verschiedenster Art, mit dem Ziel, latente Fundamente, (konstitutive) Widersprüchlichkeiten und Inkonsistenzen aufzudecken. In philosophischen Texten geht es dabei v.a. darum, „zu zeigen, inwiefern explizit Ausgeschlossenes gerade das Fundament der betreffenden Argumentation bildet.“ (Schönwälder-Kuntze, 2015, S.155). Nach Dreisholtkamp (1999) beruhen zahlreiche philosophisch wirksame, jedoch prima facie unsichtbare Denkmuster (von Texten) auf Ausschlüssen, die es durch dekonstruktives Fragen zu analysieren bzw. kritisieren gilt.

Wie im Rahmen der Beschäftigung mit den Begriffen/ Konzepten von (moralischer) Verantwortung (siehe: 2.1.4., 2.2.4.) und moral agency (siehe: 2.1.3., 2.2.3.) gezeigt wurde, spielen Denkmuster bzw. Mechanismen der Exklusion und Inklusion (von [moralischen] Handlungssubjekten, [moralischen] Verantwortungssubjekten) vielfach eine Rolle. Eine dekonstruktive Herangehensweise an philosophische Texte kann in diesem Zusammenhang dabei helfen, explizite sowie implizite Argumentationslinien, Denkvoraussetzungen und Geltungsansprüche in der Auseinandersetzung mit dem Konzept der (moralischen) Verantwortung bzw. moral agency aufzudecken sowie kritisch zu prüfen, inwieweit in einem Text explizit Dargestelltes auch entsprechend argumentativ eingeholt wird. Durch diese In-Frage-Stellung kann die Bedeutung/ Bestimmung des Kriteriums der (moralischen) Verantwortung letztlich v.a. auf argumentativer Ebene hinsichtlich seiner philosophischen Überzeugungskraft (Stimmigkeit versus Widersprüchlichkeit) analysiert werden.

Die konkrete Berücksichtigung der Methode des dekonstruktiven Fragens kommt im Kontext der Fragestellung in mehreren Punkten im Analyseschema zum Einsatz, konkret in: 3.2., Tab.5, B I-III; C I-IV.

iii) *Analytisches Verdeutlichen*: Die Methode des analytischen Verdeutlichens hat ganz allgemein das Ziel, Struktur, Inhalt, Intention und Wirkung von diversen Texten bzw. Theoriemodellen zu untersuchen, u.a., zum Zweck der Herausarbeitung einer (philosophischen) Charakterisierung und

gegenseitigen Abgrenzung (Schönwälder-Kuntze, 2015). Theorien, Konzepte und Begriffe können hierbei – je nach Analyseschwerpunkt – hinsichtlich unterschiedlicher Aspekte analysiert werden, z. B. syntaktisch, semantisch, in Bezug auf historische Divergenzen, den Verwendungskontext etc. (Schönwälder-Kuntze, 2015). Philosophische Sub-Methoden, die im Zuge dessen zum Einsatz kommen können, sind u.a. die Begriffsexplikation (d.h. Klärung von Begriffen nach relevanten Aspekten, z. B. Etymologie, Gebrauch, Hierarchie, Begriffselemente, Begriffsrelationen etc.) sowie Textkritik, Interpretation (d.h. eine begründete, reflektierte Auseinandersetzung mit philosophischen Texten) und Argumentationskritik (d.h. logische Analyse von Theorien) (Damschen & Schönecker, 2013).

Da in dieser Arbeit vorwiegend mit einem (moralischen) Verantwortungsbegriff bzw. Verantwortungskonzept in einem formalen bzw. etymologisch abgeleiteten Sinne gearbeitet wird (siehe: 2.1.4.), kann die Miteinbeziehung der oben beschriebenen Methode hierbei v.a. dann unterstützen, wenn es darum geht, das Kriterium der (moralischen) Verantwortung im Zuge der Textanalysen hinsichtlich formaler Charakteristika zu beleuchten. Diese Aspekte (z. B. Bedingungen, Definition, Relata, Struktur, Prämissen, Abgrenzung zu anderen Begriffen etc. [siehe v.a.: 2.1.4.]) auf der Grundlage theoretischer Vorarbeiten außerhalb (siehe: 2.1.4.) und innerhalb (siehe: 2.2.4.) des spezifisch roboterethischen Diskurses (vergleichend) zu analysieren, erscheint daher wesentlich, um (komparative) Einsichten in die Bedeutung des Kriteriums der (moralischen) Verantwortung im roboterethischen moral agency-Diskurs zu bekommen.

Der konkreten Berücksichtigung der Methode des analytischen Vergleichens – im oben beschriebenen Sinne – kommt besondere Relevanz zu, was sich u.a. im konkreten Ausmaß an Fragen im Analyseschema erkennen lässt, die im Kontext dieses methodischen Verfahrens bearbeitet werden: 3.2., Tab.5, B I-III; C I-IV. V.a. im Rahmen der formalen Analyse (siehe: 3.2., Tab.5, C I-IV) ist die Anwendung dieser Methode von besonderer Bedeutung.

3.2. Analyseschema: Die Anwendung der Methode im Kontext der Fragestellung

Das in diesem Kapitel vorgestellte Analyseschema gilt für die nachfolgende Analysearbeit (siehe: 4.) als eine Art Leitfaden, der über zwei philosophische Texte gelegt wird. Nachdem theoretische sowie methodische Hintergründe bereits geklärt wurden (siehe: 2., 3.1.), soll im Folgenden das daraus resultierende Schema nur noch in kompakter und übersichtlicher Form tabellarisch dargestellt werden (siehe: Tab.5). Es ist dabei in wesentliche Analyseschritte (siehe: Tab.5, Spalte 1),

und dazugehörige Analysefragen (siehe: Tab.5, Spalte 2) gegliedert sowie um Verweise auf methodisch/ theoretisch relevante Abschnitte der Arbeit ergänzt (siehe: Tab.5, Spalten 3–4).

(Anm.: Nachdem die Texte anhand des folgenden Schemas separat analysiert wurden [siehe: 4.1.-4.2.], erfolgt eine vergleichende Diskussion der Analyseergebnisse [siehe: 5.])

Tab.5: Schema zur (textspezifischen) Analyse der Bedeutung des Kriteriums der (moralischen) Verantwortung im moral agency-Diskurs der Roboterethik: Bezeichnung des Analyseschritts (Spalte 1); Analysefrage(n) des jeweiligen Analyseschritts (Spalte 2); philosophische Methode, die dem jeweiligen Analyseschritt im Rahmen der triangulierenden Herangehensweise zu Grunde liegt (Spalte 3); Verweise auf Abschnitte der theoretischen (bzw. methodisch-theoretischen) Vorarbeit, die dieser jeweiligen Analysefrage [inhaltlich] zu Grunde liegen (Spalte 4);

<i>Analyseschritt</i>	<i>Analysefrage(n) des Analyseschrittes</i>	<i>Methodischer Verweis</i>	<i>Theoretischer Verweis</i>
A) Kontextualisierung	I) In welchem (speziellen) Arbeitsfeld bzw. mit welchen Schwerpunkten beschäftigt bzw. beschäftigen sich der/ die AutorIn bzw. AutorInnen des Textes?	Hermeneutisches Verstehen	3.1.3.i;
	II) Welche Aktualität und Relevanz hat der Text im roboterethischen Diskurs?	Hermeneutisches Verstehen	3.1.3.i;
B) Argumentative Analyse	I) Welche Auffassung liegt bezüglich Moral bzw. (moral) agency vor? I, a) Allgemeine Charakterisierung I, b) Inklusivität/ Exklusivität I, c) Gradualität, Abstraktionslevel, alternative Konzepte	Dekonstruktives Fragen, Analytisches Verdeutlichen;	2.1.2., 2.1.3., 2.2.2. 2.2.3. 3.1.3.ii, 3.1.3.iii;
	II) Wie werden Roboter charakterisiert? II, a) Allgemeine Charaktersisierung II, b) Charakterisierung in Bezug auf (moral) agency	Dekonstruktives Fragen, Analytisches Verdeutlichen;	2.1.1., 2.1.2., 2.1.3., 2.2.1., 2.2.2., 2.2.3., 3.1.3.ii, 3.1.3.iii;
	III) Welche Auffassung liegt bezüglich (moralischer) Verantwortung – im Zusammenhang mit Robotern – hinsichtlich folgender Aspekte vor: III, a) Allgemeine Charakterisierung III, b) Inklusivität/ Exklusivität III, c) Gradualität, Abstraktionslevel, alternative Konzepte III, d) Zusammenhang mit moral agency	Dekonstruktives Fragen, Analytisches Verdeutlichen;	2.1.4.1., 2.2.3., 2.2.4. 3.1.3.ii, 3.1.3.iii;

C) Formale Analyse	I) Inwiefern spielt Verantwortung im Sinne einer etymologisch abgeleiteten Minimaldefinition eine Rolle? (allgemein bzw. in Bezug auf Roboter)	Analytisches Verdeutlichen (Dekonstruktives Fragen);	2.1.4.2., 3.1.3.ii;3.1.3.iii;
	II) Inwiefern wird Verantwortung in Gleichsetzung/ Anlehnung/ Abgrenzung zu folgenden eng verwandten Begriffen/ Konzepten verstanden? (allgemein bzw. in Bezug auf Roboter) II, a) Zurechnung II, b) Pflicht II, c) Schuld	Analytisches Verdeutlichen (Dekonstruktives Fragen);	2.1.4.3., 3.1.3.ii, 3.1.3.iii;
	III) Inwiefern wird Verantwortung im Sinne eines relationalen Konzepts verstanden? (allgemein bzw. in Bezug auf Roboter)	Analytisches Verdeutlichen (Dekonstruktives Fragen);	2.1.4.4., 3.1.3.ii, 3.1.3.iii;
	IV) Inwiefern spielen folgende Aspekte als Bedingungen der Zuschreibung von Verantwortung eine Rolle, wenn es um die Charakterisierung von moral agents geht? (allgemein bzw. in Bezug auf Roboter) IV, a) Kommunikationsfähigkeit IV, b) Handlungsfähigkeit IV, c) Urteilkraft	Analytisches Verdeutlichen (Dekonstruktives Fragen);	2.1.4.5., 3.1.3.ii, 3.1.3.iii;

3.3. Auswahl der zu analysierenden Texte

Da nicht sämtliche zur Verfügung stehende Literatur zum spezifischen Diskurs eingehender analysiert werden kann, beschränkt sich diese Arbeit auf zwei ausgewählte Texte (siehe: unten). Nach welchen Kriterien die Auswahl erfolgt ist, wird im Folgenden beschrieben:

- i) *Philosophisches Thema bzw. Arbeitsbereich und AutorIn/ AutorInnen:* Der Text bzw. der Autor/ die Autorin (die Autoren/ Autorinnen) sind dem Bereich der Roboterethik oder zumindest einem damit enger verwandten Bereich (wie z. B. Maschinenethik, Technikethik, Computerethik etc.) zuzuordnen.

- ii) *Inhaltliche Aspekte*: Der publizierte Text beschäftigt sich vorwiegend mit der Frage des moralischen Status von Robotern bzw. artifiziellen Systemen und nimmt dabei auch mehrfach Bezüge auf das Kriterium der (moralischen) Verantwortung.
- iii) *Spezifisch inhaltliche Aspekte*: Aus dem Text lassen sich direkt (explizit) oder indirekt (implizit [siehe: 2.2.4.]) mehrfach Charakteristika zur Bedeutung/ Bestimmung des Kriteriums der (moralischen) Verantwortung im roboterethischen moral agency-Diskurs ermitteln (z. B. Prämissen, zu Grunde liegende Denkfiguren, historische Verweise, begriffliche/ formale Verwendung, argumentatives Vorgehen etc.).
- iv) *Aktualität*: Die Texte sind aktuell (d.h. nicht älter als 0-15 Jahre) und stammen aus einem vergleichbaren Zeitraum.

Anm.I: Die oben genannten Kriterien sollen die Auswahl der zur Analyse herangezogenen Texte (siehe: unten) im Hinblick auf die Fragestellung der Arbeit eingrenzen und – soweit möglich – Vergleichbarkeit und Objektivität herstellen. Es sollte jedoch klar sein, dass auch andere Texte, die diesen Kriterien genügen, für eine nähere Auseinandersetzung in Frage kommen.

Anm.II: Die Beschränkung auf (nur) zwei Analysetexte ist dem begrenzten Rahmen einer Masterarbeit geschuldet. Für ein aussagekräftigeres Ergebnis ist in jedem Fall eine umfassendere Arbeit vonnöten (siehe: 3.4.).

Im Folgenden (sowie im Literaturverzeichnis, siehe: 7.) sind die Quellenangaben jener Texte aufgelistet, die es – unter Berücksichtigung der oben angeführten Kriterien – im Zuge der weiteren Arbeit eingehend zu analysieren gilt (Anm.: in der Reihenfolge der Analyse angegeben):

1. Irrgang, B. (2006). „Ethical Acts in Robotics“. *Ubiquity*, 7 (34), S. 1–15.
(Anm.: Analyse des Textes, siehe: 4.1.)
2. Sullins, J. P. (2006). „When is a Robot a Moral Agent?“. In: *International Review of Information Ethics*, 6 (12), S. 23-30.
(Anm.: Analyse des Textes, siehe: 4.2.)

Anm.: Während das Kriterium der Aktualität aus dem Erscheinungsjahr der oben genannten Texte direkt erschlossen werden kann, sind die übrigen Kriterien (bzw. deren Entsprechung auf die gestellten Anforderungen) erst den genaueren Analysen (siehe: 4.) zu entnehmen.

3.4. Kritische Reflexion des methodischen Ansatzes

Eine Begründung, warum der oben beschriebene methodische Ansatz als für diese Arbeit adäquat erachtet wurde, ist bereits erfolgt (siehe: 3.1.). In diesem Abschnitt soll es nun darum gehen, dieses Verfahren hinsichtlich seiner Vor- sowie Nachteile bzw. Grenzen zu beleuchten und gegebenenfalls zu relativieren:

Wenn es darum geht, die Bedeutung des Kriteriums der (moralischen) Verantwortung im moral agency-Diskurs der Roboterethik herauszuarbeiten, dann wird der prinzipielle Zugang einer philosophischen Textanalyse als sinnvoll erachtet, da sich die spezifische Debatte zu großen Teilen in diesem Leitmedium abspielt (siehe: 3.1.1.). Es stellt sich jedoch die Frage, mit welchem Zugang ausgewählte Texte analysiert werden sollten, um das Ziel der Arbeit in angemessener Art und Weise erreichen zu können. Für die Klärung dieses Anliegens wird aus diversen Gründen (siehe: unten) auf den Ansatz einer methodologischen Triangulation zurückgegriffen, der es erlaubt, mehrere (in diesem Fall philosophische) Methoden zu integrieren, um ein Thema zu beleuchten (siehe: 3.1.2.): Dies hat den Vorteil, dass dem Analysegegenstand (zugleich) von mehreren Perspektiven aus – und somit differenzierter – nachgegangen werden kann, als dies im Falle einer einzelnen Methode möglich wäre. Gerade bei der Analyse uneinheitlicher, voraussetzungsgeladener, historisch gewachsener, philosophischer Konzepte und Begriffe – wie etwa *moral agency* oder (moralische) Verantwortung (siehe: 2.1.3., 2.1.4., 2.2.3., 2.2.4.) –, scheint eine multiperspektivische Annäherung sinnvoll zu sein. Diese erlaubt es etwa, sich mit kontextuellen, argumentativen und formalen Aspekten im Rahmen einer Textanalyse gleichzeitig auseinanderzusetzen (siehe: 3.2, Tab.5), um somit tiefgehende Einsichten in die Bedeutung/ Bestimmung des zu untersuchenden Kriteriums im Kontext des spezifischen Diskurses zu erlangen. Der daran anschließende Diskussionsteil, der die Analyseergebnisse der separat beleuchteten Texte komparativ betrachtet, bietet die Möglichkeit, auf eventuelle Tendenzen (z. B. bezüglich der Gradualität des Verantwortungsbegriffs, Hervorhebung gewisser Bedingungen für Verantwortungszuschreibung etc.) oder Unterschiede aufmerksam zu machen, die Indizien für eine(n) gewisse(n) Charakterisierung/ Konsens/ Dissens etc. bezüglich des Kriteriums der (moralischen) Verantwortung darstellen können (siehe: 5.). Dies kann die Arbeit zusätzlich bereichern, muss jedoch auf Grund der geringen Zahl an Analysetexten mit Vorsicht betrachtet werden (siehe: unten).

Die Kehrseite eines Zugangs, der mehrere (philosophische) Methoden integriert, liegt darin, dass dieser somit immer zugleich auch die Gefahr mit sich bringt, dass im Rahmen dieser Arbeit nicht

jede Methode in jener eigentlichen Tiefe und Gründlichkeit durchgeführt werden kann, wie dies beispielsweise bei einem singulären Methodenzugang möglich wäre. Außerdem ist die Auswahl der Einzelmethoden (bzw. deren jeweilige Schwerpunktsetzung) – die im Zuge der Triangulation zum Einsatz kommen (siehe: 3.1.3.) – selbst wiederum eine Angelegenheit, die stets kritisch hinsichtlich der Angemessenheit befragt werden sollte (Flick, 2004). Um die eben genannten Herausforderungen zu berücksichtigen, wird der Anspruch, jede Methode in vollem Ausmaß zu implizieren dahingehend begrenzt, die drei begründet ausgewählten Verfahren (siehe: 3.1.3) schwerpunktmäßig anzuwenden, d.h. jeweils nur jene Aspekte herauszugreifen, die im Umfang der Arbeit realistisch durchzuführen sind und zugleich bei der Beantwortung der Fragestellung dienlich sein können. So wird beispielsweise bei der Charakterisierung der hermeneutischen Methode vorwiegend auf den Aspekt der Kontextualisierung Bezug genommen und im Zuge dessen auch erklärt, warum diese im Sinne der Arbeit als nützlich erachtet wird (siehe: 3.1.3.i), während eine umfassende hermeneutische Bearbeitung der Texte den Rahmen (und somit die Möglichkeit der Implikation weiterer Methoden) sprengen würde. Warum genau die in 3.1.3. beschriebenen drei philosophischen Methoden – hermeneutisches Verstehen, analytisches Verdeutlichen und dekonstruktives Fragen – als adäquat zur Bearbeitung der Problemstellung erachtet werden, wurde bereits detaillierter an entsprechender Stelle beschrieben (siehe: 3.1.3.i – 3.1.3.iii). Trotzdem muss auch in dieser Angelegenheit die Tatsache kritisch im Hinterkopf behalten werden, dass die Schwerpunktsetzung – d.h. die Beachtung bestimmter Aspekte einer Methode bzw. Fragestellung – immer auch exklusiven Charakter hat, soll heißen: Es wird (sowohl im theoretischen Teil, als auch im Rahmen des Analyseschemas [siehe: 2., 3.2.]) großes Gewicht auf die Analyse von (moralischer) Verantwortung im Sinne eines formalen Konzepts bzw. eines etymologisch hergeleiteten Begriffs gelegt, während hierbei aus begrifflicher/ konzeptueller Sicht auch andere Zugänge möglich wären.

Im Abschnitt 3.2. wurde – basierend auf theoretischen und methodischen Vorüberlegungen – ein praktikables Schema entworfen, dass letztlich zur Analyse roboterethisch relevanter Texte eingesetzt werden soll (siehe: 4.). Ein Vorteil eines solchen Analyseschemas besteht darin, dass es einen klaren und somit gut vergleichbaren Rahmen vorgibt, mit dessen Hilfe das Textmaterial beleuchtet werden kann. Es sollte hierbei jedoch kritisch angemerkt werden, dass – ähnlich wie im letzten Absatz beschrieben – auch in einem Frageschema zu Analysezielen letztlich immer ein gewisser Grad an Subjektivität (beispielsweise bei der Auswahl jener Themen, Fragen, Schwerpunkte etc., die [nicht] behandelt/ mit aufgenommen werden) hineinspielt. Um dieses Pro-

blem zu minimieren und somit für größtmögliche Objektivität zu sorgen, wurde dem Analyse-schema eine umfassende theoretische (sowie methodische) Vorarbeit mit relevanten Konzepten und Begriffen vorangestellt (siehe: 2. bzw. 3.1.), die klar machen sollte, warum genau die gewählten Analyseaspekte/ Fragestellungen (siehe: 3.2., Tab.5) als adäquat erachtet werden, um somit die Bedeutung/ Bestimmung des Kriteriums der (moralischen) Verantwortung im roboterethischen moral agency-Diskurs zu ermitteln. Um in diesem Zusammenhang für zusätzliche Nachvollziehbarkeit und Orientierung zu sorgen, wurden die jeweiligen Punkte des Analyseschemas um (mehrfache) theoretische sowie methodische Verweise ergänzt (siehe: Tab.5, Spalten 3-4;).

Da es dem begrenzten Ausmaß einer Masterarbeit geschuldet ist, nur eine überschaubare Anzahl an Texten einer umfassenden Analyse zu unterziehen, beschränkt sich diese auf zwei (siehe: 3.3. bzw. 4.1., 4.2.). Auf die Gründe bzw. Kriterien für die Auswahl der konkreten Analysetexte wurde bereits im entsprechenden Kapitel näher eingegangen (siehe: 3.3.). Eine eingehende Analyse zweier Texte und deren vergleichende Diskussion (siehe: 4., 5.) wird Ergebnisse (und vermutlich auch Trends oder Tendenzen) zeigen können, jedoch bleibt kritisch anzumerken, dass die Aussagekraft – ob der geringen Anzahl – begrenzt ist und für validere Ergebnisse eine größer angelegte Arbeit nötig wäre.

4. DURCHFÜHRUNG DER TEXTANALYSE

Nach einer ausführlichen theoretischen sowie methodischen Vorarbeit (siehe: 2., 3.) geht es in diesem Abschnitt darum, die Textanalysen durchzuführen. Die beiden ausgewählten Texte werden dazu unabhängig voneinander (siehe: 4.1., 4.2.) mit Hilfe des Analyseschemas (siehe: 3.2., Tab.5) behandelt. (Anm.I: Die konkreten Textanalysen halten sich dabei in ihrer Struktur konsequent an das zu Grunde liegende Analyseschema [vgl. Architektur des Analyseschemas, siehe: 3.2., Tab. 5;]. Anm.II: Zur Übersicht werden größere Analyseabschnitte jeweils abschließend kurz zusammengefasst. Anm.III: Im Diskussionsteil [siehe: 5.] kommt es zu einer vergleichenden Betrachtung der Analyseerkenntnisse.)

4.1. Analyse von: *Ethical Acts in Robotics* (Irrgang, 2006)

4.1.1. Kontextualisierung

4.1.1. I) In welchem (speziellen) Arbeitsfeld bzw. mit welchen Schwerpunkten beschäftigt bzw. beschäftigen sich der/ die AutorIn bzw. AutorInnen des Textes?

(Anm.: Die folgende Beschreibung bezieht sich auf die Autorenbiographie, die dem von Irrgang [2006] verfassten Artikel in *Ubiquity* beigelegt ist.)

Der Autor des zu analysierenden Textes – Dr. Bernhard Irrgang – ist Professor für Technikphilosophie an der Technischen Universität Dresden (Deutschland). Am dortigen Institut für Philosophie lehrt er v.a. in den Bereichen der Technikphilosophie und der (hermeneutischen) Ethik. Dr. Irrgang beschäftigt sich innerhalb der eben genannten Bereiche mit Themen, wie Gentechnik, Kulturtheorie der Technik, Informationstechnologien, AI und Medizinethik. Seine Schwerpunkte liegen in philosophischen Fragen rund um Technologie und Kultur (z. B. technologische Entwicklung und frühe technologische Kulturen, STS-research [Science, Technology and Society]). (Irrgang, 2006)

4.1.1. II) Welche Aktualität und Relevanz hat der Text im roboterethischen Diskurs?

Der Text in der vorliegenden Version erschien im September 2006 in *Ubiquity* (Irrgang, 2006). Eine frühere Ausgabe wurde ein Jahr zuvor auf der Konferenz CEPE2005 (Conference of Computer Ethics: Philosophical Inquiry, Enschede [Niederlande]) veröffentlicht (Irrgang, 2006).

Innerhalb der roboterethischen (bzw. nahestehenden) Debatte(n) wird Irrgangs Text von unterschiedlichen AutorInnen zitiert bzw. darauf verwiesen, was auf eine gewisse Bedeutung/ Relevanz hindeutet, beispielsweise bei:

Anderson, M., Anderson S., L. (2011). *Machine Ethics*. New York: Cambridge University Press.

Sullins, J. P. (2006). „When is a Robot a Moral Agent?“. In: *International Review of Information Ethics*. 6 (12), S. 23–30.

Wallach, W., Allen, C. (2009). *Moral Machines. Teaching Robots Right from Wrong*. New York: Oxford University Press.

4.1.2. Argumentative Analyse

4.1.2. I) Welche Auffassung liegt bezüglich Moral bzw. (moral) agency vor?

4.1.2. I, a) Allgemeine Charakterisierung

Die Begriffe *Moral* bzw. *moralisch* (engl. *moral*, *morally*, *morality*) werden im Text 17 Mal explizit erwähnt, *agency* (im Sinne von Handlungsfähigkeit, engl. *act*, *action*) lässt sich nicht nur aus dem Titel erschließen (Anm.: *Ethical Acts in Robotics*), sondern wird durchgehend zum Thema gemacht (über 100 Mal im Text genannt). Die Begriffe *moral agency* bzw. *moral agent* gebraucht der Autor hingegen nicht expressis verbis, dennoch wird gleich zu Beginn klar, dass die Frage nach der moralischen Akteurschaft – und somit *moral agency* – zentral ist: „*To consider the question, can Computers/Robots act morally [...]*“ (Irrgang, 2006, S.1). An mehreren Stellen des Textes wird darüber hinaus deutlich (z. B. S.1, 10, 12, 13), dass, wenn es ausdrücklich um *Moral* bzw. moralische Angelegenheiten geht – wie im obigen Zitat –, auch die Frage der Handlungsfähigkeit gestellt wird.

Dass *Moral* etwas ist, das artifiziellen Systemen nicht zukommt, schreibt Irrgang nur explizit in Bezug auf Computer („*Computers are neither moral nor social beings, although their common employment can produce extensive social effects.*“ [2006, S.3]), jedoch macht die weitere Auseinandersetzung mit dem Text deutlich, dass der Autor auch Robotern höchstens moralische Signifikanz zusprechen würde (siehe: 4.1.2. I, b-c; II;).

Irrgang bezeichnet eine Handlung als „*an Interpretation construct*“ (2006, S.11) und weist mehrfach (z. B. S.4, 5) auf eine grundsätzliche Unterscheidung zwischen Handlung und Verhalten (engl. *behaviour*) hin, wobei er Verhaltensweisen als „*Similar and repetitive actions*“ (2006, S.4) bezeichnet und hierbei von einer gewissen Routine und/ oder Regelmäßigkeit ausgeht (S.4).

Wenn es um die Frage der moral agency geht – d.h., wenn es darum geht, in einem moralischen Sinne zu handeln –, weist Irrgang (in Bezug auf Ricoeur) auf eine Hermeneutik des Selbst (engl. *self*) hin (S.1): Geht es in einem solchen Konzept darum, auszumachen, was es bedeutet (moralisch) zu handeln, dann müssen Irrgang zu Folge die Fragen geklärt werden, ob eine Entität selbst handeln und ob diese Handlung evaluiert werden kann. Es gehe hier letztlich um eine „*determination of self -- i.e., the subject of spoken, action, narration and of the moral attribute.*“ (Irrgang, 2006, S.2). Eine ausführlichere Charakterisierung von (moralischer) Akteurschaft bei Irrgang (z. B. menschliche Formen von agency, Handlung im abstrakten Sinne etc.) findet sich in den nächsten Abschnitten.

4.1.2. I, b) Inklusivität/ Exklusivität

An mehreren Stellen wird deutlich, dass Irrgang in Bezug auf Moral bzw. (moralische) Handlungsfähigkeit eine exklusive, anthropozentrische Position einnimmt: Die zu Beginn erwähnte Frage, ob Computer/ Roboter moralisch handeln können (S.1), wird schon wenig später zur Frage: „*Now, are there human actions in Computer Science or in Robotics?*“ (Irrgang, 2006, S.2). Irrgang betont einen substanziellen Unterschied zwischen menschlichen Handlungen und der Art und Weise, wie Maschinen handeln (S.5). Eine nicht-menschliche Handlung könne dem Autor zufolge nur in einem abstrakten Sinne als eine Handlung gelten: „*However, a non-embodied (disembodied) (non-human) action is only in the abstract sense an action*“ (Irrgang, 2006, S.11).

Zahlreiche Stellen greifen menschliche Charakteristika bzw. Vermögen auf und beschäftigen sich im Zuge dessen damit, inwiefern diese von nicht-menschlichen Entitäten simuliert, modelliert, kopiert oder nachgemacht werden können (z. B. die Seele des Menschen [S.6], die kommunikative Fähigkeit des Menschen [S.6], menschliche Gefühle [S.6, 7], körperliche Funktionen des Menschen [S.8], menschliches Bewusstsein [S.8,9], menschliche Würde [S.9], kognitive Strukturen des Menschen [S.9], menschliche Freiheit [S.11] etc.). Dies weist stark darauf hin, dass menschliche Akteurschaft/ Kompetenzen zum Maßstab der Beurteilung von (moralischer) Handlungsfähigkeit überhaupt gemacht werden. An nichtmenschlichen Entitäten weist Irrgang mehrfach auf Mängel im Vergleich zum (moralischen) Gradmesser Mensch hin (z. B. Roboter haben

keine dem Menschen vergleichbaren Gefühle [S.7]; Unmöglichkeit, menschliche Akteure/ Akteurinnen allgemein bzw. kognitive Strukturen des Menschen zu simulieren [S.9]; Unmöglichkeit, menschliche Freiheit zu simulieren [S.11]; siehe auch: 4.1.2. II, b [Blockzitat]). An folgender Stelle wird die Kluft zwischen der moralischen Handlungsfähigkeit des Menschen und dem damit in Vergleich stehenden Unvermögen nicht-menschlicher Entitäten besonders deutlich, wenn Irrgang beschreibt, dass letztere: „*nevertheless do not finally become humans, which are after all equipped with knowledge in a theoretically philosophical sense or with action competence in a moral sense.*“ (2006, S.10).

Trotz einer derartigen Charakterisierung wird von Irrgang zwar auch thematisiert, inwiefern eine nicht-menschliche Handlung als eine abstrakte Form einer Handlung verstanden werden kann (siehe: 4.1.2. I, c), was jedoch auf eine anthropozentrische Erstrangigkeit im Bereich von (moral) agency keine direkte Auswirkung zu haben scheint (siehe: 4.1.2. I, c; II) bzw. nicht-menschlichen Entitäten keine direkte Form von (moralischer) Akteurschaft ermöglicht (siehe: 4.1.2. II).

4.1.2. I, c) Gradualität/ Abstraktionslevel/ alternative Konzepte

Irrgangs Auffassung von (moralischer) Handlungsfähigkeit (siehe: 4.1.2. I, a-b) scheint wenig Spielraum für ein graduelles Konzept, ein Konzept mit einem herabgesetzten Abstraktionslevel bzw. ein alternatives (d.h. nicht im traditionellen Sinne exklusiv auf den Menschen bezogenes) Konzept zuzulassen.

Der Autor unterscheidet (wie in 4.1.2. I, a bereits bemerkt) jedoch nicht nur zwischen Handlung und Verhalten, sondern auch zwischen einer eigentlichen Handlung (engl. *action*) und einem Handlungsmuster („*action pattern*“ [Irrgang, 2006, S.2, 8, 9, 10, 11, 12]) und charakterisiert ein solches folgendermaßen: „*an action pattern (action-scheme) is present, if at least the intentionality and the goal of action are preliminarily modelled.*“ (2006, S.2). Ein solches Handlungsmuster braucht nach Irrgang kein Handelndes Ich (S.11) und erinnert in einer derartigen Charakterisierung prima facie eher an Verhalten (siehe: 4.1.2., I, a). Wenn man ein solches Handlungsmuster – im Sinne einer Herabsetzung des Abstraktionsniveaus – als eine sehr basale, einfache Form einer (menschlichen) Handlung verstehen möchte, bleibt die Frage offen, inwiefern diese im moralischen Kontext als Handlungsfähigkeit bezeichnet werden könnte. Darauf bezieht sich Irrgang z. B. an folgender Stelle: „*Without acting I (handelndes Ich), the goal structure of action pattern can also be ethically evaluated.*“ (2006, S.11). Ein Handlungsmuster von gewisser Zielgerichtetheit/ Intention („*goal structure*“ [Irrgang, 2006, S.11]) kann also ethisch evaluiert werden.

Moral agency bzw. die moralische Beurteilung einer Handlung bleiben jedoch ausdrücklich auf Seiten des Menschen: „*If a goal structure can be reconstructed, an action is present, also if no acting I (handelndes Ich) is to be identified I [Anm.: human being] can judge whether a robot carries out a morally positive or not-positive action pattern.*“ (Irrgang, 2006, S.11).

Nach der bisherigen Auseinandersetzung ist davon auszugehen, dass Irrgang im Rahmen von agency neben menschlichen Handlungen auch nicht-menschliche Handlungen prinzipiell anerkennt, letztere jedoch nur im Sinne von abstrakten Handlungsmustern. Nicht-menschliche Handlungsmuster sind selbst nur indirekt von moralischer Relevanz, da letztlich immer nur Menschen dazu in der Lage sind, deren ethische Bedeutung zu beurteilen (vgl. Zitat: letzter Absatz).

Die Auffassung von Akteurschaft könnte bei Irrgang als schwach graduell bezeichnet werden, wobei zumindest folgende drei Stufen ausfindig gemacht werden können (siehe: oben): keine agency – abstrakte agency [vgl. Handlungsmuster] – eigentliche/ menschliche/ moral agency). Moralische Akteurschaft selbst spricht der Autor wohl exklusiv dem Menschen zu (siehe: 4.1.2. I, b), und es scheinen hierbei auch keinerlei Möglichkeiten für Gradualität, die Herabsetzung des Abstraktionsniveaus oder alternative Modelle zu bestehen, um nicht-menschliche Entitäten im Kontext moralischer Akteurschaft (auf zumindest niedrigem Niveau) zu thematisieren.

Zusammenfassung 4.1.2. I

Irrgang beschäftigt sich – wenn auch nicht immer explizit – mit Fragen rund um Moral und (moral) agency. Zahlreiche Stellen machen deutlich, dass der Autor menschliche Handlungsfähigkeit (bzw. menschliche Fähigkeiten) als den einzigen Maßstab zur Beurteilung moralischer Akteurschaft heranzieht und nicht-menschlichen Entitäten (moralische) Handlungsfähigkeit im eigentlichen Sinne abspricht, auch wenn letztere von moralischer Relevanz sein können. Insofern kann Irrgangs Auffassung von Moral und (moral) agency als exklusiv bzw. anthropozentrisch bezeichnet werden.

Nicht-menschliche Entitäten fasst der Autor bloß als agents im Kontext von Handlungsmustern auf. In diesem Sinne sind sie auch in der Lage, abstrakte Handlungen (von moralischer Relevanz) auszuführen, ohne dabei jedoch selbst moralfähige Akteure darzustellen.

Akteurschaft könnte bei Irrgang insgesamt als ein schwach graduelles, dreistufiges Konzept angesehen werden (nicht handlungsfähig – handlungsfähig im abstrakten Sinne – moralisch handlungsfähig). Moralische Akteurschaft selbst scheint exklusiv im Bereich des Menschlichen verortet sowie nicht abstrahierbar oder alternativ aufgefasst zu sein.

4.1.2. II) Wie werden Roboter charakterisiert?

4.1.2. II, a) Allgemeine Charakterisierung

Irrgang nimmt in Bezug auf Roboter bzw. Computer z. T. eine relationale Sichtweise ein: Der Autor fasst artifizielle Systeme als Teil unserer Kultur auf (S.4), deren Wirkung bzw. Einfluss nur im Kontext sozialer Gruppen zu verstehen sei (S.5). Im Text werden Roboter jedoch meist in Bezug auf isolierte Fähigkeiten charakterisiert (Anm.: häufig in Bezug auf den Maßstab Mensch [siehe: 4.1.2. I, b]). Auf der einen Seite werden Robotern an mehreren Stellen gewisse Fähigkeiten und Entwicklungspotenziale eingeräumt. So spricht Irrgang in Bezug auf artifizielle Systeme von „*an increasing authority and competence of machines*“ (2006, S.4) oder davon, dass „*sentimental robots will be designed*“ (2006, S.7) etc. Bei der Ausführung mancher routinemäßiger Handlungsmuster (v.a. unter stabilen, regelhaften Bedingungen stattfindend) sieht der Autor Roboter Menschen sogar überlegen (S.6).

Trotz dieser Anerkennung gewisser Fortschritte und Fähigkeiten artifizieller Systeme bleibt der Verfasser auf der anderen Seite einer traditionellen, instrumentalisierenden Sichtweise verhaftet, in der deutlich wird, dass Roboter Werkzeuge (des Menschen) bleiben: „*They [Anm.: robots] remain tools [...]*“ (Irrgang, 2006, S.8; siehe auch: 4.1.2. II, b [Blockzitat]). Den aktuellen Status von Robotern beurteilt Irrgang außerdem nüchtern: „*Today our machines are still simple creatures, which require (like all newborn babies and children) the parental care and welfare service, and are hardly as intelligent as they are thought to be.*“ (Irrgang, 2006, S.7). Inwiefern Roboter für den Autor als (moral) agents gelten, diskutiert v.a. der nächste Abschnitt.

4.1.2. II, b) Charakterisierung in Bezug auf (moral) agency

Wenn es um die Frage geht, ob bzw. inwiefern Roboter (aus moralischer Sicht) handlungsfähige Wesen sind (bzw. sein könnten), charakterisiert Irrgang diese vorrangig, indem er sie im Vergleich zu Menschen als Mängelwesen ausmacht (siehe: 4.1.2. I, b): Es gibt Hinweise dafür, dass

der Autor Robotern einfache Formen von Handlungsfähigkeit im Sinne einer Ausführung vorab modellierter Handlungsmuster zuspricht (siehe: S.2, 11; 4.1.2. I, c). Moralische Akteurschaft (bzw. sehr wahrscheinlich Moralfähigkeit überhaupt) bleibt Irrgang zufolge – im Gegensatz zu Menschen – jedoch selbst hochentwickelten Robotern vorenthalten, wie diese Stelle deutlich macht:

„I also assume that such universal and unspecified robots with great flexibility in the behavior patterns and a due portion at learning competence, as well as a certain humanlikeness, nevertheless do not finally become humans, which are after all equipped with knowledge in a theoretically philosophical sense or with action competence in a moral sense. Even if they are possibly a useful simulation for the everyday life of human beings. The potential of humans is more comprehensive than that of a common-sensical understanding of the technologized everyday life.“ (2006, S.10).

Da davon auszugehen ist, dass Irrgang nur Menschen, nicht aber Robotern moralische Akteurschaft zuspricht (siehe: Zitat oben bzw. 4.1.2. I, c) und der Autor zudem eindeutig betont: *„They [Anm.: robots] remain tools, and do not become doer (acting) subjects“* (2006, S.8) bzw. in seiner Konklusion festhält: *„Robots do not become humans“* (2006, S.12), ist davon auszugehen, dass er – trotz der Anerkennung gewisser Entwicklungspotenziale artifizierter Entitäten – jetzt sowie in Zukunft keine Möglichkeit darin sieht, Roboter selbst als moralische Akteure aufzufassen.

Der Argumentation des Textes zufolge ist davon auszugehen, dass Irrgang Roboter wohl auch nicht als funktionale moral agents betrachten würde, denn zuvor (siehe: 4.1.2.1, c) wurde angemerkt, dass der Autor die Beurteilung moralischer Herausforderungen als etwas ansieht, das im Bereich des Menschlichen (bzw. menschlicher Vermögen [siehe: 4.1.2.1, b]) zu verbleiben scheint. Irrgang betont mehrfach die Notwendigkeit menschlicher Kontrollierbarkeit maschineller Handlungen (siehe: 4.1.2. III, b; 4.1.3. II, b; 4.1.3. III; 4.1.3. IV, b-c). Auch aus der Perspektive einer solchen Fremdkontrolle (d.h. Maschinen bleiben unter der Kontrolle des Menschen) finden sich keine Hinweise, die dafür sprechen, dass der Autor Robotern Kapazitäten zum Fällen moralischer Urteile zugestehen und sie in diesem Sinne als operationale moral agents ansehen würde.

Zusammenfassung 4.1.2. II

Irrgang betont stellenweise eine relationale Sichtweise im Kontext artifizierter Systeme und spricht Robotern auch gewisse Fähigkeiten und Entwicklungspotenziale zu. Zugleich bleibt der Autor jedoch auch traditionellen, instrumentalisierenden Denkmustern verhaftet, und es scheint so, als könnten Roboter unter diesen Prämissen – im Gegensatz zu Menschen – weder jetzt noch

zukünftig selbst als moralische Akteure gelten (auch nicht in einem funktionalen oder operationalen Sinne).

4.1.2. III) Welche Auffassung liegt bezüglich (moralischer) Verantwortung – im Zusammenhang mit Robotern – hinsichtlich folgender Aspekte vor:

4.1.2. III, a) Allgemeine Charakterisierung

Irrgang führt eine grundsätzliche Unterscheidung in Bezug auf Verantwortung ein: „*Indeed, one must differentiate between the descriptive analysis of causality on the one hand and the normative expression of responsibility on the other hand, in the sense of the someone-real-responsible-person to be made*“ (2006, S.3). Diese Stelle macht deutlich, dass der Autor Verantwortung im normativen (d.h. nicht rein kausalen) Sinne eng an ein Konzept von Personalität knüpft und lässt vermuten, dass er den Begriff auch in einem individualistischen Sinne versteht (vgl. Zitat oben: *person, someone*). Immer dann, wenn Irrgang Verantwortung behandelt (Anm.: Im Text kommt 20 Mal das Wort *responsibility* vor), bezieht er sich auf einen normativen Gebrauch dieses Begriffs (Anm.: außer auf Seite drei, wo er Verantwortung im kausalen Sinne einführt, um diese von normativer Verantwortung zu unterscheiden). Teilweise versteht der Autor Verantwortung in Anlehnung an nah verwandte Begriffe, wie *Zurechnung*, *Pflicht* oder *Schuld* (siehe: 4.1.3. II).

Der Autor weist explizit in Zusammenhang mit Robotern bzw. deren Entwicklung auf eine gewisse Dringlichkeit hin, Verantwortungsfragen zu stellen, wie beispielsweise in folgendem Zitat: „*the degree of automization [sic] is growing. Thus, it raises the responsibility question for the behaviour of robots*“ (Irrgang, 2006, S.3). Wenn es jedoch im Zusammenhang mit artifiziellen Entitäten um (Fragen rund um) Verantwortung im normativen Sinne geht, dann thematisiert Irrgang an keiner Stelle die Möglichkeit, ob bzw. inwiefern Roboter selbst Subjekte normativer oder moralischer Verantwortung sein könnten. Letztlich geht es bei der Auseinandersetzung mit nicht-kausaler Verantwortung in Verbindung mit Robotern stets um die Klärung menschlicher Verantwortung (siehe: S.6, 10, 11, 12), wie auch folgende Stelle verdeutlichen soll: „*In order to be able to take responsibility [Anm.: as a human being], technical systems must be controllable by humans.*“ (Irrgang, 2006, S.6). (Anm.: Dass [moralische] normative Verantwortung im Text eindeutig exklusiv bzw. anthropozentrisch charakterisiert wird, beschreibt u.a. der nächste Abschnitt [siehe: 4.1.2. III, b].)

4.1.2. III, b) Inklusivität/ Exklusivität

Irrgang spricht im Zusammenhang mit Informationssystemen zwar eingangs von Verantwortung („*Today, ever more decisions in a society are left to individual information systems, because human-beings are not able to make appropriate and irresponsible decisions within extremely short response times.*“ [2016, S.3]), um diese jedoch wenig später als kausale – im Vergleich zur normativen – Verantwortung auszumachen (siehe: 4.1.2. III, a). Wie bereits im letzten Abschnitt erwähnt, werden Roboter auch im Rahmen normativer Verantwortung thematisiert, jedoch ausschließlich, um damit in Zusammenhang stehende Fragen rund um menschliche Verantwortung(sfähigkeit) zu klären. In dieser Beziehung beschäftigt sich Irrgang u.a. mit einer gewissen Notwendigkeit menschlicher Kontrolle über artifizielle bzw. technische Systeme (S.6, 7, 10; „*In order to be able to take responsibility [Anm.: as a human being], technical systems must be controllable by humans.*“ [2006, S.6]) und spricht zudem auch von Gefahren und Risiken quasi-autonomer Roboter („*autonomous‘ robots carry an actual potential of danger that can arise from them*“ [2006, S.9]). Dass menschliche Kontrollierbarkeit technischer Systeme zu einer Art Norm erhoben wird, thematisiert u.a. 4.1.3. II, b.

Des Weiteren charakterisiert Irrgang normative Verantwortung auf eine Art und Weise, die darauf schließen lässt, dass er diese als typisch menschliches Vermögen ansieht, beispielsweise, wenn er behauptet: „*human-beings are not able to make appropriate and irresponsible decisions within extremely short response times*“ (2006, S.3). Da Irrgang Subjektivität als etwas Menschliches ausmacht („*Selfness and subjectivity can be understood as a dimension of human action and the unity of self.*“ [2006, S.2]), macht folgende Stelle ziemlich deutlich, dass der Autor moralische Verantwortung wohl nicht als etwas versteht, das auch außerhalb des Menschlichen zu verorten wäre: „*In order to be morally responsible, however, an act needs a participant, who is characterized by personality or subjectivity.*“ (2006, S.12). Außerdem wird diese Vermutung auch durch die bisherige Auseinandersetzung bekräftigt, beispielsweise, wenn Irrgang Computern (bzw. artifiziellen Systemen) Moral(fähigkeit) generell abspricht (siehe: 4.1.2 I, a).

Der Autor betont, dass es sinnvoll ist, Roboter zu bauen, die spezielle Aufgaben erledigen können, macht aber auch hier in Bezug auf Verantwortung seine Position – in der auch seine generelle Sichtweise auf Roboter (als vom Menschen abhängige Werkzeuge [siehe: 4.1.2. II, a]) offenkundig wird – ein weiteres Mal deutlich: „*In such specialized functions, robots are the tools that remain dependent on humans in a certain way. [sic] and belong to the area of responsibility of humans, just like traditional technological applications.*“ (Irrgang, 2006, S.10).

4.1.2. III, c) Gradualität, Abstraktionslevel, alternative Konzepte

Wie schon erwähnt, differenziert der Autor klar eine ursächliche (kausale) von einer normativen Verantwortung, wobei letztere ausschließlich im Bereich des Menschlichen zu verbleiben scheint (siehe: 4.1.2. III, a-b). Bei einer solchen Auffassung könnte man von einem nicht-graduellen, nichtabstrakten, traditionellen Modell von Verantwortung sprechen. Wenn man Irrgangs Text genauer studiert, stößt man darüber hinaus auf weitere in Zusammenhang mit Verantwortung stehende Begriffe bzw. Unterscheidungen, wie etwa: „*aresponsible*“ (2006, S.3), „*collective responsibility of humans*“ (2006, S.10) oder „*morally responsible*“ (2006, S.12). Dies mag wiederum prima facie – ob der diversen Termini – den Eindruck einer gewissen Gradualität oder verschiedenartiger Abstraktionsniveaus/ alternativer Konzepte von Verantwortung erwecken. Wie sind diese Begriffe einzuordnen?

Der Begriff „*aresponsible*“ (Irrgang, 2006, S.3) kann als Negation von (normativ) verantwortlich aufgefasst werden. Er wird vom Autor verwendet, um festzustellen, dass Menschen – im Gegensatz zu Informationssystemen – nicht in der Lage sind, derart zu entscheiden (S.3; siehe: 4.1.2. III, b). Die Negation von (normativ) verantwortlich kann nicht als Abstufung bzw. als eine weniger abstrakte Form menschlicher Verantwortung(sfähigkeit) angesehen werden.

Die Bezeichnung „*morally responsible*“ (Irrgang, 2006, S.12) wurde schon im letzten Abschnitt (siehe: 4.1.2. III, b) als eine speziell menschliche Form/ Art normativer Verantwortung ausgemacht. Ebenso ist es naheliegend, Irrgangs Bezeichnung „*collective responsibility of humans*“ (Irrgang, 2006, S.10) als eine spezifisch menschliche Form/ Art normativer Verantwortung aufzufassen. Beide Termini bleiben somit im Bereich der menschlichen normativen Verantwortung und ermöglichen es, hierin etwaige Differenzierungen (z. B. verantwortlicher Mensch, moralisch verantwortlicher Mensch, kollektive Verantwortung von Menschen) vorzunehmen. Im Kontext nicht-menschlicher Entitäten finden sich jedoch keine Indizien dafür, dass Irrgang normative Verantwortung in abgestufter, weniger abstrakter oder in konzeptuell alternativer (d.h. nicht traditioneller) Weise auffasst. In diesem Sinne kann normative Verantwortung bei Irrgang wohl als etwas exklusiv Menschliches bezeichnet werden (siehe: 4.1.2. III, b), womit keine Möglichkeit besteht, Roboter in diesem Zusammenhang direkt (auch nicht auf einem geringen Level) zu thematisieren. Normative Verantwortung kann somit als binär/ nicht-graduell angesehen werden und umfasst die beiden Pole nicht normativ verantwortlich (nicht-menschliche Entität) und normativ verantwortlich (Mensch).

4.1.2. III, d) Zusammenhang mit moral agency

Durch die Auseinandersetzung mit Irrgangs Perspektive hinsichtlich (moral) agency und (moralischer) Verantwortung (siehe: 4.1.2. I, II, III) lassen sich einige Zusammenhänge zwischen diesen beiden Begriffen/ Konzepten erkennen bzw. vermuten:

Abgesehen von nicht-menschlichen Handlungsmustern – die der Autor höchstens in einem abstrakten Sinne als Handlungen bezeichnet – und kausaler Verantwortung besteht keine Möglichkeit, Roboter im Sinne von (moralischer) Handlungsfähigkeit bzw. (moralischer) normativer Verantwortungsfähigkeit direkt zu berücksichtigen. Beide Konzepte ([moral] agency und [moralische] normative Verantwortung) werden von Irrgang vorwiegend in einer exklusiven, anthropozentrischen, traditionellen (siehe: 4.1.2. I, b; III, b;) bzw. nicht graduellen, nicht alternativen Art (siehe 4.1.2. I, c; III, c) verstanden.

Ein exklusives (moral) agency-Konzept – wie es bei Irrgang vorzuliegen scheint – gibt der Bedeutung (moralischer) normativer Verantwortung als ein potenzielles Kriterium für (moral) agency eine restriktive Richtung: Sofern es um nicht-menschliche Entitäten (z. B. Roboter) oder Handlungen (z. B. Auswahl eines Schalters durch einen Roboter) geht, scheint die Auseinandersetzung mit Kriterien für (moral) agency automatisch nicht von direkter Relevanz zu sein (siehe: 4.1.2. III, b; vgl. „*In order to be morally responsible, however, an act needs a participant*“ [Irrgang, 2006, S.12]; Anm.: [moralische] HandlungsteilnehmerInnen im eigentlichen Sinne sind für Irrgang nur Menschen [siehe: 4.1.2. I, a-b]). Roboter können letztlich bloß kausal verantwortlich oder in Bezug auf menschliche Verantwortung(sfragen) von indirekter Bedeutung sein.

Zusammenfassung 4.1.2. III

Es gibt zahlreiche Hinweise dafür, dass Irrgang in Bezug auf (moralische) normative Verantwortung eine exklusive, anthropozentrische, traditionelle Position einnimmt. Während Roboter nur in einem kausalen Sinn als verantwortungsfähige Wesen gelten, ist davon auszugehen, dass in dieser Auffassung normative Verantwortungsfähigkeit ausschließlich im Bereich des Menschlichen zu verorten ist. In diesem Kontext erwähnt Irrgang auch unterschiedliche Formen (z. B. moralische Verantwortung, kollektive Verantwortung). Es lässt sich im Text jedoch kein Anhaltspunkt dafür finden, dass auch nicht-menschliche Entitäten als (zumindest rudimentär oder graduell) normativ verantwortlich gelten könnten, weshalb in Bezug auf normative Verantwortung ein binäres Konzept (nicht normativ verantwortungsfähig – normativ verantwortungsfähig) vorzuliegen scheint.

Normative (moralische) Verantwortung als Kriterium für moral agency scheint bei Irrgang nur im Kontext von Menschen sinnvoll diskutabel zu sein. Bei Vorliegen nicht-menschlicher Entitäten sieht es so aus, als bestünde – als logische Folge anthropozentrischer Prämissen – automatisch keine Möglichkeit, (graduelle, abstrakte) Formen von normativer (moralischer) Verantwortung(sfähigkeit) als potenzielle Kriterien für (moralische) Akteurschaft zu diskutieren.

4.1.3. Formale Analyse

4.1.3. I) Inwiefern spielt Verantwortung im Sinne einer etymologisch abgeleiteten Minimaldefinition eine Rolle? (allgemein bzw. in Bezug auf Roboter)

Da Irrgang Roboter ausdrücklich nicht als Subjekte ansieht („*They [Anm.: robots] remain tools, and do not become doer (acting) subjects.*“ [2006, S.8]), ist es auch nicht möglich, ein Rede-und-Antwort-stehendes Subjekt als VerantwortungsträgerIn bzw. Hinweise auf Intersubjektivität in diesem Kontext auszumachen. Im Vergleich dazu charakterisiert der Autor Menschen als Subjekte (S.1) und schreibt in Bezug auf moralische Verantwortung: „*In order to be morally responsible, however, an act needs a participant, who is characterized by personality or subjectivity.*“ (Irrgang, 2006, S.12). Hier werden Personalität und Subjektivität gewissermaßen als Vermögen für (moralische) Verantwortung verstanden. Irrgang spricht an einer Stelle außerdem von einer kollektiven Verantwortung des Menschen (S.10), was die Vermutung nahelegt, dass er normative Verantwortung als etwas Intersubjektives, Dialogisches begreift.

Für Irrgang hat Verantwortung mit einem normativen Geschehen zu tun (siehe z. B. S.3): Schon in 4.1.2. III, a wurde darauf hingewiesen, dass der Autor in Zusammenhang mit Verantwortung allgemein zwischen: „*the descriptive analysis of causality on the one hand and the normative expression of responsibility on the other hand, in the sense of the someone-real-responsible-person*“ (S.3) unterscheidet. Es konnte außerdem bereits ermittelt werden (4.1.2. III, b-c), dass der Verfasser normative Verantwortung sehr wahrscheinlich exklusiv (und nicht graduell) im Bereich des Menschlichen verortet sieht und Robotern hingegen höchstens kausale/ deskriptive Verantwortung zuschreibt.

Irrgang geht in Zusammenhang mit Verantwortung(sfähigkeit) wenig bzw. nur indirekt auf hierfür als nötig erachtete Kompetenzen bzw. eine bestimmte psychomotivationale Verfasstheit ein:

Es wurde bereits herausgearbeitet, dass der Autor ein anthropozentrisches Verständnis von normativer Verantwortung hat (siehe: 4.1.2. III, b). Insofern scheint naheliegend (wenn auch nicht immer explizit gemacht), dass Irrgang typisch menschliche Kompetenzen im Kontext von Verantwortung als nötig erachtet (siehe: oben). Der Autor betrachtet Menschen – im Gegensatz zu Robotern – als bewusste bzw. bewusstseinsfähige Entitäten: *„The potential of humans is more comprehensive than that of a common-sensical understanding of the technologized everyday life [...] by no means is it the case that these [Anm.: highly complex/ quasi autonomous] robots can also reach a consciousness of their actions“* (Irrgang, 2006, S.10). Bewusstsein für eine Handlung zu haben scheint für Irrgang (mit)entscheidend zu sein, um überhaupt handeln zu können (S.10). Das Konzept von agency steht wiederum in enger Verbindung mit Verantwortungsfähigkeit (siehe: 4.1.2. III, d) und insofern kann hier Bewusstsein als eine mögliche Kompetenz im Zusammenhang mit Verantwortung betrachtet werden (siehe auch: 4.1.3. IV, b). Weitere Kompetenzen, wie Reflexionsfähigkeit, Urteilkraft oder Nachdenken sind nur indirekt im Sinne von Rationalität – als einer menschlichen Kompetenz – an folgender Stelle mitangesprochen: *„they [Anm.: artificial systems like automats] lie within a range of pre-rational intelligence, within which certain human competence is to be simulated.“* (Irrgang, 2006, S.5). Hieraus lässt sich vermuten, dass Irrgang Robotern jene Kompetenzen, die mit Rationalität in Verbindung stehen (Reflexionsfähigkeit, Urteilkraft, Nachdenken) wohl nicht zusprechen würde. Inwiefern diese Fähigkeiten von Irrgang als notwendige Kompetenzen für Verantwortung aufgefasst werden, lässt sich jedoch auf Grund mangelnder Informationen nicht eindeutig aus dem Text erschließen (siehe auch: 4.1.3. IV, b).

Zusammenfassung 4.1.3. I

Es ist davon auszugehen, dass Irrgang Verantwortung als etwas Intersubjektives bzw. Dialogisches versteht, das ein Subjekt als VerantwortungsträgerIn benötigt und das eng mit dem Begriff der Persönlichkeit zusammenhängt. Diese Auffassung scheint positiv auf Menschen, jedoch nicht auf Roboter anwendbar zu sein. Der Autor fasst Verantwortung im eigentlichen Sinne als ein normatives Geschehen auf. Irrgang geht (fast) nur indirekt bzw. kaum auf Kompetenzen für Verantwortungsfähigkeit ein: Es ist jedoch – ob der anthropozentrischen Prämissen – die Vermutung naheliegend (wenn auch nicht immer explizit gemacht), dass Bewusstsein oder Rationalität hier von gewisser Bedeutung sein könnten. In einer solchen Sichtweise scheint kein Raum dafür zu bestehen, Roboter als direkt verantwortungsfähige Wesen zu betrachten.

4.1.3. II) Inwiefern wird Verantwortung in Gleichsetzung/ Anlehnung/ Abgrenzung zu folgenden eng verwandten Begriffen/ Konzepten verstanden? (allgemein bzw. in Bezug auf Roboter)

4.1.3. II, a) Zurechnung

Irrgang verwendet das Wort *Zurechnung* bzw. *zurechnen/ zurechenbar* (engl. *attribution, to attribute, attributable, ascription, to ascribe*) mehrfach – v.a. in Zusammenhang mit Robotern/ artifiziellen Systemen (S.2, 3, 8, 9, 11, 13). Beispielsweise spricht er eine „*responsibility attributed to computers*“ (2006, S.3) an und meint hier kausale Verantwortung, die manchen Computern zugerechnet werden kann. An folgender Stelle bezieht sich Irrgang auch auf eine weniger komplexe Form von Verantwortung, die kein Bewusstsein voraussetzen scheint (Anm.: Irrgang spricht Robotern generell Bewusstsein ab [S.10; siehe: 4.1.3. I]): „*a non-embodied (disembodied) (non-human) action is only in the abstract sense an action, for whose consequences are also attributable to the robot and held responsible*“ (2006, S.11). Im obenstehenden Zitat verwendet er daher zwar den Terminus *verantwortlich (responsible)*, jedoch wohl in einem zurechnend-kausalen Sinne. Dennoch gesteht der Autor ein, dass Menschen dazu geneigt sind, artifiziellen Systemen Vermögen zuzurechnen, die typischerweise eng mit Verantwortung in Verbindung gebracht werden, wie etwa folgende Stelle deutlich machen soll: „*we are inclined to attribute intelligence to them* [Anm.: computers].“ (2006, S.4).

Irrgang charakterisiert jedoch auch teilweise normative Verantwortung im Sinne einer deskriptiven Subjekt-Objekt-Zuschreibung, beispielsweise wenn er menschliche Handlungen (Handlungen eines Subjektes) als „*morally attributable (zurechenbaren) acts*“ (2006, S.13) ausmacht, um diese von „*machine regulated, implemented behavioural patterns*“ (2006, S.13) zu unterscheiden.

Dass der Autor die Begriffe *Zurechnung* und *Verantwortung* nicht unbedingt gleichsetzt, diese bei ihm jedoch in engem Zusammenhang zueinander zu stehen scheinen, macht folgende Stelle klar: „*They* [Anm. robots] *remain tools, and do not become doer (acting) subjects. How should they become subjects? (How could one attribute subject to the machines? [...]*“ (Irrgang, 2006, S.8). Der Frage, wie jemand Handlungssubjekt (und somit nach Irrgang potenziell handlungsfähig und dadurch potenziell verantwortungsfähig) werden könne, folgt sofort die Frage, wie man Maschinen Subjektstatus zurechnen könne.

4.1.3. II, b) Pflicht

Irrgang spricht nicht *expressis verbis* von Pflichten (engl. *duty, obligation*). Dennoch erinnert etwa folgende Stelle an eine Art Norm/ Regel, die es (bloß) zu befolgen gilt, wenn es darum geht, als Mensch verantwortlich sein zu können: „*In order to be able to take responsibility, technical systems must be controllable by humans.*“ (Irrgang, 2006, S.6). Unkontrollierbare Systeme zu bauen (Anm.: kann als Nichteinhaltung einer Norm aufgefasst werden), nimmt Menschen Verantwortung. Im Umkehrschluss kann es als eine Art Pflicht angesehen werden, die Regel/ Norm einzuhalten, nur durch Menschen kontrollierbare Roboter zu bauen. Artificielle Systeme werden bei Irrgang mehrfach unter dem Aspekt menschlicher Kontrollierbarkeit thematisiert (z. B. S.3, 5, 6, 10).

4.1.3. II, c) Schuld

Irrgang verwendet keine Begriffe, die in direktem Zusammenhang mit dem Schuldbegriff (engl. *guilty, culpable, [to] blame/ charge/ incriminate* etc.) stehen. Der Autor versteht artificielle Systeme wohl nur als kausal verantwortlich (siehe z. B. 4.1.2 III, d). Bei der Textstelle „*Computers are made responsible for their decisions.*“ (Irrgang, 2006, S.2) ist daher davon auszugehen, dass Irrgang Verantwortung im Sinne eines (kausalen) Schuldigseins bzw. Verantwortlichseins anspricht.

Immer dann, wenn Irrgang demnach im Zusammenhang mit Robotern (bzw. technischen Systemen) – die er als nicht normativ verantwortungsfähige Entitäten auffasst (siehe: 4.1.2. III) – von Verantwortung spricht (S.2, 3, 6, 10, 12), tut er das in enger Anlehnung an Konzepte von Kausalität und Verursachung und in dieser Hinsicht im Sinne von Schuld.

Zusammenfassung 4.1.3. II

Irrgang setzt Verantwortung nicht mit den Begriffen *Zurechnung, Pflicht* und *Schuld* gleich, es gibt jedoch einige Stellen im Text, die darauf hindeuten, dass der Autor die Begriffe teilweise in einer gewissen Nähe zueinander sieht: So wird Verantwortung (im kausalen und normativen Gebrauch) z. T. in Form einer Subjekt-Objekt-Zuschreibung begriffen, an einer Stelle im Sinne einer Einhaltung einer Norm verstanden und z. T. in Anlehnung an Konzepte von Kausalität (und in dieser Hinsicht in begrifflicher Nähe zum Schuldigsein) aufgefasst.

4.1.3. III) Inwiefern wird Verantwortung im Sinne eines relationalen Konzepts verstanden? (allgemein bzw. in Bezug auf Roboter)

Subjekt der Verantwortung: Der Autor kündigt zwar zu Beginn des Textes das Vorhaben an, eine Subjekt-Philosophie gegen eine Hermeneutik des Selbst zu ersetzen (S.1-2), dennoch spricht er mehrfach von einem handelnden Subjekt (z. B. S.2, 8). Irrgang charakterisiert Menschen im Sinne einer „*first person perspective*“ (2006, S.1) und macht beispielsweise an folgender Stelle klar, dass er auch im Kontext von moralischer Verantwortung die Notwendigkeit eines Subjekts sieht: „*In order to be morally responsible, however, an act needs a participant, who is characterized by personality or subjectivity.*“ (2006, S.12). Dass er zudem Autonomie (als ein möglicherweise mit Verantwortung in Verbindung stehendes Vermögen [siehe: 4.1.3. IV, b]) eng an die Frage des Vorhandenseins eines Subjekts knüpft, zeigt außerdem diese Stelle:

„*Robots that can act ,autonomously' -- or more precisely said, that can behave in many situations somewhat similar to animals -- continue to be technical products. They remain tools, and do not become doer (acting) subjects. How should they become subjects? (How could one attribute subject to the machines? Only by a miracle!)*“ (Irrgang, 2006, S.8).

Um Verantwortung übernehmen zu können, scheint der Autor (meist implizit) ausschließlich ein menschliches Subjekt (als potenzielle/r VerantwortungsträgerIn) vorauszusetzen/ in Erwägung zu ziehen, womit Robotern bei Irrgang – allgemein sowie in Bezug auf Verantwortung – wohl kein Platz zur direkten Berücksichtigung eingeräumt werden kann (siehe: Blockzitat oben).

Objekt der Verantwortung: Der Autor macht das Bauen kontrollierbarer Roboter als jene Handlung aus, für die (bzw. für deren Handlungsfolgen) Verantwortung von Menschen übernommen werden muss (S.10). Unkontrollierbare Roboter zu bauen (Anm.: entspricht der Handlung), könnte schweren Schaden (Anm.: entspricht den Handlungsfolgen) anrichten und gilt daher aus einer ethischen Sicht für Irrgang als Grenze der Verantwortung: „*a technology that is completely out of control and can cause heavy damage, should not be built, because this is the limit of a potential of responsibility*“ (2006, S.10). Dass das Subjekt (Mensch) und das eben in der Handlung und Handlungsfolge beschriebene Objekt (das Bauen unkontrollierbarer Roboter und damit in Verbindung stehende Schäden) in enger Beziehung zueinander stehen, wird an folgendem Rollenverständnis deutlich, wenn Irrgang meint, Roboter „*belong to the area of responsibility of humans*“ (2006, S.10), bzw. an dieser Stelle: „*In order to be able to take responsibility, technical systems must be controllable by humans.*“ (2006, S.6).

Inстанz der Verantwortung: Im Text wird nicht darauf eingegangen, vor welcher Instanz ein Verantwortungssubjekt (Anm.: bei Irrgang kommt als solches nur der Mensch in Frage; siehe: oben) Verantwortung zu übernehmen hat. Zwar kann der ethische Appell, keine unkontrollierbaren Technologien zu bauen, da diese schweren Schaden anrichten könnten (siehe: oben [Objekt der Verantwortung]), im Sinne eines guten Gewissens bzw. als eine Art interne, private Instanz aufgefasst werden, konkrete bzw. weitere Anhaltspunkte finden sich dazu jedoch nicht im Text.

AdressatIn der Verantwortung: Es gibt keine Stelle im Text, die auf direkte Weise ein Gegenüber der Verantwortung ausmacht. Wenn davon die Rede ist, dass unkontrollierbare technische Systeme, die schweren Schaden verursachen könnten, nicht (von Menschen) gebaut werden sollten (siehe: oben), dann kann hier indirekt – über die Beantwortung der Frage, wen (welche Entitäten, Werte, Normen etc.) es vor diesem vermeintlich schweren Schaden zu schützen gilt – auf ein Gegenüber des Verantwortungssubjekts geschlossen werden. Es kämen in diesem Kontext u.a. Menschen, Gesundheit oder andere in Betracht (Anm.: Roboter sind nicht unbedingt ausgeschlossen). Auf Grund der anthropozentrischen Position Irrgangs (siehe u.a.: 4.1.2. III, b) mag man vielleicht dazu geneigt sein, davon auszugehen, dass der Verfasser am ehesten Menschen (bzw. damit direkt in Zusammenhang stehende Normen und Werte) bzw. eher nicht Roboter (bzw. damit in Verbindung stehende Normen und Werte) als konkretes Gegenüber des Verantwortungssubjekts ausmachen würde. Angesichts mangelnder Informationen in diesem Punkt, wäre ein solcher Schluss jedoch zu hypothetisch.

Normative Kriterien der Verantwortung: Bereits in der Auseinandersetzung mit Irrgangs Auffassung von Verantwortung im Sinne eines Pflichtbegriffs (siehe: 4.1.2. II, b) wurde die Norm bzw. Regel der menschlichen Kontrollierbarkeit erwähnt. Diese gilt es – dem Autor nach – zu befolgen, wenn es darum geht, als Mensch moralisch verantwortlich sein zu können (vgl. „*In order to be able to take responsibility, technical systems must be controllable by humans.*“ [Irrgang, 2006, S.6]). Wenn Irrgang betont, dass eine Handlung – um als moralisch verantwortlich gelten zu können – eines Handlungsteilnehmers/ einer Handlungsteilnehmerin bedarf, der/ die sich durch Personalität oder Subjektivität auszeichnet, dann spricht er somit nicht nur Bedingungen für Handlungs-/ Verantwortungsfähigkeit an (siehe: 4.1.3. IV), sondern erweckt damit auch zumindest den Eindruck normativer Kriterien für Verantwortung (Anm.: die Norm personaler/ subjektiver [für Irrgang: menschlicher] Handlungssubjekte). Da Irrgang Subjektivität als etwas Menschliches ausmacht (siehe: 4.1.2. III, b) und Roboter ausdrücklich nicht als Subjekte ansieht (siehe: oben), scheint klar, dass letztere einer solchen Norm nicht genügen können.

Zusammenfassung 4.1.3. III

Es lassen sich aus dem Text in Zusammenhang mit Verantwortung mehrere Relata ermitteln und charakterisieren, manche davon konkret, andere weniger deutlich bzw. mit weniger Anhaltspunkten: Um normativ verantwortlich sein zu können, bedarf es nach Irrgang eines Subjekts, worunter für den Autor wohl nur Menschen, nicht aber Roboter fallen. Als das Objekt der Verantwortung konnten das Bauen unkontrollierbarer Roboter sowie sich dadurch (potenziell) ergebende Schäden ausgemacht werden. Eine Instanz der Verantwortung wird nicht direkt angesprochen. Als AdressatIn der Verantwortung könnte mutmaßlich der Mensch (bzw. damit in Verbindung stehende Normen, Werte etc.) gelten, wohl eher nicht (direkt) Roboter. Menschliche Kontrollierbarkeit beim Bauen technischer Systeme scheint als ein normatives Kriterium von Verantwortung eine gewisse Bedeutung für Irrgang darzustellen.

4.1.3. IV) Inwiefern spielen folgende Aspekte als Bedingungen der Zuschreibung von Verantwortung eine Rolle, wenn es um die Charakterisierung von moral agents geht? (allgemein bzw. in Bezug auf Roboter)

4.1.3. IV, a) Kommunikationsfähigkeit

Es wurde bereits darauf eingegangen, warum die Vermutung naheliegt, dass Irrgang normative Verantwortung als etwas Intersubjektives bzw. Dialogisches begreift (siehe: 4.1.3. I). Inwiefern der Autor Kommunikationsfähigkeit als wesentliche Verantwortungsbedingung ausmacht, lässt sich jedoch nicht eindeutig bzw. nur schwierig aus dem Text erschließen, da er diese nicht direkt in diesem Kontext anspricht. An einer Stelle bezieht sich Irrgang (unabhängig von Verantwortung) auf die Kompetenz der Kommunikation: „*It is objected that the communicative competence of humans is not in principle technically replaceable. [...] But it ignores the role of humans, who can not be simulated.*“ (2006, S.6). Da davon auszugehen ist, dass der Autor Menschen im Gegensatz zu Robotern als normativ verantwortungsfähige Wesen auffasst (siehe: 4.1.2. III, b) und hier die technische Unersetzbarkeit eines menschlichen Vermögens – das im Kontext des Dialogischen steht – angesprochen wird (vgl. Zitat), kann zumindest vermutet werden, dass Irrgang Kommunikationsfähigkeit in Zusammenhang mit Verantwortung eine gewisse Relevanz beizumisst. Aus der Tatsache, dass der Autor von menschlicher Kommunikation spricht (vgl. Zitat), ist abzuleiten, dass er diese Kompetenz wohl im sprachlichen Sinne auffasst.

4.1.3. IV, b) Handlungsfähigkeit

Inwiefern (moralische) Verantwortung und (moralische) Handlungsfähigkeit (konzeptuell bzw. kriteriologisch) zusammenhängen, wurde bereits in 4.1.2. III, d näher beleuchtet. Daraus, dass Irrgang die generelle Fähigkeit, handeln zu können, als Bedingung für Verantwortungsfähigkeit überhaupt ansieht, kann geschlossen werden, dass mehrfach dann, wenn von Verantwortung die Rede ist, auch gewisse Handlungen mitangesprochen sind (z. B. S.6, 11, 12), beispielsweise an folgender Stelle: „*In order to be morally responsible, however, an act needs [...]*“ (2006, S.12). Im Gegensatz zu Menschen besitzen Roboter dem Autor zufolge Handlungsfähigkeit nur im Sinne abstrakter Handlungen bzw. Handlungsmuster (siehe: 4.1.2 I, b).

Im Folgenden soll kurz geklärt werden, inwiefern Irrgang Handlungsfähigkeit in Zusammenhang mit – in diesem Kontext als bedeutsam erachteten – nachstehenden Aspekten/ Kompetenzen bringt:

Folgebewusstsein: Das Bewusstsein für/ von Handlungsfolgen wird nur indirekt in Zusammenhang mit Verantwortung angesprochen: Wenn Irrgang das Bauen unkontrollierbarer Roboter und damit in Verbindung stehende Schäden als „*the limit of a potential of responsibility*“ (2006, S.10) bezeichnet, dann weist dies darauf hin, dass er ein gewisses vorausblickendes Wissen (des Menschen) um schädigende Folgen bestimmter Handlungen annimmt bzw. als wichtig erachtet (siehe: 4.1.3. III). Irrgang geht eindeutig davon aus, dass Roboter – im Gegensatz zu Menschen – kein Bewusstsein für eine Handlung (siehe: 4.1.3. I), ergo wohl auch kein Folgebewusstsein haben können.

Kontextwahrnehmung: Die Bedeutung einer gewissen Wahrnehmung in Bezug auf den Kontext kann daraus erschlossen werden, dass Irrgang betont, dass die Wirkung von Computern (für Menschen) nur im Kontext sozialer Gruppen verstehbar sei (S.5; siehe: 4.1.2. II, a). Es scheint daher die Vermutung naheliegend zu sein, dass der Autor ebenso auch die Folgen einer (menschlichen) Handlung in Zusammenhang mit Robotern nur dann als ermittelbar ausmachen würde, wenn die jeweilige Situation – in einem relationalen Sinne – vom Menschen (als Handlungssubjekt; siehe: 4.1.3. III) wahrgenommen wird.

Personalität: Irrgang setzt (moralische) Verantwortung ausdrücklich in Beziehung mit Personalität: „*In order to be morally responsible, however, an act needs a participant, who is characterized by personality or subjectivity.*“ (2006, S.12).

Einflussmöglichkeit: Irrgang geht im Zusammenhang mit artifiziellen Entitäten davon aus, dass „an action pattern (action-scheme) is present, if at least the intentionality and the goal of action are preliminarily modelled.“ (2006, S.2; siehe auch: 4.1.2. I, c). Für Handlungsmuster eines Roboters scheint es also zu genügen, wenn eine gewisse Absichtlichkeit vorab modelliert wurde. Damit jedoch verantwortliches Handeln möglich ist, sieht Irrgang die Notwendigkeit (vollständiger?) menschlicher Kontrolle über technische Systeme als unerlässlich (vgl. „In order to be able to take responsibility, technical systems must be controllable by humans.“ [2006, S.6]) und fasst in diesem Kontext absichtsvolles Einflussnehmen in einem relativ strengen Sinne auf.

Freiheit: Es wurde bereits mehrfach betont, dass Irrgang Menschen im Vergleich zu Robotern als handlungsfähige Entitäten versteht (siehe: oben). Der Maßstab Mensch scheint artifiziellen Systemen in vielerlei Hinsicht unerreichbar zu sein (siehe: 4.1.2. I, b), ausdrücklich auch im Kontext menschlicher Freiheit, wenn Irrgang im Zusammenhang mit Maschinen schreibt: „human liberty or freedom can't be simulated.“ (2006, S.11). Der Autor bringt Freiheit – in der Bedeutung von Autonomie – mehrfach in Verbindung mit Handlungsfähigkeit (z. B. S.3, 8, 9), was darauf hindeutet, dass diese hierfür von gewisser Bedeutung sein könnte. Robotern spricht Irrgang Autonomie höchstens in einem abstrakten Sinne zu (Anm.: Dies zeigt sich u.a. auch darin, dass Irrgang *Autonomie* in Verbindung mit Robotern z. T. unter Anführungszeichen setzt [siehe z. B. S.8]). Inwiefern der Autor Freiheit unter dem Aspekt kognitiver Kompetenzen (wie Rationalität oder Urteilskraft) als wichtig erachtet, lässt sich nicht eindeutig aus dem Text erschließen. In jedem Fall macht Irrgang deutlich, dass artifizielle Systeme nur prärationale Entitäten darstellen (S.5) bzw. dass kognitive Kompetenzen des Menschen nicht simuliert werden können (S.6).

4.1.3. IV, c) Urteilskraft

Wie bereits thematisiert (siehe: 4.1.3. II, b; 4.1.3. III), spricht Irrgang in Verbindung mit Verantwortung indirekt von einer gewissen Norm (Anm.: menschliche Kontrollierbarkeit bei der Entwicklung von Technik). Die Frage, welche Vermögen der Autor zur Ausbildung solch einer Norm (bzw. eines Normensystems allgemein) für notwendig erachtet, lässt sich nicht eindeutig aus dem Text erschließen. Da es nach Irrgang im Kontext von Verantwortungsübernahme jedoch expressis verbis um menschliche Kontrollierbarkeit geht (vgl. „In order to be able to take responsibility, technical systems must be controllable by humans.“ [2006, S.6]), könnten im Sinne einer Urteilskraft menschliche Fähigkeiten in Betracht gezogen werden. Dass der Verfasser kognitive Kompetenzen des Menschen als elaborierter (im Vergleich zu artifizieller Prärationaleität [S.5]) ansieht,

wurde bereits im letzten Abschnitt erwähnt (siehe: 4.1.3. IV, b [Freiheit]). Möglicherweise würde der Autor auch der Notwendigkeit bestimmter Vermögen (wie etwa Reflexionsfähigkeit oder Rationalität) zur Ausbildung handlungsleitender Gründe zustimmen. Darüber kann jedoch in Anbetracht mangelnder Informationen an dieser Stelle kein sicheres Urteil gefällt werden.

Zusammenfassung 4.1.3. IV

Inwiefern Irrgang gewisse Aspekte als Bedingungen der Zuschreibung von Verantwortung ausmacht, ist nicht immer leicht bzw. eindeutig festzustellen (v.a. da dieser häufig nicht [explizit] darauf eingeht): Klar scheint, dass für den Autor (gewisse Aspekte von) Handlungsfähigkeit eine Prämisse für (moralische) Verantwortungsfähigkeit darstellt (hier v.a. Personalität). Inwiefern Kommunikation und Urteilkraft Bedingungen für Verantwortungszuschreibung darstellen, lässt sich nicht klar aus dem Text erschließen: Es wäre jedoch wenig überraschend, wenn Irrgang in diesen Punkten menschliche Kommunikationsfähigkeit und kognitive Kompetenzen des Menschen (im Vergleich zu Vermögen artifizieller Systeme) präferieren würde.

4.2. Analyse von: *When Is a Robot a Moral Agent* (Sullins, 2006)

4.2.1. Kontextualisierung

4.2.1. I) In welchem (speziellen) Arbeitsfeld bzw. mit welchen Schwerpunkten beschäftigt bzw. beschäftigen sich der/ die AutorIn bzw. AutorInnen des Textes?

(Anm.: Die folgende Beschreibung bezieht sich auf die Autorenbiographie, die von John Sullins in *Academia* verfasst wurde [Sullins, 2019].)

Der Autor des zu analysierenden Textes – Dr. John P. Sullins – ist Professor für Philosophie an der Sonoma State University in Kalifornien (USA). Seine Spezialgebiete liegen in der Technikphilosophie, Computerethik, Cognitive Science und Philosophy of Science. Sullins' Forschungsschwerpunkte umfassen u.a. die Bereiche der AI, Robotik, Philosophy of Mind, Computerethik etc. Der Autor hat speziell in den Feldern von AI und Robotik über ethisch-praktische Probleme nachgedacht und publiziert (beispielsweise über autonome Kriegssysteme, selbstfahrende Autos

etc.). Seine Arbeit erfährt immer wieder Berührungspunkte mit den Gebieten der Computer- und IT-Ethik sowie der Frage nach dem Design autonomer ethischer Akteure. (Sullins, 2019)

4.2.1. II) Welche Aktualität und Relevanz hat der Text im roboterethischen Diskurs?

Der Text in der vorliegenden Version erschien im Jahr 2006 in *International Review of Information Ethics* (IRIE) (Sullins, 2006). Er ist außerdem u.a auch im Sammelband *Machine Ethics* (Anderson & Anderson, 2011) zu finden.

Innerhalb der roboterethischen (bzw. nahestehenden) Debatte(n) wird Sullins' Text von unterschiedlichen – z. T. auch sehr aktuellen – AutorInnen bzw. im Rahmen angewandter ethischer Arbeiten/ Werke zitiert bzw. darauf verwiesen, was auf eine gewisse (praktisch-ethische) Bedeutung/ Relevanz hindeutet, beispielsweise bei:

Anderson, M., Anderson S., L. (2011). *Machine Ethics*. New York: Cambridge University Press.

Gunkel, D. J. (2012). *The Machine Question. Critical Perspectives on AI, Robots, and Ethics*. Cambridge/ London: The MIT Press.

Scalet, S., Arthur, J. (2016). *Morality and Moral Controversies. Readings in Moral, Social and Political Philosophy*. New York: Routledge.

Van Wynsberghe, A. (2015). *Healthcare Robots. Ethics, Design and Implementation*. London/ New York: Routledge.

4.2.2. Argumentative Analyse

4.2.2. I) Welche Auffassung liegt bezüglich *Moral* bzw. (*moral*) *agency* vor?

4.2.2. I, a) Allgemeine Charakterisierung

Die Begriffe *Moral*, *agency* und *moral agency* spielen durchgehend bzw. eine entscheidende Rolle in Sullins' Text, was sich bereits in der Häufigkeit ihrer expliziten Verwendung (das engl. Wort *moral* wird über 100 Mal erwähnt, *moral agent* 42 Mal, *moral agency* 20 Mal) bzw. im Titel des Papers (*When is a Robot a Moral Agent?*) andeutet. Wenn es um die Frage moralischer Akteurschaft geht, werden häufig auch Menschen bzw. menschliche Vermögen angesprochen (z. B.

S. 23, 25, 28, 29 etc.), dennoch besteht das primäre Ziel Sullins' darin, zu klären, inwiefern bzw. unter welchen Umständen Roboter als *moral agents* gelten können, wie der Autor einleitend festhält: „*In this paper I argue that in certain circumstances robots can be seen as real moral agents.*“ (2006, S.23).

Sullins verwendet den Begriff der *Moral* bzw. *Moralität* (engl. *moral* bzw. *morality*) – abgesehen von moralischer Akteurschaft – in einer Bandbreite unterschiedlichster Zusammenhänge und spricht etwa von: Moralität gewisser Situationen, Situationen mit moralischem Charakter, moralischen Rechten und Pflichten, moralischen Interessen, moralischen Werten, dem moralisch Guten/ Lobenswerten (alle: S.24), moralischen Beziehungen, moralischer Verantwortung (alle: S.25), moralischer Schuld, moralischer Urteilskraft, moralischen Entscheidungen (alle: S.26), Moraltheorie, mind-less-Moralität, moralischen Entscheidungen, dem moralischen Status (alle: S.27), dem moralisch Nützlichen/ Schädlichen (S.28), moralischen Fragen und Rätseln, dem moralisch Gleichen (alle: S.29); Obwohl der Autor in verschiedensten Kontexten von *Moral* spricht, gibt er jedoch nicht an, was er konkret darunter versteht und deutet nur an einer Stelle darauf hin, dass *Moral* mit Schaden/ Leid bzw. dem Guten/ Wohl (von Verhalten/ Handlungen) zu tun hat (S.28; siehe: unten). Sullins gibt im Gegensatz dazu jedoch sehr deutlich zu verstehen, dass er sich (im Rahmen von Robotern) für eine mind-less Moralität ausspricht (S.27), weshalb anzunehmen ist, dass er *Moral* nicht unbedingt in einem streng philosophischen Sinne versteht.

Die prinzipielle Möglichkeit zu handeln lässt sich für den Autor nicht allein auf den Bereich des Menschlichen beschränken, sondern ist – bei einer gewissen Toleranz niedriger Abstraktionsniveaus (siehe: 4.2.2. I, c) – auch bei anderen Entitäten denkbar: Sullins fasst den Begriff *agency* sehr weit als „*self motivated, goal driven [sic] behavior*“ (2006, S.28) auf. Außerdem knüpft sein Konzept von (moral) *agency* eng an einen (ebenfalls sehr breiten) Begriff von Autonomie (S.28): Sobald Autonomie auf einem sehr basalen Niveau vorhanden ist (Anm.: d.h. für Sullins, eine Entität steht nicht unter direkter Kontrolle eines anderen Akteurs/ einer anderen Akteurin [S.28]), kann bereits die Rede von praktisch unabhängiger Akteurschaft sein (S.28). Wenn eine derart autonome Handlung zudem wirkungsvoll im Erreichen/ Erfüllen gewisser Ziele/ Aufgaben ist, dann bezeichnet Sullins eine solche als effektiv autonom (Anm.: Autonomie bei Sullins, siehe u.a.: 4.2.3. IV, b), wobei einer Entität – je nach Effektivität der jeweiligen Handlung – folglich mehr oder weniger Akteurfähigkeit zugeschrieben werden kann (S.28). Wenn eine Entität mittels einer effektiv autonomen Handlung außerdem Schaden/ Leid oder Gutes/ Wohl in einem moralischen Sinne verursacht, dann kann diese – dem Autor nach – als *moral agent* aufgefasst werden

(S.28). (Anm.: Dieses stark kausale [verursachende] Verständnis von moral agency rückt den Begriff der Verantwortung letztlich in eine gewisse Nähe zum Schuldbegriff [siehe: 4.2.3. II, c].)

Sullins' Auffassung von moralischer Akteurschaft konkretisiert sich darüber hinaus folgendermaßen: Damit eine Entität als moral agent gelten kann, muss dem Autor zufolge nicht notwendig Personalität vorliegen (S.23, 26), sondern zumindest ein gewisser Grad an autonomem Vermögen (S.26-28; siehe: oben) sowie drei Voraussetzungen: signifikante Autonomie, Intentionalität und Verantwortung (siehe u.a.: 4.2.3.; v.a.: 4.2.3. IV, b).

4.2.2. I, b) Inklusivität/ Exklusivität

Sullins' Perspektive auf Moral bzw. (moralische) Handlungsfähigkeit kann als inklusiv, nicht-anthropozentrisch und als mathenozentrische bezeichnet werden, worauf im Folgenden näher eingegangen werden soll:

Der Autor meint, dass typischerweise die Auffassung bestehe, moralische Akteure/ Akteurinnen müssten Personen sein (S.26). Entitäten, die nicht als Personen gelten, fallen demnach automatisch auch aus dem Bereich der moral agency heraus. Demgegenüber versucht Sullins zu argumentieren „*that personhood is not required for moral agency.*“ (2006, S.26). Eine solche Perspektive kann – entgegen traditioneller Positionen – den Blick für die Diskussion/ Betrachtung moralischer Akteure/ Akteurinnen erweitern und insofern als inklusiv bezeichnet werden.

Sullins' Perspektive kann außerdem auch dahingehend als inklusiv bezeichnet werden, da der Autor neben Personalität weitere, häufig als notwendig erachtete anthropozentrische Prämissen – die traditionell in Zusammenhang mit Moral bzw. (moralischer) Handlungsfähigkeit stehen – auf ein niedrigeres Abstraktionsniveau herunterbricht (siehe: 4.2.2. I, c): So vertritt der Verfasser des Textes das Konzept einer mind-less Moralität (S.27) und betont, dass speziell menschliche Vermögen, wie Bewusstsein, Seele, Geist etc. nicht notwendig sind, um von moral agency (bzw. damit in Zusammenhang stehenden Kriterien) sprechen zu können (S.29).

Als mathenozentrisch kann Sullins' Perspektive auf moral agency (bzw. Moral) bezeichnet werden, da dieser ausdrücklich darum bemüht ist, die Frage zu klären, inwiefern Roboter (als künstliche Entitäten) in den Horizont moralischer Akteurschaft miteinbezogen werden können (vgl. Titel [*When Is a Robot a Moral Agent?*]). Der Autor akzeptiert, dass Menschen in der Regel als moralische Akteure/ Akteurinnen gelten, geht darauf jedoch höchstens beiläufig bzw. indirekt ein (z. B. „*Certainly a human nurse is a moral agent.*“ [2006, S.29]), während sein primäres Anliegen

wohl darin besteht, jene Umstände zu klären, unter denen Roboter als moralische Akteure angesehen werden können (S.23).

4.2.2. I, c) *Gradualität/ Abstraktionslevel/ alternative Konzepte*

Der Art und Weise, wie Sullins (moralische) Handlungsfähigkeit auffasst (siehe: 4.2.2. I, a-b) liegt ein graduelles Konzept von Moral bzw. (moral) agency zu Grunde, das in Zusammenhang mit einem herabgesetzten Abstraktionslevel bzw. einem alternativen (d.h. einem nicht im traditionellen Sinne exklusiv auf den Menschen bezogenen) Moral- bzw. Handlungsverständnis zu stehen scheint:

Es wurde bereits darauf eingegangen, dass Sullins (wenn auch nicht immer explizit) wohl davon ausgeht, dass Menschen als moral agents gelten (siehe: 4.2.2. I, b). Der Autor bleibt dabei aber nicht stehen, sondern thematisiert, dass – um als moralisch handlungsfähige Entität gelten zu können – es u.a. notwendig sei, zumindest einen „*significant degree of autonomous ability to reason and act on those reasons.*“ (S.26, 2006) ausmachen zu können, bzw. „*that full subjectivity is not needed*“ (S.27, 2006). Diese beiden Stellen deuten auf ein graduelles Verständnis von moral agency hin.

Auch wenn sich Sullins mit Kriterien für moral agency beschäftigt, tut er dies augenscheinlich in einem abgestuften Sinne, da er in diesem Kontext neben streng philosophischen auch einfachere Formen von Autonomie, Intentionalität und Verantwortung toleriert (siehe: unten). Im Zuge der Auseinandersetzung Sullins' mit komplexen Maschinen meint dieser außerdem, dass deren ProgrammiererInnen nicht als einziger Lokus moralischer Akteurschaft aufgefasst werden können (S.26), was als weiterer Hinweis für ein graduelles Verständnis für moral agency gelten kann. Die eben gebrachten Argumente legen nahe, dass der Autor in diesem Kontext keine binäre Sichtweise (Anm.: moral agent – nicht moral agent) einnimmt, sondern vielmehr ein Kontinuum zwischen voller moral agency und basalen Formen vertritt, wobei – je nach gradueller Bestimmung der hierfür als notwendig erachteten Kompetenzen (siehe u.a.: 4.2.3. IV) – von sehr vielen bzw. feinen Abstufungen ausgegangen werden kann.

Sullins spricht an zwei Stellen von einem „*reasonable level of abstraction*“ (2006, S.28/ 29), von dem ausgehend zu beurteilen ist, ob es sich bei einer gewissen Entität um einen moral agent handelt. Der Autor erklärt indirekt, was er unter diesem angemessenen Abstraktionsniveau versteht,

indem er angibt, wie er dieses im Verhältnis zu traditionellen Sichtweisen von moralischer Akteurschaft versteht: So spricht Sullins davon, dass er Personalität (S.23, 26) und Subjektivität (S.27) sowie allgemein Begriffe/ Konzepte, die mit der Geistesphilosophie in Zusammenhang stehen (S.27), nicht als nötig erachtet, um über moralische Handlungsfähigkeit zu diskutieren. Ebenso fasst er auch jene Kriterien, die er als notwendig für das Vorhandensein von moral agency erachtet (Anm.: Autonomie, Intentionalität, Verantwortung; siehe u.a.: 4.2.2. I, a; III, c) in einer wenig abstrakten, d.h. nicht streng philosophischen Form auf: Autonomie in einem technischen bzw. nicht radikalen Sinne (S.28), Intentionalität sowie Verantwortung in einem abgeschwächten bzw. nicht streng philosophischen Sinne (S.28, 29).

Sullins' Konzept von moral agency kann als alternativ bezeichnet werden, da der Autor – entgegen traditioneller Auffassungen zu moralischer Akteurschaft – ausdrücklich darum bemüht ist, geistesphilosophische Fragen nicht (bzw. höchstens auf einem sehr niedrigen Abstraktionslevel) zu berücksichtigen (S.27-29). Die Umsetzung dieses Vorhabens wird v.a. bei der näheren Beschäftigung des Verfassers mit den Schlüsselkriterien für moral agency deutlich, die dort vorwiegend im Sinne einer funktionalen Äquivalenz verstanden werden (S.28-29; siehe: 4.2.2. III, c [Verantwortung]; 4.2.3. IV, b [Autonomie und Intentionalität]). Insofern kann Sullins' Konzept von moralischer Akteurschaft als ein funktionales bezeichnet werden (siehe: 4.2.2. II, b; III, d).

Zusammenfassung 4.2.2. I

Sullins befasst sich explizit mit der moral agency-Thematik und dabei v.a. mit der Frage, inwiefern Roboter als moralische Akteure gelten können. Während der Autor nicht konkret bestimmt, was er allgemein unter Moral versteht, charakterisiert er (in engem Zusammenhang mit Autonomie) ein sehr breites Konzept von (moral) agency. Moralische Akteurschaft konkretisiert der Verfasser außerdem anhand dreier Kriterien (signifikante Autonomie, Intentionalität, Verantwortung), die er hierfür als notwendig erachtet.

Sullins' Perspektive auf moralische Akteurschaft kann als inklusiv, nicht-anthropozentrisch und als mathenozentrisch bezeichnet werden, da der Fokus v.a. darauf gelegt wird, inwiefern Roboter in den Horizont moralischer Akteurschaft miteinbezogen werden können und hierbei traditionelle Voraussetzungen (v.a. aus dem Bereich der Geistesphilosophie) keinen Prämissenstatus einnehmen.

Es ist außerdem deutlich erkennbar, dass der Autor ein Kontinuum zwischen voller moral agency und basalen Formen vertritt, das in engem Zusammenhang mit der Toleranz eines niedrigen

Abstraktionsniveaus zu stehen scheint. Moralische Akteurschaft bzw. deren Kriterien werden dennoch in einer – in Bezug auf geistesphilosophische Vermögen – äquivalenten Weise bzw. unter einer Perspektive funktionaler Äquivalenz betrachtet.

4.2.2. II) Wie werden Roboter charakterisiert?

4.2.2. II, a) Allgemeine Charakterisierung

Mehrere Stellen in Sullins' Text weisen darauf hin, dass dieser eine relationale Sichtweise in Bezug auf Roboter einnimmt. So beschäftigt sich der Autor gleich zu Beginn mit Moralität und Mensch-Roboter-Interaktionen und meint im Rahmen dieses Abschnitts:

„At the present time, these interactions [Anm.: interactions between robots and humans] are no different than the interactions one might have with any piece of technology, but as these machines become more interactive they will become involved in situations that have a moral character that may be uncomfortably similar to the interactions we have with other sentient animals.“ (Sullins, 2006, S.24).

Situationen, in denen Roboter mit Menschen in Beziehung stehen, bezeichnet Sullins als *„real moral situations that must be taken seriously“* (2006, S.24). Der Autor kritisiert instrumentalisierende Standardmodelle, in denen Roboter als bloße Werkzeuge betrachtet werden und in denen seiner Meinung nach ein zu einfacher Blick auf die Art und Weise, wie Technologien (schon jetzt) ethische bzw. moralische Werte des Menschen beeinflussen, gelegt wird (S.24). Stattdessen schlägt Sullins – in Anlehnung an ein Verständnis von Technologie als *„manipulation of nature to human ends“* (2006, S.24) ein Modell vor, das komplexe Beziehungen zwischen natürlichen (Anm.: Menschen) und artifiziellen AkteurInnen (Anm.: u.a. Robotern) berücksichtigt (S.25). Dabei unterscheidet er zwischen zwei unterschiedlichen Formen von Robotertechnologien (S.25-26): Während er Teleroboter als vorwiegend (von Menschen) kontrollierte Maschinen mit nur minimaler Autonomie bezeichnet, können autonome Roboter in einer völlig anderen Form aufgefasst werden, in der sie in einer Art Beziehungsnetz (u.a. bestehend aus: ProgrammiererInnen, anderen Robotern, BenutzerInnen, VermarkterInnen etc.) mit anderen Entitäten in (moralischer) Verbindung stehen (S.26).

Sullins schreibt v.a. autonomen Robotern gewisse Kompetenzen und Entwicklungspotenziale zu (siehe u.a.: 4.2.2. II, b; 4.2.3. IV) und erachtet es auch prinzipiell als möglich – wenn auch nicht als wahrscheinlich in näherer Zukunft – dass Roboter in gewissen Punkten (z. B. Autonomie,

Klarheit von Intentionen) – sogar Menschen überlegen sein könnten (S.29). Dennoch bewertet der Autor den aktuellen Status autonomer Robotertechnologien als bescheiden (S.27) und weist beispielsweise darauf hin, dass gewisse Kompetenzen noch weit von menschlichen bzw. tierischen Vermögen entfernt sind, wie etwa an folgenden Textstellen klargemacht werden soll: „*No robot in the real world or that of the near future is, or will be, as cognitively robust as a guide dog.*“ (Sullins, 2006, S.25); „*It is going to be some time before we meet mechanical entities that we recognize as moral equals*“ (Sullins, 2006, S.29). Auf die Frage, inwiefern Sullins Roboter als (moral) agents betrachtet, bezieht sich der nächste Abschnitt (siehe: 4.2.2. II, b).

4.2.2. II, b) Charakterisierung in Bezug auf (moral) agency

Der letzte Abschnitt mag zwar den Eindruck einer gewissen Bescheidenheit Sullins' in Bezug auf Vermögen/ Kompetenzen von Robotern gemacht haben, dennoch betrachtet der Autor Roboter unter gewissen Voraussetzungen als moral agents:

Sullins bezeichnet Teleroboter ausdrücklich als nicht-moralische Akteure (S.26), während autonome Roboter (siehe: 4.2.2. II, a) moral agents darstellen können: „*For the robot to be a moral agent, it is necessary that the machine have a significant degree of autonomous ability to reason and act on those reasons.*“ (2006, S.26). Autonomie bezeichnet der Autor in diesem Kontext als die Fähigkeit, zumindest einige wichtige Handlungsentscheidungen auf der Basis einer dem Roboter zu Grunde liegenden Programmierung ausführen zu können (S.26).

Handelt es sich um einen autonomen Roboter im oben beschriebenen Sinne, dann gilt es nach Sullins folgende drei Fragen zu klären: „*Is the robot significantly autonomous? Is the robot's behaviour intentional? Is the robot in a position of responsibility?*“ (2006, S.28). Werden diese allesamt positiv beantwortet, so kann ein Roboter dem Autor zu Folge als moralischer Akteur bezeichnet werden. Maschinen, die diesen Kriterien nicht entsprechen, bezeichnet Sullins ausdrücklich als nicht moralisch akteurfähig (S.29). Wichtig scheint hier die Bemerkung zu sein, dass der Verfasser die drei oben genannten Schlüsselkriterien in Anlehnung an ein Konzept von moral agency versteht, das (im Vergleich zu traditionellen Auffassungen bzw. geistesphilosophischen Vermögen) niedrigere Abstraktionsniveaus bzw. funktionale Äquivalenzen auf einem hohen Toleranzlevel zulässt (siehe: 4.2.2. I, c). Dadurch wird einerseits die Thematisierung gradueller Formen von moralischer Akteurschaft bei Robotern ermöglicht. Andererseits charakterisiert der Autor moral agency somit auch in einer funktionalen Form (siehe: 4.2.2. I, c; III, d),

Sullins betrachtet autonome Roboter – die oben genannte Kriterien erfüllen – nicht als alleinige moralische Akteure, jedoch zumindest in gewisser Weise als moral agents, wie diese Stelle verdeutlichen soll: „*I am not trying to argue that robots are the only locus of moral agency in such a community* [Anm.: a community of interaction between robots and other entities, like programmers, users etc.], *only that in certain situations they can be seen as fellow moral agents in that community.*“ (2006, S.26). Dass der Verfasser des Textes Menschen im Vergleich zu rezenten Robotern wohl ein höheres Maß an moralischer Akteurschaft zuspricht, wurde bereits im letzten Abschnitt angedeutet (4.2.2. II, a).

Zusammenfassung 4.2.2. II:

Bei der Charakterisierung von Robotern nimmt Sullins eine Perspektive ein, die als relational bezeichnet werden kann: Die moralische Betrachtung von Robotern bzw. deren moralische Handlungsfähigkeit scheint nur sinnvoll in Mensch-Roboter-Beziehungen thematisierbar zu sein. Roboter werden in diesem Kontext von Standardmodellen unzureichend berücksichtigt.

Sullins unterscheidet zwischen Telerobotern und autonomen Robotern, wobei er nur letztere als potenzielle moral agents bezeichnet, sofern sie den drei Schlüsselkriterien (Autonomie, Intentionalität und Verantwortung) – im Sinne eines angemessenen Abstraktionsniveaus bzw. funktionaler Äquivalenz (siehe: 4.2.2. I, c) – genügen. Ist dies der Fall, so kann autonomen Robotern zwar wohl keine moralische Akteurschaft im menschlichen Sinne zugesprochen werden, aber zumindest ein gewisser Teil an (funktionaler) moral agency innerhalb eines Beziehungsnetzwerks.

4.2.2. III) Welche Auffassung liegt bezüglich (moralischer) Verantwortung – im Zusammenhang mit Robotern – hinsichtlich folgender Aspekte vor:

4.2.2. III, a) Allgemeine Charakterisierung

Für Sullins gilt Verantwortung – neben Autonomie und Intentionalität – als eines von drei zentralen Kriterien für moral agency bei Robotern (S.28; siehe: 4.2.2. III, d). Obwohl er diese Prämissen separat behandelt, gibt es mehrere Hinweise dafür, dass auch Verantwortung im Sinne des Autors mit Autonomie und Intentionalität in Zusammenhang steht (siehe u.a.: 4.2.3. IV). Sullins versteht (moralische) Verantwortung aber nicht nur als ein Kriterium für moral agency, sondern auch als etwas, das moralischen Akteuren/ Akteurinnen gewissermaßen zukommt bzw. zusteht (siehe: 4.2.2. III, d).

Bei seiner Auseinandersetzung mit dem Begriff *Verantwortung* beschreibt Sullins dieses Kriterium zunächst sehr allgemein und unscharf (S.28): Verantwortung liege demnach dann vor, wenn sich ein Roboter in einer Art und Weise verhält, sodass gilt: „*we can only make sense of that behaviour by assuming it has a responsibility to some other moral agent(s).*“ Danach konkretisiert Sullins sein Verständnis für Verantwortung (Anm.: im Kontext von Pflegerobotern):

„*If the robot behaves in this way [Anm.: vgl. letztes Zitat] and it fulfils some social role that carries with it some assumed responsibilities, and only way we can make sense of its behaviour is to ascribe to it the ‘belief’ that it has the duty to care for its patients, then we can ascribe to this machine the status of a moral agent.*“ (2006, S.28)

Für Sullins scheint Verantwortung in Zusammenhang mit Robotern notwendigerweise an einen sozialen Kontext gebunden zu sein, wenn der Autor meint, dass erst das Ausführen/ Erfüllen einer gewissen sozialen Rolle Verantwortungen mit sich bringt (vgl. Zitat oben). Außerdem wird Verantwortung in Zusammenhang mit Überzeugungen (vgl. Zitat oben bzw. siehe: 4.2.2. III, c) und in Anlehnung an einen gewissen Zurechnungs- (vgl. Zitat oben „*ascribe to it*“ [Sullins, 2006, S.28]; siehe: 4.2.3. II, a) und Pflichtbegriff (vgl. Zitat oben „*that it has the duty to*“ [Sullins, 2006, S.28]; siehe: 4.2.3. II, b) betrachtet.

4.2.2. III, b) Inklusivität/ Exklusivität

Sullins’ Verständnis von (moralischer) Verantwortung kann schon insofern als inklusiv bezeichnet werden, als der Autor durch eine Auseinandersetzung mit Robotern (als potenziell verantwortliche Entitäten) traditionell aus dem Verantwortungsdiskurs ausgeschlossene Entitäten mit in die Debatte aufnimmt. Auf ein damit in Zusammenhang stehendes graduelles bzw. tolerantes (in Bezug auf Abstraktionslevel) Konzept von Verantwortung wird im nächsten Abschnitt (siehe: 4.2.2. III, c) näher eingegangen.

Sullins bemerkt zwar an einigen Stellen, dass Menschen (und auch Tiere) rezenten Robotern in mehreren Aspekten überlegen sind (z. B. hinsichtlich kognitiver Vermögen [S.25], bezüglich moralischem Level [S.29] etc.; siehe: 4.2.2. II, a), jedoch nimmt er keine anthropozentrische bzw. traditionelle Perspektive ein: Bei der Beschreibung von Situationen, in denen autonome Roboter involviert sind, meint der Autor: „*The programmers of these machines are somewhat responsible but not entirely*“ (Sullins, 2006, S.26). Da Menschen (in diesem Fall ProgrammiererInnen) – zumindest in Fällen, in denen sie mit autonomen Robotern in einem Netz von Beziehungen stehen (siehe: 4.2.2. II, a) – nicht als vollständig verantwortlich bezeichnet werden können (vgl. Zitat

oben), ist es naheliegend (Anm.: auch wenn von Sullins an dieser Stelle nicht explizit gemacht), dass der Autor Robotern hier (zumindest einen geringen Anteil an) Verantwortung zugestehen würde.

4.2.2. III, c) *Gradualität, Abstraktionslevel, alternative Konzepte*

Das Zitat im letzten Kapitel (Anm.: „*The programmers of these machines are somewhat responsible but not entirely*“ [Sullins, 2006, S.26]) unterstreicht nicht nur eine inklusive Perspektive auf Verantwortung, sondern deutet auch darauf hin, dass der Verfasser des Textes ein graduelles Konzept in dieser Angelegenheit vertritt. Sullins' Ausführungen zum Beispiel von Pflegerobotern deuten auch auf ein Kontinuum zwischen unterschiedlichen Graden von Verantwortung(sfähigkeit) hin: Damit Pflegeroboter in ihrem Kontext als vollwertige moral agents (vgl. Zitat unten: „*full moral agent*“ [Sullins, 2006, S.29]) zählen können, müssen diese dem Autor zufolge – neben Autonomie und Intentionalität – auch eine Programmierung aufweisen, die komplex genug ist „*that it* [Anm.: the robotic caregiver] *understands its role in the responsibility of the health care system*“ (Sullins, 2006, S.29). Ist dies nicht der Fall, dann meint Sullins: „*Any machine with less capability than would not be a full moral agent*“ (2006, S.29). Ein geringerer Grad an Vermögen (z. B. Verantwortungsfähigkeit) geht hier mit einem geringeren Grad an moralischer Akteurschaft einher. Wie viele Zwischenstufen das Kontinuum zwischen basalen Formen und einem sehr hohen Grad an moral agency bzw. Verantwortung(sfähigkeit) dabei annehmen kann, wird von Sullins nicht konkret angesprochen. (Anm.: Die [weitere] Auseinandersetzung mit Sullins' Text deutet jedoch darauf hin, dass der Autor auch Bedingungen für die Zuschreibung von Verantwortung[sfähigkeit] z. T. graduell auffasst bzw. auf verschiedenen Abstraktionsniveaus toleriert, weshalb wohl von zahlreichen Zwischenstufen in einem solchen Verantwortungskonzept ausgegangen werden kann [siehe u.a.: 4.2.3. IV].).

Eine graduelle Perspektive hinsichtlich Verantwortung scheint mit einer gewissen Toleranz niedriger(er) Abstraktionsniveaus für dieses Kriterium zusammenzuhängen (Anm.: ähnlich zum moral agency-Konzept, siehe: 4.2.2. I, c): Sullins spricht nicht nur im Zuge von moral agency von einem angemessenen Abstraktionslevel, sondern auch in konkretem Bezug auf die Kriterien für moralische Akteurschaft, wenn er schreibt, dass „*this questions* [Anm.: considering autonomy, intentionality and responsibility of a robot] *have to be viewed from a reasonable level of abstraction*“ (2006, S.28). Was versteht der Autor im Rahmen von (moralischer) Verantwortung als angemessenes Abstraktionsniveau? In 4.2.2. III, a wurde bereits darauf eingegangen, dass

Sullins im Kontext von Verantwortung von der Notwendigkeit gewisser Überzeugungen (engl. *beliefs*) spricht, die er folgendermaßen konkretisiert:

„Again, the beliefs do not have to be real beliefs, they can be merely apparent. The machine may have no claim to consciousness, for instance, or a soul, a mind, or any of the other somewhat philosophically dubious entities we ascribe to human specialness. These beliefs, or programs, just have to be motivational in solving moral questions and conundrums faced by the machine.“
(2006, S.29)

Die Ausklammerung geistesphilosophischer Prämissen für Überzeugungen (und damit indirekt auch für [moralische] Verantwortung) kann als eine Herabsetzung des Abstraktionsniveaus für (moralische) Verantwortung aufgefasst werden. (Anm.: Es sei an dieser Stelle kritisch angemerkt, dass Sullins einerseits dafür argumentiert, geistesphilosophische Vermögen nicht vorauszusetzen [bzw. höchstens auf einem niedrigen Abstraktionsniveau zu berücksichtigen], im nächsten Satz jedoch von einer Motivation beim Lösen moralischer Fragen spricht, was nicht komplett unabhängig von einer mind-less-Argumentation zu sein scheint [vgl. Blockzitat oben; siehe: 5.2.1, 5.3.2.].)

Sullins' Konzept von (moralischer) Verantwortung kann als alternativ bezeichnet werden, da es traditionelle Fragen rund um Bewusstsein, Personenstatus etc. ausklammert (vgl. Blockzitat oben). Es erscheint außerdem praktisch-pragmatisch ausgerichtet zu sein, was beispielweise daran deutlich wird, dass der Autor beschreibt, dass Überzeugungen von Programmen nicht in einer streng philosophischen Form vorliegen müssen, sondern *„just have to be motivational in solving moral questions“* (Sullins, 2006, S.29). Im Sinne einer solchen funktionalen Äquivalenz genügt es Sullins anscheinend, wenn im Rahmen von Überzeugungen zumindest ein gewisses Vermögen (eine Motivation, moralische Fragen zu lösen) vorliegt. Aus dieser Perspektive ist es naheliegend von einem funktionalen Verständnis von Verantwortung zu sprechen.

4.2.2. III, d) Zusammenhang mit moral agency

Die Auseinandersetzung mit Sullins' Perspektive hinsichtlich (moral) agency und (moralischer) Verantwortung (siehe: 4.2.2. I, II, III a-c) lässt einige Zusammenhänge zwischen diesen Begriffen/ Konzepten erkennen bzw. vermuten:

(Moralische) Verantwortung und moral agency werden von Sullins konzeptuell sehr ähnlich aufgefasst, nämlich vorwiegend in einer inklusiven, nicht-traditionellen (siehe: 4.2.2. I, b; III, b;)

bzw. in einer graduellen, auf einem herabgesetzten Abstraktionsniveau beruhenden, alternativen (funktionalen) Art (siehe 4.2.2. I, c; III, c).

Sullins fasst Verantwortung als eine (von drei) Prämisse(n) für moral agency auf: Die Klärung der Frage „*Is the robot in a position of responsibility?*“ (Sullins, 2006, S.28) stellt ein Schlüsselkriterium dar, wenn es darum geht, (moralische) Verantwortung autonomer Roboter zu klären. (Anm.: Der Autor betrachtet mit den beiden anderen Kriterien [Autonomie und Intentionalität] zwei weitere Prämissen für moral agency, die jedoch selbst wiederum – mehr oder weniger eng – mit dem Begriff der Verantwortung zusammenhängen [siehe: 4.2.3. IV].)

Sobald ein Roboter die drei oben genannten Kriterien in ausreichendem Maße erfüllt, kann dieser als moral agent bezeichnet werden. Sollte jedoch eine dieser Voraussetzungen nicht hinlänglich erfüllt sein, meint Sullins „*they* [Anm.: robots lacking a requirement] *are not robust moral agents for whom we can credibly demand moral rights and responsibilities equivalent to those claimed by capable human adults.*“ (2006, S.28). Der Autor scheint Verantwortung einerseits als Prämisse für moral agency zu verstehen, andererseits – betrachtet man die eben angeführte Stelle – aber auch als etwas, worauf ein/e moralische/r AkteurIn als solche/r gewissermaßen Anspruch hat.

Zusammenfassung 4.2.2. III

Verantwortung stellt bei Sullins eines von drei Schlüsselkriterien für moral agency dar und wird dabei auch in konzeptuell ähnlicher Weise charakterisiert. Darüber hinaus scheint der Autor Verantwortung auch als etwas aufzufassen, worauf moralische AkteurInnen als solche Anspruch haben.

Sullins' Konzept von (moralischer) Verantwortung kann – im Kontext autonomer Roboter – als relational, inklusiv/ nicht-traditionell, graduell, tolerant gegenüber niedrigen Abstraktionsniveaus und als alternativ bzw. funktional bezeichnet werden: Als ein relationales Konzept wurde Verantwortung ausgemacht, da für den Autor Verantwortung in Zusammenhang mit Robotern notwendigerweise an einen sozialen Kontext gebunden zu sein scheint. Inklusiv und nicht-traditionell ist es v.a. deshalb, da Sullins entgegen anthropozentrischer Denkmuster argumentiert. Das Konzept lässt sich als graduell auffassen, da sich im Text ein Kontinuum/ Zwischenformen zwischen nicht verantwortlich bzw. nicht verantwortungsfähig und einer starken Form von Verantwortung(sfähigkeit) erkennen lässt/ lassen. Diese Möglichkeit zur Abstufung scheint dabei eng an eine gewisse Toleranz für niedrige Abstraktionsniveaus (v.a. bezüglich geistesphilosophischer Aspekte,

wie Überzeugungen) geknüpft zu sein. Als funktional lässt sich Sullins' Konzept auffassen, da es mit Argumenten funktionaler Äquivalenz arbeitet.

4.2.3. Formale Analyse

4.2.3. I) Inwiefern spielt Verantwortung im Sinne einer etymologisch abgeleiteten Minimaldefinition eine Rolle? (allgemein bzw. in Bezug auf Roboter)

Sullins gibt im Rahmen seiner Auffassung von moral agency zu verstehen, „*that full subjectivity is not needed*“ (2006, S.27). Der Autor hält zudem Personalität für nicht notwendig im Zusammenhang mit moralischer Akteurschaft (siehe v.a.: 4.2.2. I, a; 4.2.3. IV, b). Auf Grund der konzeptuell ähnlichen Charakterisierung des Kriteriums der Verantwortung (siehe: 4.2.2. III, d) ist davon auszugehen, dass Sullins wohl auch Verantwortung als nicht notwendigerweise an Personalität bzw. eine starke Form von Subjektivität geknüpft sieht. Dennoch scheint sein Konzept von Verantwortung (im Rahmen von autonomen Robotern) nicht ohne eine Beziehung zu anderen Verantwortungssubjekten auszukommen, wie folgende Stelle verdeutlicht: „*Finally, we can ascribe moral agency to a robot when the robot behaves in such a way that we can only make sense of that behaviour by assuming it has a responsibility to some other moral agent(s).*“ (Sullins, 2006, S.28) oder jenes Zitat, in dem es darum geht, Verantwortung zwischen verantwortlichen Parteien aufzuteilen bzw. zu verorten: „*The programmers of these machines [Anm.: autonomous robots] are somewhat responsible but not entirely*“ (Sullins, 2006, S.26). In dieser (zumindest abgeschwächten) Form kann Verantwortung bei Sullins – auch weitgehend unabhängig von traditionellen bzw. individualistischen Prinzipien – als etwas Dialogisches bzw. Intersubjektives bezeichnet werden, wobei es keine Hinweise dafür gibt, dass der Autor dies im Sinne von Sprachlichkeit verstehen würde. Verantwortung(sfähigkeit) (bzw. Handlungsfähigkeit als ein Teilaspekt davon [siehe: 4.2.3. IV, b]) scheint für den Verfasser auch an einen gewissen Zweck gebunden zu sein: Wenn sich ein Roboter im oben beschriebenen Sinne verhält (siehe: vorletztes Zitat) und dieses Verhalten nur sinnvoll erklärbar ist, wenn man einer artifiziellen Entität die Überzeugung zuschreibt „*that it has the duty to care for its patients, then we can ascribe to this machine the status of a moral agent.*“ (Sullins, 2006, S.28).

Wenn Sullins in Zusammenhang mit Verantwortung eine gewisse (einem Gegenüber bestehende) Sorgfaltspflicht erwähnt (vgl. letztes Zitat; S.28), dann weist dies darauf hin, dass der Autor hierbei ein normatives Verständnis von Verantwortung vertritt. Der Verfasser spricht im Rahmen von Pflegerobotern an einer weiteren Stelle von einer Art normativem Kriterium, das deren Verantwortungsbereich definiert: So kann ein solcher Roboter nur dann als moral agent gelten, wenn dieser Autonomie und Intentionalität aufweist und wenn es zudem der Fall ist, dass dessen „*programming is complex enough that it understands its role in the responsibility of the health care system*“. Sullins versteht hier die Prämisse, ein Roboter müsse seine jeweilige Verantwortungsrolle im Gesundheitssystem verstehen, als eine Art normatives Kriterium (bzw. auch als Kompetenz für Verantwortungsfähigkeit, siehe: 4.2.3. IV, b). Da der Autor jedoch ein herabgesetztes Abstraktionsniveau bezüglich verschiedenster Kriterien für moral agency bzw. Verantwortung vertritt (siehe: 4.2.2. I, c; III, c), ist nicht klar, inwieweit prima facie normativ charakterisierte Verantwortung nicht auch in einer gewissen Nähe zu kausal-deskriptiver Verantwortung (oder eng verwandten Begriffen [siehe: 4.2.3. II]) verstanden wird.

Verantwortung hat bei Sullins mit gewissen Kompetenzen der angesprochenen Entität (meist Roboter) zu tun: Wie schon mehrfach erwähnt (siehe v.a.: 4.2.2. III, c), handelt es sich hierbei um Vermögen, die geistesphilosophische bzw. typischerweise Menschen zugeschriebene Aspekte ausklammern bzw. auf einem niedrigen Abstraktionslevel tolerieren. So bezeichnet der Autor etwa menschliches Bewusstsein, Geist, Seele etc. als nicht notwendig, damit von Überzeugungen ausgegangen werden kann, auf denen Verantwortung basiert (siehe: 4.2.2. III, a, c; S.28-29). Indem Sullins bemerkt, dass Überzeugungen lediglich motivational im Lösen moralischer Herausforderungen sein müssen (S.29) bzw. ein Verstehen der jeweiligen Verantwortungsrolle vorhanden sein müsse (S.29), weist der Verfasser auf eine bestimmte psychomotivationale Verfasstheit von Robotern hin, die einen gewissen Charakter der Bewusstheit zu tragen scheinen (vgl. motivational) bzw. auch eine gewisse Reflexions- und Urteilsfähigkeit voraussetzen scheinen (vgl. Verstehen; siehe: 4.2.3. IV, b-c). (Anm.: Es sei an dieser Stelle [ähnlich wie schon in 4.2.2. III, c] kritisch angemerkt, dass Sullins einerseits dafür argumentiert, geistesphilosophische Voraussetzungen auszuklammern [bzw. höchstens auf einem niedrigen Abstraktionsniveau zu berücksichtigen], jedoch zugleich von – damit unvereinbar scheinenden – Vermögen, wie Motivation und Verstehen spricht [siehe: 5.2.1, 5.3.2.]

Zusammenfassung 4.2.3. I

Sullins vertritt im Kontext von Verantwortung zwar kein individualistisches Prinzip bzw. Konzept von Personalität, dennoch spricht seine relationale Sicht auf Roboter – als partiell verantwortungsfähige Subjekte – dafür, dass er Verantwortung (zumindest in einer abgeschwächten Form) als etwas Dialogisches bzw. Intersubjektives versteht, das auch an einen gewissen Zweck gebunden zu sein scheint. Es gibt Hinweise dafür, dass der Autor Verantwortung bei Robotern als etwas Normatives begreift (vgl. Sorgfaltspflicht in Zusammenhang mit Pflegerobotern). Da dem jedoch ein – in Bezug auf niedrige Abstraktionsniveaus – tolerantes Konzept zu Grunde liegt, ist nicht klar, inwiefern eine solche Form von Verantwortung letztlich nicht eher kausal-deskriptiv verstanden werden kann/ sollte. Wenn es darum geht, Roboter als verantwortungsfähige Entitäten auszumachen, bezeichnet Sullins Kompetenzen philosophischer bzw. typisch menschlicher Natur als nicht notwendige Prämissen, wobei dennoch Indizien dafür vorliegen, dass der Autor nicht komplett unabhängig von geistesphilosophischen Aspekten argumentiert.

4.2.3. II) Inwiefern wird Verantwortung in Gleichsetzung/ Anlehnung/ Abgrenzung zu folgenden eng verwandten Begriffen/ Konzepten verstanden? (allgemein bzw. in Bezug auf Roboter)

4.2.3. II, a) Zurechnung

Sullins verwendet das Wort *Zurechnung* bzw. *zurechnen/ zurechenbar* (engl. *attribution, to attribute, attributable, ascription, to ascribe*) mehrfach und auch explizit in Zusammenhang mit (moralischer) Verantwortung des Menschen/ von Robotern (z. B. S.28, 29), beispielsweise an folgender Stelle: „*The machine may have no claim to consciousness, for instance, or a soul, a mind, or any of the other somewhat philosophically dubious entities we ascribe to human specialness.*“ (2006, S.29). Der Autor charakterisiert hier Verantwortung aber nicht nur angesichts des Gebrauchs des Zurechnungsbegriffs in einer Art und Weise, die eher an einen Zuschreibungsakt erinnert: So macht das obige Zitat zudem auch deutlich, dass Sullins Verantwortung konzeptuell weniger komplex (als beispielsweise klassische Konzepte) und nicht unbedingt an ein Bewusstsein für ein Gegenüber geknüpft sieht (siehe auch: 4.2.3. IV, b [Folgenbewusstsein]).

Folgende Stelle ist ein weiterer Hinweis dafür, dass der Autor, wenn er von Verantwortung spricht, dies in enger Anlehnung an ein Zurechnungsverständnis tut: „*Finally, we can ascribe moral agency to a robot when the robot behaves in such a way that we can only make sense of that behaviour by assuming it has a responsibility to some other moral agent(s).*“ (2006, S.28).

Wenn Verantwortung als etwas beschrieben wird, das der Mensch erst auf Grund eines bestimmten Verhaltens des Roboters annehmen muss (vgl. Zitat oben), dann ist Verantwortung hier höchstens in einer retrospektiven Form möglich, während der prospektive Charakter des Begriffs vernachlässigt scheint.

4.2.3. II, b) Pflicht

Von Pflicht (engl. *duty, obligation*) ist im Text expressis verbis die Rede, wenn Sullins davon spricht, dass einem verantwortlichen Roboter (bzw. einem Roboter, der sich scheinbar so verhält, dass man von Verantwortung ausgehen kann) die Überzeugung zu Grunde liegen müsse, „*that it* [Anm.: the robot] *has the duty to care for its patients*“ (2006, S.28). Hier fasst der Autor Verantwortung in Verbindung mit einer gewissen (dem Gegenüber bestehenden) Sorgfaltspflicht und somit eng in Zusammenhang mit einer Verbindlichkeit bzw. Notwendigkeit auf. Inwiefern darunter eine bloße Einhaltung einer Regel verstanden oder Verantwortung in Form einer Beachtung von Handlungsfolgen aufgefasst wird, ist nicht eindeutig zu sagen. Da Sullins jedoch ein stark herabgesetztes Abstraktionsniveau als ausreichend (für Verantwortung bzw. damit in Verbindung stehende Prämissen) erachtet, Handlungsfähigkeit (von Robotern) generell nah an Verhalten angelehnt betrachtet (siehe: 4.2.2. I) und Verantwortung (zumindest an einigen Stellen) prospektive Aspekte vermissen lässt (siehe: 4.2.3. II, a; 4.2.3. IV, b [Folgenbewusstsein]), ist wohl eher davon auszugehen, dass hier nicht von einem Entstehen eines Roboters für die Folgen seiner Handlung (in Bezug zur Norm) die Rede sein kann.

4.2.3. II, c) Schuld

In Sullins' Text finden sich zwar mehrere Begriffe, die mit Schuld in Verbindung gebracht werden können (z. B. engl. *guilty* [S.26], *culpable* [S.26], *[to] blame* [S.24, 26, 28]), der Autor verwendet diese jedoch nicht direkt, um sein eigenes Verständnis von Verantwortung näher auszuführen. Es wurde bereits mehrfach darauf hingewiesen, dass Sullins ein herabgesetztes Abstraktionsniveau bezüglich (verschiedenster Kriterien für) Verantwortung toleriert (siehe v.a.: 4.2.2. I, c; III, c). Verantwortung auf einem sehr niedrigen Abstraktionsniveau betrachtet, befindet sich in einer gewissen Nähe zu Konzepten von Verursachung bzw. Kausalität, weshalb manche Textstellen eher an den Begriff der Schuld erinnern, wenn von Verantwortung die Rede ist (Anm.: Außerdem betrachtet Sullins Akteurschaft allgemein als ein selbst-motiviertes, zielgerichtetes Verhalten, worin auch eine gewisse Kausalvorstellung zum Ausdruck kommt [siehe: 4.2.2. I, a]). Es liegen

jedoch nicht genügend bzw. keine starken Hinweise vor, um behaupten zu können, dass der Autor Verantwortung generell als Schuld versteht.

Zusammenfassung 4.2.3. II

Einige Stellen sprechen dafür, dass Sullins Verantwortung (zumindest teilweise) in einer gewissen Nähe zu Zurechnung und Pflicht versteht: So charakterisiert der Autor Verantwortung explizit in Zusammenhang mit diesen Begriffen oder beschreibt Verantwortung stellenweise in einer retrospektiven (jedoch nicht prospektiven) Form. Da Verantwortung mehrfach in gewisser Nähe zu Kausalkonzepten formuliert wird, erinnert der Begriff teilweise auch an Schuld.

4.2.3. III) Inwiefern wird Verantwortung im Sinne eines relationalen Konzepts verstanden? (allgemein bzw. in Bezug auf Roboter)

Subjekt der Verantwortung: Sullins verwendet den Begriff *Subjekt* (engl. *subject*) an keiner Stelle in direktem Zusammenhang mit Verantwortung (weder bei Menschen noch bei Robotern). Er spricht im Kontext von moral agency bei Robotern außerdem davon, „*that full subjectivity is not needed*“ (S.27, 2006). Dennoch charakterisiert er Verantwortung auf eine Art und Weise, die ohne eine/n TrägerIn der Verantwortung nicht auszukommen scheint: Zum einen beschreibt Sullins das Kriterium in einer gewissen Nähe zum Begriff der Zurechnung (siehe: 4.2.3. II, a) und somit in gewisser Weise in Form einer Subjekt-Objekt-Zuschreibung. Zum anderen hat der Autor (in Verbindung mit Robotern) ein relationales Verständnis von Verantwortung: „*it [Anm.: the robot] has a responsibility to some other agent(s).*“ (Sullins, 2006, S.28), was auf eine Beziehung zwischen Verantwortungssubjekten hindeutet. Das zuletzt gebrachte Zitat legt nahe, dass Sullins Roboter im Zusammenhang mit Verantwortung in einer Art und Weise auffasst, die stillschweigend eine Verantwortung tragende Entität vorauszusetzen scheint. Dass es sich bei Robotern jedoch nicht um ein Subjekt im traditionellen Sinne handelt (d.h. mit Vermögen/ Charakteristika, wie etwa Personalität oder Bewusstsein), ist wohl auf die Art und Weise zurückzuführen, wie der Autor Verantwortung (bzw. moral agency) charakterisiert (siehe u.a.: 4.2.2. I, III).

Objekt der Verantwortung: Sullins macht indirekt die Sorgfalt bzw. das Kümmeren um das Gegenüber im Kontext des Gesundheits- bzw. Pflegesystems als diejenige Handlung aus (S.29), für die (bzw. für deren Handlungsfolgen [Anm.: z. B. ein kranker Mensch wird geheilt/ stirbt etc.]) einem moralischen Akteur/ einer moralischen Akteurin Verantwortung zugeschrieben werden

kann. In diesem Zusammenhang thematisiert der Autor nicht nur menschliche PflegerInnen, sondern auch Pflegeroboter, die – sofern ein ausreichendes Maß an Autonomie und Intentionalität vorhanden ist – zumindest auch über ein Programm/ eine Programmierung verfügen müssen, für das/ für die gilt, dass *„the programming is complex enough that it understands its role in the responsibility of the health care system that it is operating“* (Sullins, 2006, S.29). Da Sullins autonome Roboter innerhalb eines Beziehungsnetzwerks auffasst (S.26, siehe: 4.2.2. II, a) und auch davon spricht, dass es bei einem Roboter – um moral agent sein zu können – u.a. gegeben sein muss, dass *„it [Anm.: the robot] fulfils some social role that carries with it some assumed responsibilities [...]“* (2006, S.28), ist davon auszugehen, dass der Autor als allgemeines Objekt der Verantwortung (Anm.: eines Roboters, eines Menschen) wohl diejenigen Handlungen bzw. Handlungsfolgen ausmachen würde, die sich aus dem jeweiligen, konkreten Beziehungsgefüge bzw. der zu Grunde liegenden sozialen Rolle erschließen lassen. An den beiden Zitaten dieses Absatzes lässt sich auf jeden Fall erkennen, dass das Subjekt (z. B. Pflegeroboter) und das Objekt der Verantwortung (z. B. Kümern um die Heilung/ Gesundheit der pflegebedürftigen Person) über gewisse Rollen(vorstellungen) miteinander in Beziehung stehen. (Anm.: Mit der Thematik, inwiefern davon auszugehen ist, dass/ ob Roboter bei Sullins über Folgenbewusstsein [für deren Handlungen] verfügen, beschäftigt sich v.a. 4.2.3. IV, b [Folgenbewusstsein]).

Inстанz der Verantwortung: Im Text wird in Bezug auf Verantwortung nicht direkt darauf eingegangen, vor welcher Instanz sich ein Verantwortungssubjekt zu verantworten hat. Es wurde bereits erwähnt, dass Sullins auch Roboter als potenziell verantwortliche Subjekte ausmacht (siehe: oben) und hierbei ein niedriges Abstraktionsniveau in Bezug auf als notwendig erachtete Prämissen toleriert (S.28-29). Daher ist (zumindest bei Robotern) nicht anzunehmen, dass der Autor von einer internen Instanz ausgeht (z. B. gutes Gewissen), da diese mit einem mind-less-Ansatz – den Sullins vertritt (S.27) – unvereinbar wäre.

AdressatIn der Verantwortung: Sullins spricht davon, dass Verantwortung eine Art von Verhalten ist, *„that shows and [sic] understanding of responsibility to some other moral agent.“* (2006, S.23) bzw. ein Verhalten, bei dem anzunehmen ist, dass *„it has a responsibility to some other moral agent(s).“* (2006, S.28). Als AdressatIn bzw. Gegenüber der verantwortlichen Entität wird also allgemein von anderen (potenziell) verantwortungstragenden Entitäten ausgegangen. Bei der Beschäftigung mit Pflegerobotern erwähnt Sullins eine dem Gegenüber bestehende Sorgfaltspflicht (S.28; siehe: 4.2.3. II, b), wobei naheliegend ist, dass er damit pflegebedürftige Menschen meint, denn wenig später spricht er davon, dass *„robotic caregivers are being designed to assist in the*

care of the elderly.“ (2006, S.29). Betrachtet man nun die Frage, wem (d.h. welchen Entitäten, Werten, Normen etc.) gegenüber ein Kümmern/ eine Sorgfalt bestehen sollte, dann können in Bezug auf das obige Beispiel etwa Menschen, menschliche Gesundheit, menschliche Heilung etc. als Gegenüber in Betracht gezogen werden.

Normative Kriterien der Verantwortung: Es wurde bereits mehrfach auf Normen bzw. ein normatives Verständnis bezüglich Sullins' Konzept von Verantwortung eingegangen (siehe v.a.: 4.2.3. I). Eine konkrete Norm erwähnt der Autor am Beispiel der Pflegeroboter, wenn er davon spricht, dass hier eine dem Gegenüber bestehende Sorgfaltspflicht innerhalb des Gesundheitssystems vorliegt, die „*towards the patient under its direct care.*“ (Sullins, 2006, S.29) gilt. Hier wird deutlich, dass ein normatives Kriterium (Anm.: Sorgfaltspflicht) den Verantwortungsbereich – in dem eine verantwortungsfähige Entität agiert (Anm.: Gesundheitsbereich bzw. eine konkrete, pflegebedürftige Entität unter direkter Sorgfalt) – definiert. Sullins geht außerdem anhand eines weiteren praktischen Beispiels (indirekt) auf spezifisch menschliche Normen in Bezug auf Maschinen ein: Er bezeichnet partiell fähige Roboter, die im Kontext des Polizei- oder Kriegswesens von Menschen eingesetzt werden, als problematisch und weist darauf hin, dass Entscheidungen dieser artifizienten Entitäten möglicherweise tragisch sein könnten (S.29). In diesen Fällen meint Sullins: „*Here we will have to treat the machines the way we might do for trained animals*“ (2006, S.29), was zumindest nach einer schwachen Form einer auf den Menschen bezogenen Norm bzw. Regel klingt (vgl. „*we will have to treat [...]*“ [2006, S.29]). (Anm.: Der Autor geht in der Folge darauf ein, dass in solchen Fällen der Großteil moralischer Verantwortung bzw. Schuld beim Besitzer/ bei der Besitzerin eines solchen Roboters liegt.)

Zusammenfassung 4.2.3. III

In Sullins' Text lassen sich in Zusammenhang mit Verantwortung – mehr oder weniger konkret bestimmt – mehrere Relata ermitteln und charakterisieren: Der Autor geht nicht direkt auf die Notwendigkeit eines Verantwortungssubjekts ein, mehrere Stellen sprechen jedoch dafür, dass dieser bei der Rede von Robotern – wenn auch in einer nicht traditionellen Weise – eine Verantwortung tragende Entität annimmt. Als Objekt der Verantwortung konnten allgemein jene Handlungen (bzw. Handlungsfolgen) ausgemacht werden, die sich aus dem jeweiligen konkreten Beziehungsgefüge bzw. der zu Grunde liegenden sozialen Rolle erschließen lassen, im Falle eines Pflegeroboters beispielsweise das Kümmern um eine pflegebedürftige Person. Sullins geht nicht direkt auf eine Instanz der Verantwortung ein, jedoch ist (zumindest im Bereich von Robotern) auszuschließen, dass es sich hierbei um eine interne handelt. Als AdressatIn der Verantwortung

werden allgemein andere (potenziell verantwortungstragende) Entitäten ausgemacht, im konkreten Kontext der Pflegeroboter beispielsweise (die Gesundheit des) Menschen. Ebenso wird an diesem Beispiel klar, dass Verantwortung auch mit normativen – den Verantwortungsbereich eingrenzenden – Kriterien zusammenhängt.

4.2.3. IV) Inwiefern spielen folgende Aspekte als Bedingungen der Zuschreibung von Verantwortung eine Rolle, wenn es um die Charakterisierung von moral agents geht? (allgemein bzw. in Bezug auf Roboter)

4.2.3. IV, a) Kommunikationsfähigkeit

Roboter werden vom Autor in bestimmten Situationen als Teil eines Beziehungsgefüges bzw. „*as fellow moral agents in that community* [Anm.: composed of robot, programmers, builders etc.].“ (Sullins, 2006, S.26) verstanden. Sullins spricht auch mehrfach von Mensch-Maschine- bzw. Mensch-Roboter-Interaktionen (S.24-26), von einer Interaktion zwischen Roboter und Umwelt (S.28) bzw. von zukünftigen, hochkomplexen und interaktiven Robotern (S.29). Dies mag den Anschein erwecken, dass (partiell) verantwortliche, artifizielle Entitäten wohl auch in irgendeiner Weise fähig sein müssen, zu kommunizieren (Anm.: Es wurde bereits thematisiert, inwiefern Verantwortung bei Sullins als etwas Intersubjektives bzw. Dialogisches aufgefasst werden kann [siehe: 4.2.3. I]). Die Fähigkeit der Kommunikation wird in Sullins' Text jedoch an keiner Stelle als eine konkrete/ wesentliche Verantwortungsbedingung thematisiert (weder beim Menschen, noch bei Robotern). Der Autor spricht davon, dass einem Roboter – wenn sich dieser auf eine gewisse Weise verhält – Verhalten bzw. gewisse Überzeugungen zugeschrieben werden können (S.28). Dies könnte dahingehend gedeutet werden, dass Roboter zumindest über gewisse Verhaltensweisen und auf einem sehr basalen Niveau etwas mitteilen und insofern kommunizieren können, genauer lässt sich dies jedoch nicht aus dem Text erschließen. Mit großer Sicherheit kann wohl davon ausgegangen werden, dass Sullins Kommunikationsfähigkeit in jedem Fall nicht im Sinne von Sprachlichkeit versteht (siehe: 4.2.3. I).

4.2.3. IV, b) Handlungsfähigkeit

Die Frage, was der Autor unter (moralischer) Handlungsfähigkeit versteht bzw. inwiefern diese (konzeptuell bzw. kriteriologisch) mit (moralischer) Verantwortung zusammenhängt, wurde be-

reits beleuchtet (siehe: 4.2.2. III, d). Im Folgenden sollen genauere Einsichten in Sullins' Verständnis von Handlungsfähigkeit erworben werden, indem in diesem Kontext als bedeutsam erachtete Aspekte näher beleuchtet werden:

Folgenbewusstsein: In 4.2.3. III (Objekt der Verantwortung) wurde erwähnt, dass Sullins indirekt das Kümern um das Gegenüber im Kontext des Gesundheits- bzw. Pflegesystems als diejenige Handlung ausmacht, für die bzw. für deren Handlungsfolgen einem/ einer moralischen AkteurIn Verantwortung zugeschrieben werden kann. Da der Autor jedoch Bewusstsein nicht als notwendig für Verantwortung(sfähigkeit) erachtet (vgl. „*The machine may have no claim to consciousness [...]*“ [Sullins, 2006, S.29]), ist auch nicht davon auszugehen, dass er Folgenbewusstsein (in Zusammenhang mit Handlungs- bzw. Verantwortungsfähigkeit) hierfür als obligat erachtet. Wie bereits beschrieben, charakterisiert Sullins Verantwortung auch z. T. in einer gewissen Nähe zum Pflichtbegriff bzw. zu Konzepten von Kausalität, womit z. T. prospektive Aspekte von (verantwortlichen) Handlungen zu fehlen scheinen (siehe: 4.2.2. I, a; 4.2.3. II, b). Daher ist wohl nicht davon auszugehen, dass der Autor vorausblickendes Handeln (eines Roboters) als unbedingte Prämisse für Handlungsfähigkeit erachtet.

Kontextwahrnehmung: Die Wahrnehmung einer jeweiligen Situation scheint in konkreten Handlungssituationen eine gewisse Rolle zu spielen, worauf beispielsweise folgende Stelle hindeutet: Sullins zu Folge müsse das Programm bzw. die Programmierung von Pflegerobotern komplex genug sein, dass „*it understands its role in the responsibility of the health care system*“ (2006, S.29). Ein Verstehen der Verantwortungsrolle im Gesundheitssystem (Anm.: entspricht einer gewissen Kontextwahrnehmung) scheint hier vom Autor vorausgesetzt zu werden, damit ein Pflegeroboter (bzw. eine pflegende Entität allgemein) letztlich (verantwortlich) agieren kann (S.29). Wie jedoch eine solche Wahrnehmung des Kontexts – durch den Roboter selbst – in Verbindung mit Sullins' mind-less Perspektive (siehe u.a.: 4.2.3. I) bzw. einem Mangel an Folgenbewusstsein (siehe: oben) möglich sein soll, bleibt fraglich und lässt sich aus dem Text nicht erschließen. Wenn der Autor Verantwortung als etwas charakterisiert, das der Mensch erst auf Grund eines bestimmten Verhaltens des Roboters annehmen muss (siehe: 4.3.2. III, a; S.28), dann macht dies viel eher den Eindruck, dass Kontextwahrnehmung (bzw. prinzipiell auch Handlungsfähigkeit und Verantwortung) in einer Art und Weise verstanden wird, die darauf hindeutet, dass die Einordnung/ Einschätzung einer jeweiligen (normativen) Situation eine menschliche Angelegenheit bleibt.

Personalität: Sullins macht mehrfach explizit klar (z. B. S.23, 25), dass er Personalität nicht als eine prinzipielle Voraussetzung für (moralische) Akteurschaft (von Robotern) erachtet, beispielsweise an dieser Stelle: „*I will now have to argue that personhood is not required for moral agency.*“ (2006, S.25). Der Autor betont zudem auch, dass er strikt zwischen den Begriffen *Person* und *moral agent* unterscheidet (S.23). Dennoch spricht Sullins an einer Stelle in Zusammenhang mit Intentionalität u.a. von der Notwendigkeit eines gewissen Grades an persönlicher Intentionalität (S.28; siehe: unten [Einflussmöglichkeit/ Freiheit]), was in diesem Kontext inkonsistent zu sein scheint.

Einflussmöglichkeit: Der Autor macht die Möglichkeit, Einfluss zu nehmen in Form von intentionalem Verhalten, als eine Schlüsselkompetenz für moralische Handlungsfähigkeit (von Robotern) aus (S.28). Darunter versteht Sullins jedoch keine höhergradigen Absichten (S.26) oder Intentionalität in einem strengen (menschlichen) Sinne (S.28), sondern eine viel basalere Form (siehe auch: unten [Freiheit]): Solange das Verhalten eines Roboters komplex genug ist, „*that one is forced to rely on standard folk psychological notions of predisposition or ‘intention’ to do good or harm, then this is enough to answer in the affirmative to this question* [Anm.: Is the robot’s behaviour intentional?].“ (Sullins, 2006, S.28). Sullins bringt Intentionalität mehrfach in Zusammenhang mit moral agency (S.28-29) und spricht in dieser Beziehung einerseits von intentionalem Verhalten (siehe: oben bzw. S.28), andererseits aber auch von einem gewissen Grad an persönlicher Intentionalität (siehe: unten [Freiheit] bzw. S.28), womit er eine von ihm als nicht notwendig erachtete Voraussetzung für Akteurfähigkeit (siehe: oben [Personalität]) anspricht.

Freiheit: Sullins bezeichnet philosophisch zweifelhafte sowie typisch menschliche Konzepte/ Vermögen als nicht notwendig, um über Verantwortung bzw. moral agency (von Robotern) diskutieren zu können (S.29). Im Rahmen seiner Auseinandersetzung mit Intentionalität (als einem Kriterium für moralische Handlungsfähigkeit) ist außerdem folgende Textstelle zu lesen:

„*There is no requirement that the actions [Anm.: of a robot] 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, at the level of the interaction between the agents involved, there is a comparable level of personal intentionality and free will between all the agents involved.*“ (Sullins, 2006, S.28)

Handlungen setzen laut Sullins also keinen freien Willen im strengen/ traditionellen Sinne voraus, sondern nur ein vergleichbares Niveau von freiem Willen aller an einer Handlung beteiligten Entitäten. Was der Autor unter einem solchen „*comparable level of personal intentionality and free will*“ (Sullins, 2006, S.28) versteht, erläutert er jedoch nicht weiter in diesem Zusammenhang.

Außerdem bleibt die Frage offen, inwiefern dies mit seiner grundlegenden Auffassung von Handlung als selbst-motiviertem, zielgerichtetem Verhalten (siehe: 4.2.2. I, a) in Konsens zu bringen ist.

Freiheit im Sinne von (einem signifikanten Level an) Autonomie bezeichnet der Autor in zweifacher Hinsicht als eine Voraussetzung für (moralische) Handlungsfähigkeit, denn er erachtet einerseits nur autonome Roboter als potenzielle moral agents (siehe: 4.2.2. II), und andererseits stellt die positive Beantwortung der Frage „*Is the robot significantly autonomous?*“ (Sullins, 2006, S.28) eines von drei zu erfüllenden Schlüsselkriterien dar, damit Roboter für den Verfasser als moralische Akteure gelten können (siehe: 4.2.2. III, d). Sullins definiert autonome Roboter auf einem niedrigen Abstraktionsniveau folgendermaßen: „*[They] must be capable of making at least some of the major decisions about their actions using their own programming.*“ (2006, S.26). Signifikante Autonomie fasst der Autor in einer technischen, nicht radikalen, abgeschwächten Form auf: „*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.*“ (2006, S.26).

Sullins bringt Autonomie, (moralische) Handlungsfähigkeit und Verantwortung in Zusammenhang mit gewissen kognitiven Kompetenzen, wie folgende zwei Stellen zeigen: „*For the robot to be a moral agent, it is necessary that the machine have a significant degree of autonomous ability to reason and act on those reasons.*“ (2006, S.26); „*robot moral agency requires the robot to behave in a way that shows and [sic] understanding of responsibility to some other moral agent.*“ (2006, S.23). Wie diese Vermögen (Anm.: ein Verstehen von Verantwortung, ein auf Gründen basierendes Handeln [vgl. Zitate oben]) vorzustellen sind, beschreibt der Autor hier nicht weiter, allerdings sprechen die Toleranz eines niedrigen Abstraktionsniveaus (S.29) und folgende Stelle dafür, dass er Kognitionen schon auf basalstem Niveau für (eine zumindest rudimentäre Form von moralischer) Handlungsfähigkeit als genügend erachtet:

„*No robot in the real world or that of the near future is, or will be, as cognitively robust as a guide dog. But even at the modest capabilities robots have today some have more in common with the guide dog than a hammer.*“ (Sullins, 2006, S.25)

4.2.3. IV, c) Urteilskraft

Wie bereits thematisiert wurde (siehe: 4.2.3. II, b; 4.2.3. III), spricht Sullins in Verbindung mit Verantwortung bzw. anhand des konkreten Beispiels der Pflege indirekt von einer gewissen Norm (Anm.: Sorgfaltspflicht eines Pflegers/ einer Pflegerin bzw. eines Pflegeroboters gegenüber einer

zu pflegenden Entität). Die Frage, welche Vermögen der Autor zur Ausbildung einer solchen Norm (bzw. von Normen/ eines Normensystems allgemein) für notwendig erachtet, lässt sich nicht eindeutig aus dem Text erschließen. In jedem Fall ist jedoch moralische Akteurschaft unabdingbar, die sich – im Fall von Pflegerobotern – u.a. durch ein Verstehen der Verantwortungsrolle im Gesundheitssystem auszeichnet (S.29; siehe: 4.2.3. IV, b). Allgemein kann man davon ausgehen, dass Sullins wohl ein gewisses Verständnis für den jeweiligen Kontext – in dem eine verantwortungstragende Entität agiert – als nötig erachtet. Es wurde bereits im letzten Abschnitt darauf eingegangen, inwiefern der Autor Verstehen bzw. kognitive Kompetenzen in Zusammenhang mit Handlungsfähigkeit (bzw. Verantwortung) betrachtet (siehe: 4.2.3. IV, b [Freiheit]). Auch wenn sich Sullins nicht explizit dazu äußert, ob/ wie bzw. mittels welcher Vermögen handlungsleitende Gründe letztlich entstehen können, kann nach der bisherigen Betrachtung die Vermutung angestellt werden, dass er von der Notwendigkeit einer Urteilskraft – wenn überhaupt – wohl generell höchstens auf einem sehr basalen Niveau ausgehen würde.

Zusammenfassung 4.2.3. IV

Inwiefern Sullins gewisse Aspekte als Bedingungen der Zuschreibung von (moralischer) Verantwortung(sfähigkeit) ausmacht, wird teilweise sehr klar/ explizit vom Autor festgelegt, zum Teil bleiben jedoch einige Punkte ungeklärt/ offen: Unklar erscheint, in welcher Hinsicht Kommunikationsfähigkeit oder Urteilskraft vom Autor als Verantwortungsbedingungen verstanden werden, da er nicht explizit auf diese Vermögen eingeht. Es ist jedoch davon auszugehen, dass er diese – wenn überhaupt – dann wohl nur auf einem sehr basalen Niveau als notwendige Bedingungen für Verantwortung(sfähigkeit) erachten würde. Betrachtet man Handlungsfähigkeit bzw. deren Teilaspekte als mögliche Prämisse(n) für die Zuschreibung von Verantwortungs- bzw. Handlungsfähigkeit, so bezeichnet der Autor Personalität ausdrücklich als nicht notwendig, während Folgenbewusstsein, Kontextwahrnehmung, Einflussmöglichkeit und Freiheit zum Großteil in einer Art und Weise charakterisiert werden, die darauf hindeutet, dass der Autor diese Vermögen ebenfalls höchstens auf einem sehr niedrigen Abstraktionslevel berücksichtigen würde.

5. VERGLEICHENDE DISKUSSION

Während es im vorherigen Kapitel darum ging, zwei Texte separat zu analysieren (siehe: 4.1., 4.2.), versteht sich dieser Abschnitt als eine prägnante und vergleichende Diskussion jener Erkenntnisse, die durch vorausgehende Betrachtungen gewonnen werden konnten. Im Methodenteil wurde bereits angemerkt (siehe: 3.1., 3.4.), inwiefern ein solcher Zugang wertvoll – in Bezug auf die Beantwortung der Fragestellung dieser Arbeit – sein kann. Es soll hierbei nicht darum gehen, die ausführlichen Inhalte der Analyse im Detail zu diskutieren, sondern vermehrt darum, in einer vergleichenden bzw. zusammenfassenden Art zu betrachten. Außerdem sollen im Zuge dieses Kapitels auch jene Stellen der Textanalyse kurz aufgegriffen werden, die dort (siehe: 4.2.2. III, c; 4.2.3. I) um einen kritischen Verweis auf die vergleichende Diskussion ergänzt wurden.

Die in 1.2.3. entwickelte Fragestellung der Arbeit soll an dieser Stelle nochmals vor Augen geführt werden: Welche (einheitlichen oder unterschiedlichen) Bestimmungen (v.a. hinsichtlich: Prämissen, begrifflichen/ konzeptuellen Implikationen, [formaler] Charakteristika etc.) bzw. Bedeutungen (v.a. argumentativer, kriteriologischer, konzeptueller Natur) erfährt das Kriterium der (moralischen) Verantwortung im moral agency-Diskurs der Roboterethik? Im Rahmen dieses Kapitels wird versucht, in folgenden drei Abschnitten, eine Antwort darauf zu geben:

Im ersten Teil (siehe: 5.1.) geht es darum, einen tabellarischen Überblick/ Vergleich der Charakterisierung wesentlicher, im Zusammenhang mit (moralischer) Verantwortung bzw. dem Diskurs stehender Analyseaspekte der betrachteten Texte zu geben.

Im zweiten Teil (siehe: 5.2.) soll (vergleichend) diskutiert werden, inwiefern der Kontext/ der Diskurs – d.h. (begriffliche) Implikationen und Prämissen (z. B. Verständnis von Moral, Handlungsfähigkeit, Forschungsschwerpunkte der Autoren etc.) – die jeweilige Bedeutung/ Bestimmung von (moralischer) Verantwortung in einem Text (mit)bestimmt.

Im dritten Teil (siehe: 5.3.) wird die Bedeutung/ Bestimmung von (moralischer) Verantwortung vergleichend hinsichtlich folgender Aspekte diskutiert: kriteriologisch-argumentative (siehe: 5.3.1.) / konzeptuelle (siehe: 5.3.2.) / formale (siehe: 5.3.3.) Aspekte von (moralischer) Verantwortung.

5.1. Vergleichender Überblick wesentlicher Analyseaspekte der Texte

Welche Bestimmung bzw. Bedeutung erfährt nun das Kriterium der (moralischen) Verantwortung im moral agency-Diskurs der Roboterethik? Bereits in der Einleitung wurde betont, dass eine reflektierte Auseinandersetzung mit dem Themenkomplex *Roboter(ethik) – Moral/ (moral) agency – (moralische) Verantwortung* für die Beantwortung der Fragestellung dieser Arbeit wesentlich sei (siehe: 1.2.1.). Die Textanalysen (siehe: 4.1., 4.2.) haben – basierend auf theoretischen und methodischen Vorarbeiten – ein umfangreiches Bild zur Bedeutung des Kriteriums der (moralischen) Verantwortung im moral agency-Diskurs der Roboterethik hervorgebracht, das nachstehend in übersichtlicher Form – hinsichtlich obengenannter Aspekte – tabellarisch dargestellt (siehe: Tab.6) und im Folgenden diskutiert werden soll (siehe: 5.2., 5.3.): Dazu werden zentrale Analyseaspekte aufgelistet (siehe: Tab.6, Spalte 1) und deren Charakterisierung bei Irrgang (2006) (siehe: Tab.6, Spalte 2) und Sullins (2006) (siehe: Tab.6, Spalte 3) parallel zueinander betrachtet. Die Übersicht zeigt u.a., dass hinsichtlich wesentlicher Aspekte (in Zusammenhang mit [moralischer] Verantwortung bzw. dem Diskurs) großteils sehr unterschiedliche bzw. konträre Perspektiven/ Positionen von den Autoren eingenommen werden (siehe: Tab.6; 5.2., 5.3.).

Tab.6: (Vergleichende) Betrachtung bzw. Charakterisierung wesentlicher Aspekte der Textanalysen: Analyseaspekt, der (vergleichend) betrachtet wird (Spalte 1); Charakterisierung des jeweiligen Analyseaspekts bei Irrgang (2006) (Spalte 2); Charakterisierung des jeweiligen Analyseaspekts bei Sullins (2006) (Spalte 3); (Anm.: Die Inhalte dieser Tabelle beziehen sich auf die Kapitel 4.1. [Analyse von Irrgangs {2006} Text] und 4.2. [Analyse von Sullins' {2006} Text].)

Analyseaspekt:	Irrgang (2006)	Sullins (2006)
Moral/ (moral) agency	v.a. traditionelle, exklusive, anthropozentrische, instrumentalisierende Denkmuster; Konzepte v.a. binär, nicht-graduell, nicht alternativ bzw. ohne Toleranz niedriger Abstraktionsniveaus;	v.a. alternative, inklusive, mathenozentrische, nicht-instrumentalisierende Denkmuster; Konzepte v.a. graduell, alternativ (funktional) bzw. tolerant gegenüber niedrigen Abstraktionsniveaus, mind-less-Konzept;
	<i>Moral</i> : nicht konkret definiert; Mensch als moralischer Gradmesser; ±exklusiv	<i>Moral</i> : unscharf definiert; hat mit Schaden/Leid bzw. dem Guten/ Wohl zu tun; ±inklusiv
	<i>Agency</i> : im eigentlichen Sinne nur Menschen zuschreibbar; geringe Gradualität: keine agency, abstrakte agency (Handlungsmuster), menschliche agency;	<i>Agency</i> : selbst-motiviertes, zielgerichtetes Verhalten; hängt mit Autonomie zusammen;

	<i>Moral agency</i> : nicht abstrahierbare, exklusiv menschliche Form der Akteurschaft; Personalität obligat; binäres Konzept;	<i>Moral agency</i> : liegt bei handlungsfähiger Entität vor, die signifikante Autonomie, Intentionalität und Verantwortung aufweist; Personalität nicht obligat; graduelles Konzept;
(moralische) Verantwortung	v.a. traditionelle, exklusive, anthropozentrische Perspektive; Konzept von normativer (moralischer) Verantwortung: ausschließlich im Bereich des Menschlichen, binär, nicht-graduell, nicht abstrahierbar;	v.a. nicht-traditionelle, inklusive, mathematozentrische Perspektive; Konzept von normativer (moralischer) Verantwortung: graduell, toleriert verschiedene Abstraktionsniveaus, nicht anthropozentrisch, mind-less-Zugang, Argument funktionaler Äquivalenz;
	intersubjektiv-dialogisch, im eigentlichen Sinne normativ, wenig im Kontext gewisser Kompetenzen betrachtet;	nicht typisch individualistisch bzw. nicht unbedingt an Personalität geknüpft, ±normatives (z. T. kausales [?]) Verständnis, ±unabhängig von philosophischen/ typisch menschlichen Kompetenzen;
	z. T. begriffliche Nähe zu Zurechnung, Pflicht; Nähe zu Schuld nur im Kontext von Robotern/ kausaler Verantwortung;	z. T. begriffliche Nähe zu Zurechnung, Pflicht, Schuld; (expressis verbis in Beziehung zu Zurechnung und Pflicht gestellt)
	Verantwortung als Relationsbegriff: Relata Subjekt (Mensch), Objekt, normatives Kriterium erkennbar bzw. ±deutlich (indirekt) erschließbar; Relata AdressatIn und Instanz nicht deutlich erschließbar/ thematisiert;	Verantwortung als Relationsbegriff: Relata Subjekt (auch tw. Roboter; nicht traditionell), Objekt, AdressatIn, normatives Kriterium erkennbar bzw. ±deutlich (indirekt) erschließbar; Relatum Instanz nicht (deutlich) erschließbar/ thematisiert;
	Bedingungen der Zuschreibung: Handlungsfähigkeit, v.a. Personalität bedeutend; Kommunikationsfähigkeit, Urteilstkraft unklar bzw. nicht ausreichend thematisiert (vermutlich aber anthropozentrisch aufgefasst);	Bedingungen der Zuschreibung: Handlungsfähigkeit höchstens auf sehr basalem Niveau nötig, Personalität keine Prämisse; Kommunikationsfähigkeit, Urteilstkraft unklar bzw. nicht ausreichend thematisiert (wahrscheinlich ebenfalls höchstens auf basalem Niveau als notwendige Bedingung erachtet);
Roboter (vs Mensch)	<i>Roboter</i> : z. T. relationale, instrumentalisierende Sichtweise; weder jetzt noch zukünftig moral agents, auch nicht in gradueller/ alternativer Form; höchstens kausal, jedoch nicht normativ (moralisch) verantwortungsfähig;	<i>Roboter</i> : ausgeprägte relationale Sichtweise; gewisse autonome Roboter können (graduelle, funktionale) moral agents bzw. verantwortungsfähige Entitäten darstellen;
	<i>Mensch</i> : genuiner moral agent, normativ (moralisch) verantwortungsfähig, moralischer Gradmesser;	<i>Mensch</i> : im Vergleich zu rezenten Robotern in höherem Grade moral agents bzw. moralisch verantwortungsfähig;

5.2. Die Bedeutung des Kontexts/ des Diskurses für die Bestimmung/ Bedeutung von (moralischer) Verantwortung

Sowohl in der theoretischen als auch in der methodischen Vorarbeit zur Analyse wurde mehrfach auf die Wichtigkeit der Betrachtung des jeweiligen Kontexts bzw. Diskurses – in dem (moralische) Verantwortung thematisiert wird – hingewiesen (siehe z. B. 2.2.3., 3.1.2. etc.). Nicht zuletzt wurden aus diesem Grund der Begriff der (moralischen) Verantwortung und damit in Zusammenhang stehende Begriffe/ Konzepte, wie *Moral*, *moralische Akteurschaft*, *Roboter* etc. innerhalb und außerhalb des spezifisch roboterethischen Diskurses vorab theoretisch beleuchtet (siehe: 2.1., 2.2.). In diesem Abschnitt sollen nun die Ergebnisse der Analysearbeit in Bezug auf den Kontext bzw. Diskurs vergleichend betrachtet bzw. diskutiert werden. Dies kann Hinweise dafür liefern, inwiefern diese von Relevanz für die Bestimmung/ Bedeutung von (moralischer) Verantwortung sein können.

5.2.1. *Moral* und (moral) agency als Prämissen für die Bestimmung/ Bedeutung von (moralischer) Verantwortung

Da (moralische) Verantwortung – mehr oder weniger direkt – mit einem Vorverständnis von *Moral* sowie (Aspekten von [moralischer]) Handlungsfähigkeit (z. B. *Personalität*, *Freiheit* etc.) zusammenhängt (siehe: 2.1.2.-2.1.3., 2.2.2.-2.2.4.), mag es einleuchtend erscheinen, dass die Art und Weise, wie diese Prämissen bestimmt sind, maßgeblichen Einfluss darauf haben kann, was jeweils unter bzw. wie (moralische) Verantwortung verstanden wird. Im Folgenden wird – in Rekurs auf die Textanalysen – darüber diskutiert, welche Zusammenhänge zwischen *Moral*, (moral) agency und einem jeweiligen Verantwortungsverständnis ermittelt werden konnten bzw. naheliegender scheinen:

i) *Moral*: Irrgang (2006) definiert nicht konkret, was er unter *Moral* versteht, macht aber durch zahlreiche Beispiele klar, dass er den Menschen als moralischen Gradmesser auffasst, womit man hier von einem anthropozentrischen *Moral*-verständnis ausgehen kann (siehe: 4.1.2. I bzw. Tab.6). Sullins (2006) bezeichnet *Moral* als etwas, das mit Schaden/ Leid bzw. dem Guten/ Wohl zu tun hat (siehe: 4.2.2. I bzw. Tab.6), womit er eine unscharfe und viel inklusivere Perspektive als der erstgenannte Autor einnimmt.

Welche Bedeutung kann nun das jeweilige Moralverständnis für die Bestimmung/ Bedeutung von (moralischer) Verantwortung haben? Das kommt u.a. darauf an, wie (differenziert) Verantwortung in einem jeweiligen Text verstanden wird, denn Verantwortung – im normativen, nicht rein kausalen Sinne – ist nicht unbedingt mit moralischer Verantwortung gleichzusetzen (Sombetzki, 2014; siehe: 2.1.4.2. ii). Sullins spricht in konkretem Zusammenhang mit Verantwortung zwar an keiner Stelle *expressis verbis* von moralischer Verantwortung, da der Autor Verantwortung jedoch als eine wesentliche Prämisse für moralische Akteurschaft ausmacht (siehe: 4.2.2. III, 5.3.1.) und diese in Zusammenhang mit dem Lösen moralischer Fragen stellt (siehe: 4.2.2. III, c), ist davon auszugehen, dass er (wenn auch auf einem sehr basalen Niveau) eine Art von moralischer Verantwortung meint, wenn er generell von Verantwortung spricht. In diesem Fall scheint ein derart offenes/ weites Moralverständnis letztlich ein Türöffner für ein inklusives Verantwortungsverständnis darzustellen.

Irrgang (2006) charakterisiert normative Verantwortung (die er im Gegensatz zu kausaler Verantwortung als *eigentliche Verantwortung* ausmacht), spricht auch *moralische Verantwortung* *expressis verbis* an, grenzt diese beiden in seinem Text jedoch nicht klar voneinander ab (siehe v.a.: 4.1.2. III, a), weshalb davon ausgegangen werden kann, dass er nicht eindeutig zwischen normativer und spezifisch moralischer normativer Verantwortung unterscheidet. In diesem Sinne kann das anthropozentrische Moralverständnis des Autors (siehe: oben) als exklusive Prämisse für normative Verantwortung überhaupt angesehen werden.

ii) *Agency*: Handlungsfähigkeit wurde im Rahmen der theoretischen Auseinandersetzungen mit Verantwortung als wesentliche Bedingung für die Zuschreibung von Verantwortung(sfähigkeit) ausgemacht (siehe: 2.1.4.5. ii). Die Art und Weise, wie (Aspekte von) Handlungsfähigkeit (prinzipiell bzw. in Zusammenhang mit [moralischer] Verantwortungsfähigkeit) aufgefasst wird (werden), kann somit als entscheidende Prämisse für die Bestimmung/ Bedeutung von (moralischer) Verantwortung betrachtet werden.

Vernachlässigt man abstrakte Handlungsmuster, die nach Irrgang (2006) keine direkte Relevanz in Zusammenhang mit normativer Verantwortung haben (siehe u.a.: 4.1.2. I), so charakterisiert der Autor (Aspekte von) Handlungsfähigkeit in einer Art und Weise, die wohl nur für menschliche Entitäten von direkter Relevanz ist bzw. nur von solchen verwirklicht/ eingenommen werden kann. Besonders deutlich wird das, wenn Irrgang (2006) Persönlichkeit bzw. Subjektivität als unab-

dingbare Prämissen für Handlungsfähigkeit bezeichnet und zudem weitere Vermögen, wie Folgenbewusstsein, Kontextwahrnehmung, Freiheit etc. so beschreibt, dass hier davon auszugehen ist, dass diese in einem (exklusiv) anthropozentrischen Sinne verstanden werden (siehe: 4.1.3. IV, b). Der Verfasser setzt normative bzw. auch explizit moralische Verantwortung in direkten Bezug zu Handlungsfähigkeit (siehe u.a.: 4.1.2. III, d), womit deutlich wird, dass bzw. inwiefern dessen Perspektive von agency eine wesentliche Prämisse für sein Verständnis von moralischer Verantwortung darstellt.

Sullins (2006) hingegen bezeichnet Handlungsfähigkeit als ein selbst-motiviertes, zielgerichtetes Verhalten, das mit Autonomie zusammenhängt (siehe: 4.2.2. I, a). Er betrachtet Bedingungen für die Zuschreibung von agency im Kontext von Robotern höchstens auf einem sehr basalen Niveau, auf dem (im Gegensatz zu Irrgang [2006]) Persönlichkeit ausdrücklich keine Notwendigkeit hat und weitere Aspekte für Handlungsfähigkeit (z. B. Autonomie) höchstens auf sehr niedrigem Level und in nicht typisch anthropozentrischer Weise vorausgesetzt werden (siehe u.a.: 4.2.2. I). Verantwortung und Handeln (speziell moral agency [siehe: unten]) stehen bei Sullins (2006) in mehrerer Hinsicht bzw. in enger/ konkreter Art und Weise in Zusammenhang: Verantwortungsfähigkeit ist eine Voraussetzung für moral agency (siehe u.a.: 4.2.2. III, d), grundlegende Handlungsfähigkeit ist jedoch auch erforderlich, um überhaupt verantwortlich (bzw. autonom, intentional etc.) sein zu können. In letzterem Sinne kann Sullins' (2006) Verständnis von agency als eine dem Verantwortungsverständnis zu Grunde liegende Prämisse aufgefasst werden. Es scheint jedoch nicht ganz widerspruchsfrei zu sein, dass der Autor generell basale Formen von Akteurschaft als ausreichend (in Zusammenhang mit Robotern als moral agents bzw. verantwortungsfähigen Wesen) erachtet (siehe: 4.2.3. IV, b) und dennoch bei seiner generellen Definition von Handlungsfähigkeit (siehe: dieser Absatz oben) geistesphilosophisch geladene Vokabel, wie *selbst-motiviert* oder *zielgerichtet* gebraucht (vgl. dazu: 5.3.2.).

iii) Moral agency: Da es ein Ziel dieser Arbeit ist, die Bestimmung/ Bedeutung von (moralischer) Verantwortung im moral agency-Diskurs zu betrachten, ist die Diskussion dieser Prämisse von besonderer Relevanz:

Irrgang (2006) verwendet Begriffe wie *moral agent* oder *moral agency* zwar nicht expressis verbis (siehe: 4.1.2. I, a), jedoch kann die Art und Weise, wie er Moral und Handlungsfähigkeit auffasst (siehe u.a.: oben bzw. Tab.6) bzw. (indirekt) in Verbindung mit (moralischer) Verantwortung bringt, Aufschlüsse darüber geben, inwiefern sein implizites moral agency-Verständnis für die

Bestimmung/ Bedeutung von (moralischer) Verantwortung relevant ist. Im Zuge der Analyse konnte Irrgangs (2006) Auffassung von moral agency als eine exklusiv menschliche Form von Akteurschaft ausgemacht werden sowie als nicht abstrahierbar und als binär (siehe v.a.: 4.1.2. I bzw. Tab.6). Der Autor charakterisiert auch normative bzw. moralische Verantwortung sehr ähnlich (als etwas spezifisch Menschliches, Binäres, nicht Abstrahierbares; siehe: 4.1.2. III bzw. Tab.6) und nimmt auch im Zuge seiner Bestimmung von moralischer Verantwortung direkten Bezug auf sein Konzept von Handlungsfähigkeit (vgl. *„In order to be morally responsible, however, an act needs a participant, who is characterized by personality or subjectivity.“* [Irrgang, 2006, S.12]; siehe: 4.1.2. III, d). Dies kann zumindest als ein Hinweis dafür betrachtet werden, dass die Art und Weise, wie Moral, Handlungsfähigkeit und v.a. moralische Handlungsfähigkeit aufgefasst werden, einen maßgeblichen Einfluss darauf hat, welche Bestimmung bzw. welche Bedeutung letztlich (moralischer) Verantwortung zukommt: Irrgangs (2006) anthropozentrische Perspektive auf Moral, agency und moral agency scheint eine Prämisse exklusiver Natur für sein Verständnis von normativer bzw. moralischer Verantwortung darzustellen.

Sullins (2006) bezeichnet moral agency als etwas, das handlungsfähigen Entitäten – die signifikante Autonomie, Intentionalität und Verantwortung aufweisen – zukommt (siehe: 4.2.2. I, a; II, b bzw. Tab.6). Hier gilt (moralische) Verantwortung – aus einer kriteriologischen bzw. argumentativen Perspektive – als Schlüsselkriterium/ Prämisse für moral agency, d.h. ohne Verantwortungsfähigkeit besteht keine Möglichkeit für moral agency (siehe: 5.3.1). Dennoch kann Sullins' (2006) Auffassung von moral agency umgekehrt auch als Prämisse für die Bestimmung/ Bedeutung von (moralischer) Verantwortung angesehen werden, denn dieser liegt ein generelles Moral- bzw. Handlungsverständnis zu Grunde (siehe: oben) bzw. scheint die Art und Weise, wie Verantwortung vom Autor aufgefasst wird, eng mit seinem inklusiven Verständnis von Moral bzw. agency zusammenzuhängen (siehe: oben). Kurz gesagt: Verantwortung(sfähigkeit) bestimmt bei Sullins (2006) einerseits (mit), ob jemand/ etwas moralische/r AkteurIn ist, andererseits bedingt die Art und Weise, was/ wen der Autor allgemein unter einem moral agent (bzw. unter Moral und Handlungsfähigkeit) versteht, wie Sullins (2006) letztlich Verantwortung (als ein Kriterium für moral agency) auffasst. Somit kann auch in diesem Fall davon ausgegangen werden, dass die Bedeutung und Bestimmung von (moralischer) Verantwortung maßgeblich von der Charakterisierung des zu Grunde liegenden moral agency-Konzepts abhängig ist.

5.2.2. Zusammenhang zwischen Roboter bzw. Mensch-Roboter-Verhältnis und (moralischer) Verantwortung

Die Frage, ob Roboter verantwortungsfähige Entitäten darstellen bzw. ob (moralische) Verantwortung(sfähigkeit) überhaupt von (direkter) Bedeutung für Roboter – als potenziell moralische Akteure – ist, hängt einerseits vom Verantwortungskonzept selbst (siehe u.a.: Tab.6; 5.3.2), andererseits aber auch davon ab, welche Auffassung von Robotern – in Beziehung zu Menschen, in Bezug auf (menschliche) Kompetenzen, in Bezug auf Moral und (moral) agency – in einem Text vorliegt (siehe u.a.: 2.2.). Die Art und Weise, wie Roboter (v.a. hinsichtlich obengenannter Aspekte) charakterisiert werden, stellt somit einen Teil des Kontexts dar, innerhalb dessen über (moralische) Verantwortungsfähigkeit nachgedacht werden kann. Die Analyse zweier Texte hat eine großteils konträre Sichtweise auf Roboter ergeben (siehe v.a.: 4.1.2. II, 4.2.2. II, Tab.6):

Irrgang (2006) betont stellenweise eine relationale Perspektive auf artifizielle Entitäten, spricht diesen auch gewisse Fähigkeiten und Entwicklungspotenziale zu (siehe: 4.1.2. II). Letztlich charakterisiert der Autor Roboter jedoch vorwiegend in einer instrumentalisierenden Art und Weise. Roboter werden (indirekt bzw. meist im Vergleich zu menschlicher Überlegenheit) weder als moralfähig noch als handlungsfähig im eigentlichen Sinne noch als moralisch handlungsfähige Entitäten aufgefasst (siehe v.a.: 4.1.2. II, Tab.6.). In 5.2.1. wurde bereits betont, dass Moral und (moralische) Handlungsfähigkeit bei Irrgang (2006) jedoch in engem Zusammenhang mit (moralischer) Verantwortung zu stehen scheinen. Letztlich scheint es daher wenig überraschend, dass in Irrgangs (2006) traditionellen Denkmustern in Bezug auf Roboter bzw. Moral und (moralische) Akteurfähigkeit von Robotern keine direkte Möglichkeit (Anm.: auch nicht graduell) der Betrachtung artifizieller Entitäten im Kontext von (moralischer) Verantwortung besteht. Roboter können für den Autor unter solchen Prämissen letztlich bloß kausal verantwortungsfähig bzw. von indirekter Relevanz in (menschlichen) Verantwortungsfragen sein.

Sullins (2006) nimmt – im Vergleich zu Irrgang (2006) – einen ausgeprägteren relationalen Blickwinkel auf Roboter ein (siehe v.a.: 4.2.2. II, a): Der Status bzw. die Fähigkeiten von Robotern können jeweils nur situativ beurteilt werden. In Bezug auf (Kompetenzen/ Prämissen für) Verantwortung(sfähigkeit) bedeutet das, dass der jeweilige Kontext (mit)entscheidend dafür ist, inwiefern eine artifizielle Entität als (mehr oder weniger) verantwortungsfähig (handlungsfähig, moralfähig etc.) ausgemacht werden kann. Dennoch beschäftigt sich Sullins (2006) in Zusammen-

hang mit Robotern auch mit konkreten Prämissen/ Vermögen (für Handlungsfähigkeit, Verantwortungsfähigkeit etc), bei denen er mehrfach höhere menschliche Kompetenzen anerkennt (siehe u.a.: 4.2.2. II, a; III, b). Im Gegensatz zu Irrgang (2006) bedeutet die Unterlegenheit artifizieller Systeme hinsichtlich mancher Fähigkeiten jedoch nicht, dass Roboter damit aus der Diskussion um Verantwortungsfähigkeit fallen. Dem zu Grunde scheint Sullins' (2006) inklusives/ weites Verständnis von Moral und (moral) agency (bei Robotern) zu liegen (siehe: 5.2.1.), sodass – fasst man diese als Prämissen für (moralische) Verantwortungsfähigkeit auf – die Möglichkeit nicht ausgeschlossen ist, Roboter als (zumindest graduell) verantwortungsfähige Entitäten (innerhalb eines Beziehungsnetzwerks) zu diskutieren. Das Ausmaß, in dem Roboter bei Sullins (2006) letztlich als verantwortungsfähig angesehen werden können, hängt wiederum v.a. mit dem konkreten Verantwortungskonzept des Autors zusammen (siehe u.a.: 4.2.3., 5.3.).

5.2.3. Kontextualisierender Blick auf die Autoren der Analysetexte als Prämisse für die Bedeutung/ Bestimmung von (moralischer) Verantwortung

Die Art und Weise, wie (moralische) Verantwortung verstanden wird, ob/ wie sie in Beziehung zu Moral, (moralischer) Handlungsfähigkeit, Robotern etc. gesetzt wird, kann letztlich auch davon abhängen, welchen Fokus/ Schwerpunkt bzw. vor welchem Hintergrund sich die Autoren der Analysetexte mit der Thematik auseinandersetzen. Im Folgenden wird daher vergleichend diskutiert, inwiefern Hinweise auf Zusammenhänge zwischen der Kontextualisierung (im Sinne von: 4.1.1., 4.2.1.) und der Bestimmung/ Bedeutung des Verantwortungsverständnisses in den Analysetexten vorliegen:

Sullins (2006) charakterisiert Verantwortung meist explizit und in Zusammenhang mit der Bedeutung für moralische Akteurschaft (siehe u.a.: 4.2.2. I). Dabei richtet er seinen Fokus ausdrücklich darauf, jene Bedingungen zu klären, unter denen Roboter als moral agents angesehen werden können. Eine derart mathenozentrische Perspektive auf Bedingungen für moral agency – und damit auch auf (moralische) Verantwortung – lässt sich eng in Verbindung mit den Forschungsschwerpunkten des Autors bringen: Sullins (2006) beschäftigt sich vermehrt mit Fragen/ Themen rund um Roboter/ Robotik, AI sowie im Speziellen mit dem Design autonomer ethischer Akteure (siehe: 4.2.1., I).

Irrgang (2006) befasst sich in seinem Text mit artifiziellen Systemen bzw. Robotern und mit Fragen rund um moralische Akteurschaft und Verantwortung, hingegen aus einer völlig anderen Perspektive bzw. unter anderen Gesichtspunkten (Anm.: als Sullins [2006]): Der Autor versucht nicht zu klären, unter welchen Bedingungen Roboter als moralische Akteure (bzw. verantwortungsfähige Entitäten) gelten könnten, sondern betrachtet vorwiegend, inwiefern von menschlichen Handlungen in artifiziellen Systemen ausgegangen werden kann (vgl. „*Now, are there human actions in Computer Science or in Robotics?*“ [Irrgang, 2006, S.2]; siehe v.a.: 4.1.2. I, b). Diese anthropozentrische Perspektive auf Handlungen (und letztlich auch auf [moralische] Verantwortung[sfähigkeit]) lässt sich z. T. in Verbindung mit den Forschungsschwerpunkten des Autors bringen: Irrgang (2006) beschäftigt sich v.a. mit kulturellen Fragen im Kontext von Technologie (siehe: 4.1.1., I). Dieser Fokus könnte ein Grund dafür sein, warum sich der Autor mit technischen Systemen vermehrt unter anthropozentrischen Gesichtspunkten bzw. Fragestellungen auseinandersetzt. Außerdem befasst sich Irrgang (2006) – im Vergleich zu Sullins (2006) – nicht gleichermaßen intensiv bzw. konkret mit Fragen rund um Robotik und moral agency (siehe: 4.1.1. I, 4.2.1. I) bzw. Verantwortung (siehe v.a.: 4.1.2. III, 4.2.2. III), weshalb die Bedeutung/ Bestimmung von (moralischer) Verantwortung bzw. Zusammenhänge mit/ zu weiteren Begriffen/ Konzepten im Text meist indirekter bzw. weniger explizit (als bei Sullins [2006]) ausgemacht werden konnten. (Anm.: Die Tatsache, dass die Analyse von Irrgangs [2006] Text dennoch eine umfangreiche Charakterisierung des Verantwortungsbegriffs [bzw. damit zusammenhängender Aspekte/ Bezüge zu Konzepten; siehe: 4.1.2. III, 4.1.3.] ergeben hat, ist ein Hinweis dafür, dass [moralische] Verantwortung und damit zusammenhängende Aspekte/ Konzepte vielfach latent in Texten mitschwingen können und eine tiefergehende Analyse sinnvoll ist.)

5.3. Konkrete Bestimmung und Bedeutung von (moralischer) Verantwortung

Die Analyse zweier Texte hat ein umfangreiches und größtenteils sehr konträres Bild hinsichtlich der Bestimmung und Bedeutung von (moralischer) Verantwortung (im roboterethischen moral agency-Diskurs) ergeben (siehe u.a.: 4.1.2. III, 4.1.3. bzw. 4.2.2. III, 4.2.3. III bzw. Tab.6). In diesem Kapitel geht es nun darum, dieses näher zu beleuchten, indem (moralische) Verantwortung hinsichtlich kriteriologisch-argumentativer (siehe: 5.3.1.), konzeptueller (siehe: 5.3.2.) und formaler Aspekte (siehe: 5.3.3.) betrachtet bzw. vergleichend diskutiert wird.

5.3.1. Kriteriologisch-argumentative Bestimmung/ Bedeutung

Welche argumentative Rolle spielt (moralische) Verantwortung bzw. inwiefern wird diese als ein Kriterium aufgefasst, wenn es darum geht, Roboter im moral agency-Diskurs zu beleuchten? Die Auseinandersetzung mit nur zwei Texten hat gezeigt, dass (moralischer) Verantwortung in diesem Kontext sehr unterschiedliche Funktionen zukommen können:

Sullins (2006) macht explizit klar, dass Verantwortung ein notwendiges Kriterium für moral agency (bei Robotern) darstellt (siehe u.a.: 4.2.2. II, b). Der argumentative Zusammenhang zwischen Verantwortung und moral agency lässt sich logisch-argumentativ folgendermaßen beschreiben: Wenn eine Entität (in einem gewissen Grad) verantwortungsfähig (sowie autonom und intentional) ist, dann kann diese als (zumindest graduelle[r]) moralische/r AkteurIn aufgefasst werden. Da bei Fehlen einer gewissen (vom Autor als ausreichend erachteten) Verantwortungsfähigkeit in keinem Fall von moral agency (bei Robotern) die Rede sein kann, kann diese (neben Autonomie und Intentionalität) als ein Schlüsselkriterium für moralische Akteurschaft bezeichnet werden.

Irrgang (2006) bringt (moralische) Verantwortung z. T. zwar in konkrete Beziehung zu (moralischer) Akteurschaft (siehe: 4.1.2. III, d), es zeigt sich hier jedoch weniger deutlich (Anm.: als bei Sullins [2006]), ob/ inwiefern ein (argumentativer bzw. kriteriologischer) Zusammenhang zwischen Verantwortung und moral agency besteht. Klar scheint für den Autor zu sein, dass moralische Verantwortungsfähigkeit nur im Kontext von Menschen (nicht aber bei artifiziellen Systemen) sinnvoll diskutabel ist (siehe: 4.1.2. III, d) sowie, dass diese (zumindest gewisse Aspekte von) Handlungsfähigkeit voraussetzt (siehe u.a.: 4.1.3. IV, b). Umgekehrt geht jedoch nicht klar aus dem Text hervor, inwieweit Verantwortung(sfähigkeit) eine Prämisse (bzw. ein [Schlüssel-]Kriterium) für moralische Akteurschaft darstellt.

5.3.2. Konzeptuelle Bestimmung/ Bedeutung

In 5.2. wurde diskutiert, inwiefern die jeweiligen Konzepte von (moralischer) Verantwortung auf unterschiedlichen Prämissen beruhen bzw. vom jeweiligen Kontext (z. B. Vorstellungen von Moral, [moralische] Akteurschaft, Roboter, Mensch-Roboter-Beziehung etc.) abhängig sind. In diesem Kapitel geht es nun darum, die – u.a. darauf beruhenden – durch die Textanalysen ermittelten

Auffassungen von (moralischer) Verantwortung möglichst konkret hinsichtlich konzeptueller Gesichtspunkte zu thematisieren:

Die Textanalysen haben zwei sehr unterschiedliche (bzw. größtenteils konträre) Konzepte von (moralischer) Verantwortung ergeben: ein binäres, nicht abstrahier- bzw. nicht abstufbares, teilweise relationales, exklusives, anthropozentrisches, traditionelles und nicht alternativ aufzufassendes bei Irrgang (2006) (siehe: Tab.6); ein graduelles, abstrahier- bzw. abstufbares, relationales, inklusives, mathenozentrisches, nicht-traditionelles und alternatives (funktionales) bei Sullins (2006) (siehe: Tab.6). Im Folgenden sollen diese beiden konzeptuellen Perspektiven kurz in Bezug auf deren (problematische) Implikationen diskutiert werden:

Irrgang (2006) verwendet in seinem Text zwar keines der Vokabel, die im letzten Absatz verwendet wurden, um sein Konzept zu charakterisieren (siehe: oben), die Art und Weise, wie er nicht-kausale Verantwortung in Bezug auf Menschen bzw. im Vergleich zu nichtmenschlichen Entitäten auffasst, legt eine solche Beschreibung jedoch nahe (siehe v.a.: 4.1.2. III, 5.3.3.). Die Pole seines zweiseitigen Konzepts stellen normativ bzw. moralisch verantwortungsfähig (Anm.: bei Irrgang [2006] nur Menschen) auf der einen und nicht normativ bzw. nicht moralisch verantwortungsfähig (Anm.: bei Irrgang [2006] alle nichtmenschlichen Entitäten) auf der anderen Seite dar. Problematisch kann eine solche Perspektive insofern sein, als sie stellenweise spezieisistisch argumentiert (siehe z. B. 4.1.2. III, a). Roboter (bzw. generell alle nichtmenschlichen Entitäten) fallen somit letztlich – unabhängig von deren Fähigkeiten – aus der direkten Diskussion um (zumindest rudimentäre Formen von moralischer) Verantwortung, was auf Grund der steigenden Komplexität bzw. zunehmender Kompetenzen technischer Systeme kritisch betrachtet werden sollte (siehe u.a.: 1.1., 2.2.).

In ziemlich konträrer Weise dazu befindet sich Sullins' (2006) Konzept von (moralischer) Verantwortung, das in diesem Zusammenhang v.a. Roboter betrachtet (siehe: 4.2.2. II). Es kann als graduell und abstrahier- bzw. abstufbar/ tolerant gegenüber niedrigen Abstraktionsniveaus bezeichnet werden (siehe v.a. 4.2.2. III). Hier liegt ein Kontinuum zwischen sehr basalen Formen von (moralischer) Verantwortungsfähigkeit auf der einen und hochgradigen Formen auf der anderen Seite vor (siehe v.a.: 4.2.2. III, c), das der Autor unter Berücksichtigung des jeweiligen Kontexts beschreibt (siehe u.a.: 4.2.2. II, a). Die Zwischenstufen lassen sich je nach Vorliegen bestimmter Kompetenzen – die Sullins (2006) als nötig für (moralische) Verantwortungsfähigkeit erachtet (siehe: unten) – bestimmen. Das graduelle Konzept bringt (im Gegensatz zu Irrgangs

[2006] binärem Konzept) die Möglichkeit mit sich, auch nichtmenschliche Formen von (moralischer) Verantwortung diskutieren/ beurteilen zu können (Anm.: ist daher nicht-traditionell und inklusiv) und scheint in direktem Zusammenhang mit der Herabsetzung bzw. der Toleranz des Abstraktionsniveaus zu stehen, die der Autor durch folgende Aspekte charakterisiert: genereller mind-less Zugang (siehe: 4.2.2. I, b; 5.3.3.), Absage an die Notwendigkeit geistesphilosophischer bzw. spezifisch menschlicher Prämissen (siehe u.a.: 4.2.2. III, c), Toleranz funktionaler Äquivalenz von Vermögen (siehe: 4.2.2. III, c). (Anm.: Außerdem scheinen [vermutlich mit oben beschriebenen Aspekten in Zusammenhang stehend] bei Sullins [2006] – stärker als bei Irrgang [2006] – Begriffe, wie *Zuschreibung*, *Pflicht* oder *Schuld* mit Verantwortung zu überlappen [siehe: 5.3.3., ii; Tab.6]). Bereits in der Textanalyse wurde an mehreren Stellen kritisch angemerkt, dass Sullins' (2006) Zugang in diesem Punkt problematisch bzw. z. T. widersprüchlich ist (siehe: 4.2.2. III, c; 4.2.3. I): Graduelle/ in Bezug auf das Abstraktionsniveau tolerante Konzepte stehen generell vor der Grundsatzproblematik, ob/ inwiefern/ bis zu welchem Grad eine solche Betrachtung aussagekräftig bzw. gültig sein kann. Bei Sullins' (2006) lassen sich darüber hinaus argumentative/ logische Widersprüche bzw. Schwachpunkte erkennen: Einerseits argumentiert der Autor dafür, dass geistesphilosophische Voraussetzungen nicht bzw. höchstens auf einem niedrigen Abstraktionsniveau oder im Sinne funktionaler Äquivalenz nötig für (moralische) Verantwortungsfähigkeit seien. Andererseits spricht Sullins (2006) in direktem Zusammenhang mit Verantwortung von – damit unvereinbar scheinenden – Begriffen, wie Motivation, Verstehen (siehe: 4.2.3. I) oder charakterisiert auch Handlungsfähigkeit – die eine Voraussetzung für Verantwortung darstellt (siehe: 5.2.1.) – in einer Art und Weise, die sich nicht widerspruchsfrei mit der Akzeptanz niedriger Abstraktionsniveaus zusammenbringen lässt (siehe: 5.2.1.). Die Frage, inwiefern eine äquivalente Betrachtung von Vermögen zur Bestimmung von Verantwortungsfähigkeit (siehe u.a.: 4.2.2. III, c) argumentative Gültigkeit haben kann, bleibt hier auch offen.

Die Analyse und Diskussion der Texte legt die Vermutung nahe, dass die konzeptuelle Bestimmung bzw. Bedeutung des (moralischen) Verantwortungsbegriffs einerseits hinsichtlich zahlreicher (expliziter sowie impliziter) Aspekte und Implikationen variieren kann und dass dies andererseits erheblichen Einfluss darauf hat, welche letztlich als verantwortungsfähige Entitäten ausgemacht werden können (siehe: oben).

5.3.3. *Bestimmung/ Bedeutung hinsichtlich formaler Aspekte*

Um tiefere Einsichten in die Bedeutung/ Bestimmung von (moralischer) Verantwortung zu bekommen, wurde sowohl in der theoretischen Auseinandersetzung (siehe: 2.1.4., 2.2.4.), als auch im Zuge der Textanalysen (siehe: 4.1.3., 4.2.3.) ein Schwerpunkt auf formale Aspekte des Begriffs gelegt. Bei der Analyse von Irrgangs (2006) und Sullins' (2006) Text kam es dabei zu folgenden Einsichten, die im weiteren Verlauf kurz betrachtet bzw. vergleichend diskutiert werden sollen:

i) Minimaldefinition von Verantwortung: Beide Autoren fassen Verantwortung – mehr oder weniger explizit – als etwas Normatives auf und charakterisieren diese im Kontext gewisser (Anm.: wenn auch unterschiedlicher) Kompetenzen (siehe: 4.1.3. I, 4.2.3. I, 5.3.3. iv). Außerdem bezeichnen beide (rezente) Roboter nicht (bzw. zumindest nicht ausdrücklich) als Personen, ziehen daraus jedoch verschiedene Schlüsse in Bezug auf potenzielle Verantwortungsfähigkeit artifizieller Entitäten, was wahrscheinlich auf Unterschiede hinsichtlich intersubjektiv-dialogischer Aspekte zurückzuführen ist: Während in der Art und Weise, wie sich Irrgang (2006) mit Verantwortung auseinandersetzt, intersubjektiv-dialogische Aspekte stärker betont bzw. traditioneller berücksichtigt werden (Anm.: Der Autor erachtet Personalität und Subjektivität ausdrücklich als unabdingbar für moralische Verantwortungsfähigkeit [siehe: 4.1.3. IV, b]) und rezente Roboter hier somit wohl nicht verantwortungsfähig sein können, vertritt Sullins (2006) keine typisch individualistische bzw. dialogische Position (Anm.: Der Autor erachtet u.a. Personalität als nicht notwendig für Handlungsfähigkeit [siehe u.a.: 4.2.3. I; IV, b]), wodurch die Möglichkeit bei letzterem nicht ausgeschlossen wird, dass Roboter potentiell verantwortungsfähig sein können. Dennoch ist auch bei Sullins (2006) – wenn auch in abgeschwächter Form – von einem verantwortungstragenden Subjekt (siehe: 4.2.3. III) sowie von etwas Dialogischem (Anm.: im Sinne einer relationalen Auffassung von Verantwortung [siehe: 4.2.3. I, III, 5.2.2.]) auszugehen, wenn er (moralische) Verantwortung anspricht.

Die Art und Weise, wie die beiden Autoren Verantwortung hinsichtlich (gewisser Aspekte) einer Minimaldefinition verstehen, zeigt Gemeinsamkeiten, aber auch Unterschiede, die u.a. in der Exklusivität/ Inklusivität des Konzepts zum Ausdruck kommen (siehe u.a.: 5.3.2.).

ii) Verantwortung in Abgrenzung zu ähnlichen Begriffen: Die Textanalysen haben ergeben, dass die Autoren (nicht-kausale) Verantwortung zwar nicht mit den Begriffen *Zurechnung*, *Pflicht* oder *Schuld* gleichsetzen, es konnten aber – auf Grund dessen, wie Verantwortung von diesen

charakterisiert wird – z. T. gewisse begriffliche Überschneidungen erkannt werden (siehe: 4.1.3. II, 4.2.3. II, Tab.6). Bei Irrgang (2006) ist dies – weniger stark bzw. meist indirekter, als bei Sullins (2006) – daran zu erkennen, dass er Verantwortung z. T. in Form einer Subjekt-Objekt-Zuschreibung und im Sinne der Einhaltung einer Norm begreift (siehe: 4.1.3. II, a-b). Sullins (2006) hingegen verwendet an mehreren Stellen *expressis verbis* die Begriffe *Zurechnung* und *Pflicht*, um Verantwortung zu beschreiben und charakterisiert Verantwortung auch mehrfach in retrospektiver (jedoch nicht prospektiver) Form und auf einem niedrigeren Komplexitätsgrad als Irrgang (2006) (siehe: 4.2.3. II, a-b). Diese begrifflichen Überschneidungen bzw. Vereinfachungen scheinen der Preis zu sein, den Sullins (2006) für sein graduelles bzw. niedrige Abstraktionsniveaus zulassendes Konzept von Verantwortung (siehe u.a.: 5.3.2.) zu zahlen hat. Inwiefern bzw. bis zu welcher Grenze hier mit gutem Grund (noch) von Verantwortung gesprochen werden kann, lässt sich also nicht ohne Weiteres sagen und wird z. T. in 5.3.2. diskutiert. Außerdem fasst Sullins (2006) Verantwortung teilweise auch in gewisser Nähe zu kausalen bzw. verursachenden Aspekten und somit als ein Schuldigsein auf (siehe: 4.2.3. II, c), was Irrgang (2006) nur im Rahmen von kausaler Verantwortung bzw. Robotern tut (siehe: 4.1.3. II, c; Tab.6).

iii) Verantwortung als relationaler Begriff: Bei beiden Autoren konnten im Zuge der Textanalyse (mehr oder weniger deutlich bzw. direkt erschließbar) Relata der Verantwortung ausgemacht werden (siehe: 4.1.3. III; 4.2.3. III), wobei diese – wie im Folgenden beschrieben – mit z. T. unterschiedlichen Denkvoraussetzungen und Bezügen zu den Verantwortungskonzepten (z. B. relationale Aspekte, Komplexität der Konzepte; siehe: 5.3.2.) zusammenzuhängen scheinen:

Die Textanalysen legen nahe, dass sowohl Irrgang (2006) als auch Sullins (2006) von einem verantwortungstragenden Subjekt ausgehen (siehe: 4.1.3. III; 4.2.3. III). Vor allem die unterschiedlichen Auffassungen bezüglich intersubjektiv-dialogischer Aspekte des Verantwortungsbegriffs (siehe: 5.3.3. i) scheinen jedoch zu bedingen, dass ersterer darunter nur Menschen, zweiterer darunter auch andere Entitäten (z. B. gewisse Roboter) versteht. Des Weiteren sprechen die Textanalysen dafür, dass die Autoren (implizit) auch ein Objekt der Verantwortung annehmen (siehe: 4.1.3. III; 4.2.3. III): Für Irrgang (2006) hat der Gegenstand der Verantwortung mit rein menschlichen Handlungen bzw. Handlungsfolgen (in Bezug auf Roboter: mit dem Bauen unkontrollierbarer Roboter sowie sich dadurch [potenziell] ergebender Schäden) zu tun, während Sullins (2006) in diesem Punkt eine relationalere Sichtweise einnimmt und als Objekt der Verantwortung allgemein jene Handlungen (bzw. Handlungsfolgen) ausmacht, die sich aus dem jeweiligen konkreten Beziehungsgefüge bzw. der zu Grunde liegenden sozialen Rolle erschließen lassen (siehe:

4.1.3. III; 4.2.3. III). Das Relationselement AdressatIn ist bei Irrgang (2006) wenig deutlich bzw. nur indirekt erschließbar und steht wohl in Zusammenhang mit dem Menschen bzw. menschlichen Normen und Werten (siehe: 4.1.3. III). Bei Sullins (2006) können diesbezüglich allgemein andere (potenziell verantwortungstragende) Entitäten bzw. im konkreten Kontext der Pflegeroboter beispielsweise (die Gesundheit des) Menschen ausgemacht werden (siehe: 4.2.3. III). Normative Kriterien von Verantwortung verbleiben bei Irrgang (2006) auf der Seite des Menschen/Menschlichen (z. B. menschliche Kontrollierbarkeit) (siehe: 4.1.3. III), bei Sullins (2006) können diese auch Roboter betreffen (z. B. Sorgfaltspflicht von Pflegerobotern) (siehe: 4.2.3. III). Beide Autoren gehen nicht näher auf eine Instanz der Verantwortung ein, bei Irrgang (2006) ist jedoch eher von der Möglichkeit einer internen Instanz (z. B. gutes Gewissen) auszugehen als bei Sullins (2006), was mit dem Komplexitätsgrad der jeweiligen Verantwortungskonzepte zusammenhängen scheint (siehe: 4.1.3. III; 4.2.3. III; 5.3.2).

iv) Bedingungen der Zuschreibung von Verantwortung: In 5.2. wurde bereits darauf eingegangen, inwiefern die Autoren Moral bzw. (moralische) Handlungsfähigkeit als Bedingungen für Verantwortung(sfähigkeit) thematisieren. Im Folgenden soll kurz vergleichend diskutiert werden, inwieweit Aspekte, die in 2.1.4.5. als Bedingungen der Zuschreibung von Verantwortung ausgemacht wurden, eine Rolle in diesem Kontext spielen:

Kommunikationsfähigkeit: Beide Autoren beschäftigen sich mit der Fähigkeit der Kommunikation an keiner Stelle in direktem Zusammenhang mit ihrer Bedeutung als Bedingung für Verantwortung(sfähigkeit) (siehe: 4.1.3. IV, a; 4.2.3. IV, a). Irrgang (2016) bezeichnet menschliche Kommunikationsfähigkeit – vermutlich im Sinne von Sprachlichkeit aufgefasst – als technisch unersetzbar, tut dies jedoch unabhängig von der Verantwortungsdebatte, weshalb (hier bzw. im Text allgemein) nicht klar hervorgeht, ob/ inwiefern der Autor diese als eine Bedingung hierfür auffasst (siehe: 4.1.3. IV, a). Auch in Sullins' (2006) Text wird nicht klar, inwiefern die Fähigkeit der Kommunikation von Bedeutung ist, der Autor scheint jedoch zumindest eine sehr basale Form von Mitteilungsfähigkeit vorauszusetzen, wenn von Verantwortungsfähigkeit die Rede ist (siehe: 4.2.3. IV, a). Im Gegensatz zu Irrgang (2016) würde Sullins (2006) Kommunikationsfähigkeit höchstwahrscheinlich in keinem Fall notwendig an Sprachlichkeit gebunden sehen (siehe: 4.1.3. IV, a; 4.2.3. IV, a). Der stärker/ traditioneller dialogisch-intersubjektive Charakter von Irrgangs (2006) Verantwortungsauffassung (siehe: 5.3.3. i) könnte ein Hinweis dafür sein, dass dieser Kommunikationsfähigkeit größeren Stellenwert zumisst als Sullins (2006).

Handlungsfähigkeit (siehe auch: 5.2.1. ii): Handlungsfähigkeit spielt bei beiden Autoren eine gewisse Rolle als Bedingung für die Zuschreibung von (moralischer) Verantwortungsfähigkeit, jedoch mit (deutlichen) Unterschieden in einigen Aspekten (siehe: 4.1.3. IV, b; 4.2.3. IV, b; 5.2.1. ii-iii): In Bezug auf Personalität nehmen die Verfasser einen konträren Standpunkt ein: Irrgang (2006) bezeichnet Personalität explizit als notwendig für moralische Verantwortungsfähigkeit (insofern auch für [moralische] Handlungsfähigkeit), während Sullins (2006) eindeutig zu verstehen gibt, dass dieser Aspekt nicht obligat für moralische Akteurschaft bzw. Verantwortung ist (siehe u.a.: 4.1.3. IV, b; 4.2.3. IV, b). Des Weiteren konnte aus den Texten (indirekt) erschlossen werden, dass Folgenbewusstsein bei Irrgang (2006) wahrscheinlich von gewisser Relevanz für Handlungsfähigkeit ist, bei Sullins (2006) hingegen scheint keine Unerlässlichkeit dieser Fähigkeit (bzw. von Bewusstsein allgemein) vorzuliegen (siehe: 4.1.3. IV, b; 4.2.3. IV, b). In Bezug auf weitere – in Zusammenhang mit Handlungsfähigkeit stehende – Aspekte, wie Kontextwahrnehmung, Einflussmöglichkeiten und Freiheit nehmen die Autoren relativ unterschiedliche (teilweise undeutlich/ indirekt zu ermittelnde bzw. z. T. widersprüchliche) Positionen ein, was wohl nicht zuletzt auf deren stark divergierende Handlungs- und Verantwortungskonzepte zurückzuführen ist (siehe u.a.: 4.1.2., 4.1.3. IV, b; 4.2.2., 4.2.3. IV, b; 5.2. ii-iii): Bei Irrgang (2006) kann von einer gewissen menschlichen Kontextwahrnehmung, menschlichen Einflussmöglichkeiten und einer menschlichen Freiheit – in Beziehung mit Handlungsfähigkeit – ausgegangen werden, die für artifizielle Systeme unerreichbar sind (siehe: 4.1.3. IV, b). Sullins (2006) scheint in Bezug auf Handlungs-/ Verantwortungsfähigkeit auch (implizit) von Kontextwahrnehmung, Einflussmöglichkeiten und Freiheit auszugehen, allerdings in einer nicht-traditionellen, graduellen und weniger komplexen Art und Weise (siehe: 4.2.3. IV, b), die in Zusammenhang mit seinem Handlungs-/ Verantwortungskonzept einige Fragezeichen/ Widersprüche aufwirft (siehe v.a.: 5.3.2.). (Anm.: Bei der Analyse der drei zuletzt genannten Aspekte [Kontextwahrnehmung, Einflussmöglichkeiten und Freiheit] konnten Inkonsistenzen in Bezug auf Sullins' [2006] mind-less Konzept von moral agency ermittelt werden. [siehe: 4.2.3. IV, b].)

Urteilstkraft: Inwiefern Aspekte von Urteilstkraft eine Rolle als Bedingung für die Zuschreibung von Verantwortung spielen, wird bei beiden Autoren nicht ausführlich bzw. nur unklar thematisiert und kann daher nicht mit Sicherheit angegeben werden (siehe: 4.1.3. IV, c; 4.2.3. IV, c). Die Analysen legen jedoch (in Bezug auf zu Grunde liegende Handlungs- bzw. Verantwortungskonzepte der Autoren/ Texte) nahe, dass Irrgang (2006) wahrscheinlich typisch menschliche Vermögen – wie etwa menschliche Reflexionsfähigkeit oder Rationalität – zur Ausbildung handlungsleitender Gründe anerkennen würde, während Sullins (2006) – wenn überhaupt – vermutlich

höchstens von einer sehr basalen Form von Urteilskraft ausgehen würde (siehe: 4.1.3. IV, c; 4.2.3. IV, c).

6. SCHLUSSTEIL

Nachdem die Texte ausführlich analysiert und (vergleichend) diskutiert wurden (siehe: 4., 5.), versteht sich der Schlussteil als eine kurze, zusammenfassende Betrachtung, bei der v.a. darauf eingegangen werden soll, inwiefern das Ziel der Arbeit erreicht werden konnte und welche Fragen (noch) offen bleiben.

6.1. *Schlussbemerkung zur Vorgehensweise*

Ausgehend von der Relevanz und Aktualität roboterethischer Problemstellungen (siehe: 1.1.), setzte sich die Arbeit zum Ziel, folgende Fragestellung zu bearbeiten (siehe: 1.2.3.): Welche Bestimmung bzw. Bedeutung erfährt das Kriterium der (moralischen) Verantwortung im moral agency-Diskurs der Roboterethik? Besonderer Fokus wurde hierbei auf Prämissen des spezifischen Diskurses sowie auf formale Charakteristika des Verantwortungsbegriffs gelegt (siehe: 1.2.2., 1.2.3.). Um das Ziel der Arbeit zu erreichen, wurden – nach ausführlichen theoretischen und methodischen Vorarbeiten (siehe: 2., 3.) – zwei als adäquat erachtete Texte (siehe: 3.3.) einer umfangreichen Analyse unterzogen, die die zu Grunde liegende Fragestellung v.a. unter kontextualisierenden/ argumentativen/ formalen Schwerpunkten bearbeitete (siehe u.a.: 3.2., Tab.5). Die aus den separat durchgeführten Textanalysen gewonnenen Erkenntnisse (siehe: 4.1., 4.2.) wurden abschließend vergleichend diskutiert (siehe: 5.), um eine möglichst umfassende bzw. tiefgehende Antwort auf die Fragestellung der Arbeit zu liefern (siehe u.a.: 6.3.).

6.2. *Wesentliche Erkenntnisse der Arbeit*

Die kritische Auseinandersetzung mit der Bestimmung/ Bedeutung des Kriteriums der (moralischen) Verantwortung im moral agency-Diskurs der Roboterethik hat umfangreiche Erkenntnisse geliefert, die im Folgenden knapp anhand von zwei Kernpunkten zusammengefasst werden sollen (Anm.: sämtliche Analyseergebnisse bzw. vergleichende Diskussionen dazu, siehe: 4., 5.):

i) (Moralische) Verantwortung kann im roboterethischen moral agency-Diskurs sehr unterschiedliche Bestimmungen/ Bedeutungen erfahren, v.a. hinsichtlich folgender Aspekte:

- *kriteriologisch-argumentative Aspekte* (siehe u.a.: 5.3.1.): (Moralische) Verantwortung kann etwa die Bedeutung eines Schlüsselkriteriums für moral agency (von Robotern) einnehmen oder wesentlich unbestimmter/ untergeordneter oder von weniger Relevanz in diesem Kontext sein.
- *konzeptuelle Aspekte* (siehe u.a.: 5.3.2.): (Moralische) Verantwortung kann hinsichtlich konzeptueller Aspekte sehr unterschiedlich bzw. konträr bestimmt bzw. von Bedeutung sein, wobei dies z. T. nur indirekt aus Texten erschlossen werden kann. V.a. hinsichtlich folgender Eigenschaften der (moralischen) Verantwortungskonzepte wurden Unterschiede festgestellt: Gradualität (nicht-graduelles [binäres] – graduelles Konzept), Abstraktionsniveau (nicht abstrahierbares – abstrahierbares Konzept), Relationalität (mehr oder weniger relationales Konzept), Exklusivität (exklusives – inklusives Konzept), Anthropozentrismus (stark anthropozentrisches – nicht anthropozentrisches bzw. sogar mathenozentrisches Konzept), Traditionalität (traditionelles bzw. nicht alternatives – nicht traditionelles bzw. alternatives [funktionales] Konzept). Es zeigte sich außerdem, dass die Konzepte (v.a. Sullins' [2006]) teilweise (logische, begriffliche) Inkonsistenzen/ Widersprüche/ Unklarheiten aufweisen.
- *formale Aspekte* (siehe u.a.: 5.3.3.): (Moralische) Verantwortung kann hinsichtlich formaler Aspekte unterschiedlich bestimmt bzw. von Bedeutung sein, wobei dies wohl maßgeblich von zu Grunde liegenden Prämissen/ konzeptuellen Aspekten etc. abhängig ist (siehe: unten) und z. T. nur indirekt oder nicht eindeutig aus Texten erschlossen werden kann: (Moralische) Verantwortung wurde als etwas ausgemacht, das – mehr oder weniger explizit/ deutlich bzw. je nach Autor in ungleicher Art und Weise – mit einem normativen Geschehen, Kriterien für die Zuschreibung von Verantwortung und einem Rede-und-Antwort-Stehen zu tun hat. Es konnten mehrfach begriffliche Überschneidungen zwischen (moralischer) Verantwortung und Begriffen, wie *Zurechnung* und *Pflicht* sowie z. T. eine gewisse Nähe zum Schuldbegriff erkannt werden. Die Analysen legen nahe, dass bestimmte Relata von Verantwortung (implizit) angenommen werden (v.a. Subjekt, Objekt), gewisse Relationselemente sind weniger deutlich erkennbar bzw. werden weniger deutlich thematisiert. Es bestehen große Unterschiede bezüglich Bedingungen der Zuschreibung von Verantwortungsfähigkeit, die hinsichtlich mehrerer Aspekte abweichen.

ii) *Die konkrete Bedeutung/ Bestimmung des Kriteriums der (moralischen) Verantwortung hängt wesentlich von der kontextuellen Einbettung bzw. der Bestimmung diskursprägender Prämissen ab, v.a. von:*

- *Moral, agency, moral agency* (siehe u.a.: 5.2.1.): Die Analysen (bzw. die Diskussion der Analysen) legen die Vermutung nahe, dass das Verständnis von Moral, agency und moral agency – das einem Text jeweils (explizit/ implizit) zu Grunde liegt – entscheidend für die jeweilige Bestimmung/ Bedeutung von (moralischer) Verantwortung ist. So scheint etwa ein anthropozentrisches, traditionelles Moral- bzw. Handlungsverständnis auch wenig bis keine Möglichkeit für alternative Verantwortungskonzepte zuzulassen.
- *Roboter bzw. Mensch-Roboter-Verhältnis* (siehe u.a.: 5.2.2.): Die Frage, ob Roboter (moralisch) verantwortungsfähige Entitäten sein könn(t)en, hängt einerseits vom konkreten Verantwortungskonzept ab (siehe: oben), andererseits aber auch wesentlich davon, wie Roboter (allgemein, in Bezug auf Kompetenzen/ Fähigkeiten, in Beziehung zu Menschen) in Texten aufgefasst werden.
- *Autoren (bzw. allgemein AutorInnen) der Analysetexte* (siehe u.a.: 5.2.3.): Eine kontextualisierende Auseinandersetzung (und vergleichende Diskussion) mit den Autoren der Analysetexte hat Hinweise darauf gegeben, dass das jeweilige Arbeitsfeld bzw. die Forschungsschwerpunkte der Verfasser Einfluss darauf haben können, wie (moralische) Verantwortung bestimmt wird bzw. welche Bedeutung diese als ein Kriterium für moral agency im roboterethischen Diskurs einnimmt.

6.3. Abschließende Reflexion, offene Fragen und Ausblick

Die Zielsetzung der Arbeit (siehe: 1.2.) konnte insofern erreicht werden, als eine umfassende/ elaborierte Darstellung davon erstellt werden konnte, inwiefern (moralische) Verantwortung von Relevanz (als ein Kriterium) im roboterethischen moral agency-Diskurs ist. Die Auseinandersetzung mit der konkreten Bedeutung/ Bestimmung des zu untersuchenden Kriteriums ergab ein umfangreiches, z. T. als konträr bestimmtes Bild, das nicht in allen Punkten klar ausformuliert bzw. widerspruchsfrei erscheint (siehe u.a.: 5.3.2.). Das methodische Vorgehen (siehe: 3.) be-

rücksichtigte dabei – nach theoretischen Vorarbeiten (siehe: 2.) – verschiedene (vorwiegend formale) Aspekte des Verantwortungsbegriffs bzw. Diskurs/ Kontext/ Prämissen und diskutierte die Ergebnisse vergleichend (5.), um möglichst tiefgehende Erkenntnisse zu erlangen.

Die Erkenntnisse der Arbeit sollten jedoch hinsichtlich einiger Punkte (die zum Großteil im Rahmen der Arbeit bereits diskutiert wurden) kritisch beleuchtet werden:

Einerseits sind diese zu relativieren, da – dem Umfang einer Masterarbeit geschuldet – nur zwei Texte in dieser Ausführlichkeit analysiert und vergleichend diskutiert werden konnten. In 3.4. wurde daher bereits angemerkt, dass größer angelegte Analysearbeiten noch aussagekräftiger bzw. erkenntnisbringender wären.

Die vorliegende Arbeit hat außerdem einen starken Fokus auf die Betrachtung formaler und argumentativer Charakteristika des (moralischen) Verantwortungsbegriffs gelegt (siehe z. B. 3.2., Tab.5), wobei hier anzumerken ist, dass auch andere Zugänge/ Untersuchungsschwerpunkte denkbar wären.

Außerdem lässt die Arbeit letztlich einige – während der Analyse und Diskussion angefallene – Fragen/ Problemfelder mehr oder weniger offen bzw. unbeantwortet: So konnten etwa für den Diskurs mitentscheidende Grundsatzfragen (z. B. Inwiefern können Äquivalenzargumente bei der Betrachtung von Vermögen zur Bestimmung von Verantwortungsfähigkeit argumentative Gültigkeit haben? Kann bzw. inwiefern/ bis zu welchem Grad kann bei gradueller moral agency/ moralischer Verantwortung [noch] mit gutem Grund von moralischer Handlungsfähigkeit/ moralischer Verantwortung die Rede sein? [siehe u.a.: 5.3.2.]) höchstens am Rande berührt werden.

Die Debatte um (moralische) Verantwortung, moralische Akteurschaft und Roboter ist noch längst nicht abgeschlossen. Da davon auszugehen ist, dass der Einfluss sowie die Komplexität und Möglichkeiten technischer Systeme weiter steigen werden (Floridi & Sanders, 2004; Lin, 2012), wird eine ethische Reflexion des in dieser Arbeit betrachteten Diskurses zukünftig wohl noch brisanter und relevanter.

7. LITERATURVERZEICHNIS

- Allen, C., Wallach, W., Smit, I. (2011). „Why Machine Ethics?“ In: Anderson, M., Anderson S., L. (Hrsg.): *Machine Ethics*. New York: Cambridge University Press, S. 51–61.
- Anderson, M., Anderson S., L. (2011). *Machine Ethics*. New York: Cambridge University Press
- Asaro, P. M. (2006): „What Should we Want From a Robot Ethic?“. In: *International Review of Information Ethics*. 6 (12), S. 9–16.
- Bayertz, K. (1995). „Eine kurze Geschichte der Herkunft der Verantwortung“. In: Bayertz, K. (Hrsg.): *Verantwortung – Prinzip oder Problem?* Darmstadt: Wissenschaftliche Buchgesellschaft, S. 3–71.
- Bayertz, K., Beck, B. (2017). „Der Begriff der Verantwortung in der Moderne: 19.-20. Jahrhundert“. In: Heidbrink, L., Langbehn, C., Loh, J. (Hrsg.): *Handbuch Verantwortung*. Wiesbaden: Springer VS, S. 133–147.
- Bekey, G. A. (2012). „Current Trends in Robotics: Technology and Ethics“. In: Lin, P., Abney, K., Bekey, G. A. (Hrsg.): *Robot Ethics. The Ethical and Social Implications of Robotics*. Cambridge/ London: MIT Press, S. 17–34.
- Bendel, O. (2016). 300 Keywords Informationsethik: Grundwissen aus Computer-, Netz- und Neue-Medien-Ethik sowie Maschinenethik. Wiesbaden: Springer Gabler.
- Birnbacher, D. (2003). *Analytische Einführung in die Ethik*. Berlin: Walter de Gruyter.
- Bringsjord, S. (2007). „Ethical Robots: The Future Can Heed Us“. In: *AI and Society*. 22 (4), S. 539–550.
- Brüsemeister, T. (2008). *Qualitative Forschung. Ein Überblick*. Wiesbaden: VS Verlag für Sozialwissenschaften/ GWV Fachverlage GmbH.
- Coeckelbergh, M. (2014). „The Moral Standing of Machines: Towards a Relational and Non-Cartesian Moral Hermeneutics“. In: *Philosophy and Technology*. 27, S. 61–77.

Damschen, G., Schönecker, D. (2013). *Selbst philosophieren. Ein Methodenbuch*. Berlin/ Boston: Walter de Gruyter GmbH.

Dennett, D. C. (1997). „When Hal Kills, Who’s to Blame? Computer Ethics“. In: Stork, D. (Hrsg.): *Hal’s Legacy: 2001’s Computer as Dream and Reality*. Cambridge, MA: MIT Press. S. 351–365.

Dennett, D. C. (1998): *Brainchildren. Essays on Designing Minds*. London: Penguin Books.

Dreisholtkamp, U. (1999). Jacques Derrida. München: Beck.

Feenberg, A. (1991). *Critical Theory of Technology*. New York: Oxford University Press.

Flatscher, M., Posselt, G., Weiberg, A. (2011). *Wissenschaftliches Arbeiten im Philosophiestudium*. Wien: Facultas Verlags- und Buchhandels AG.

Flick, U. (2004). *Triangulation. Eine Einführung*. Wiesbaden: VS Verlag für Sozialwissenschaften/ GWV Fachverlage GmbH.

Floridi, L. (2011). „On the Morality of Artificial Agents“. In: Anderson, M., Anderson, S. L. (Hrsg.): *Machine Ethics*. Cambridge: Cambridge University Press, S. 184–212.

Floridi, L., Sanders, J. W. (2004). „On the Morality of Artificial Agents“. In: *Minds and Machines*. 14 (3), S. 349–379.

Fonnesu, L. (2017). „Der Begriff der Verantwortung in der Neuzeit und in der Aufklärung“. In: Heidbrink, L., Langbehn, C., Loh, J. (Hrsg.): *Handbuch Verantwortung*. Wiesbaden: Springer VS, S. 111–132.

Frankfurt, H. G. (1969). „Alternate Possibilities and Moral Responsibility“. In: *The Journal of Philosophy*. 68 (1), S. 5-20.

Gunkel, D. J. (2012). *The Machine Question. Critical Perspectives on AI, Robots, and Ethics*. Cambridge/ London: The MIT Press.

Günther, K. (2006). „Aufgaben- und Zurechnungsverantwortung“. In: Heidbrink, L., Hirsch, A. (Hrsg.): *Verantwortung in der Zivilgesellschaft. Zur Konjunktur eines widersprüchlichen Prinzips*. Frankfurt/ New York: Campus, S. 295–329.

Heidbrink, L. (2007). *Handeln in der Ungewissheit. Paradoxien der Verantwortung*. Berlin: Kulturverlag Kadmos.

Heidbrink, L. (2017). „Definitionen und Voraussetzungen der Verantwortung“. In: Heidbrink, L., Langbehn, C., Loh, J. (Hrsg.): *Handbuch Verantwortung*. Wiesbaden: Springer VS, S. 3–33.

Himma, K., E. (2009). „Artificial Agency, Consciousness, and the Criteria for Moral Agency: What Properties Must an Artificial Agent Have to Be a Moral Agent?“. In: *Ethics and Information Technology*. 11 (1), S.19–29.

Höffe, O., Forschner, M. (2008). *Lexikon der Ethik*. München: Verlag C. H. Beck.

Ichbiah, D (2005). *Roboter. Geschichte, Technik, Entwicklung*. München: Knesbeck Verlag.

Irrgang, B. (2006). „Ethical Acts in Robotics“. *Ubiquity*, 7 (34), S. 1–15.

Johnson, D. G. (2011). „Computer systems: Moral entities but not moral agents“. In: Anderson, M., Anderson, S. L. (Hrsg.): *Machine Ethics*. Cambridge: Cambridge University Press, S. 168–183.

Johnson, D. G., Miller, K. M. (2008). „Un-making artificial moral agents“. In: *Ethics and Information Technology*. 10 (2-3), S. 123–133.

Johnson, E. (2001). „agent-centered morality“. In: Becker, L. C., Becker, C. B. (Hrsg.): *Encyclopedia of Ethics*. New York: Routledge, S. 40–43.

Kornwachs, K. (2013). *Philosophie der Technik. Eine Einführung*. München: Verlag C. H. Beck.

Kutz, C. (2002): „Responsibility“. In: Coleman, J. L. (Hrsg.): *The Oxford Handbook of Jurisprudence and Philosophy of Law*. Oxford: Oxford University Press, S. 548–587.

Langbehn, C. (2017). „Verantwortung, Identität und Autonomie“. In: Heidbrink, L., Langbehn, C., Loh, J. (Hrsg.): *Handbuch Verantwortung*. Wiesbaden: Springer VS, S. 315–336.

Lin, P. (2012). „Introduction to Robot Ethics“. In: Lin, P., Abney, K., Bekey, G. A. (Hrsg.): *Robot Ethics. The Ethical and Social Implications of Robotics*. Cambridge/ London: MIT Press, S. 3–15.

Lin, P., Abney, K., Bekey, G. A. (2012). *Robot Ethics. The Ethical and Social Implications of Robotics*. Cambridge/ London: MIT Press.

Lin, P., Bekey, G. A., Abney, K. (2008): *Autonomous Military Robotics: Risks, Ethics, and Design*. Prepared for: US Department of Navy, Office of Naval Research.

Loh, J. (2017a): „Roboterethik. Janina Loh über eine noch junge Bereichsethik“. In: *Information Philosophie*. 1/2017, S. 20-33.

Loh, J. (2017b): „Verantwortung und Roboterethik – ein kleiner Überblick – Teil 1“. In: *Zeitschrift zum Innovations- und Technikrecht (InTeR)*. 4/2017, S. 220–226.

Loh, W. Loh, J. (2017). „Autonomy and Responsibility in Hybrid Systems. The Example of Autonomous Cars“. In: Lin, P., Jenkins, R., Abney, K. (Hrsg.): *Robot Ethics 2.0. From Autonomous Cars to Artificial Intelligence*. New York: Oxford University Press, S. 35–50.

Matthias, A. (2004). „The responsibility gap: Ascribing responsibility for the actions of learning automata“. In: *Ethics and Information Technology*. 6, S. 175–183. Dordrecht: Kluwer Academic Publishers.

Meyer, S. S., Hause, J. P. (2017). „Der Begriff der Verantwortung in der Antike und im Mittelalter“. In: Heidbrink, L., Langbehn, C., Loh, J. (Hrsg.): *Handbuch Verantwortung*. Wiesbaden: Springer VS, S. 87–109.

Misselhorn, C. (2013): „Robots as Moral Agents?“. In: Rövekamp, F.; Bosse, F. (Hrsg.): *Ethics in Science and Society: German and Japanese Views*. München: Iudicum, S. 42–56.

Nadeau, J. E. (2006). „Only Androids Can Be Ethical“. In: Ford, K. M., Glymour, C., Hayes, P. (Hrsg.): *Thinking About Android Epistemology*. Cambridge, MA: MIT Press. S. 241–248.

Nida-Rümelin, J. (2007): „Politische Verantwortung“. In: Heidbrink, L., Hirsch, A. (Hrsg.): *Staat ohne Verantwortung? Zum Wandel der Aufgaben von Staat und Politik*. Frankfurt am Main: Campus Verlag, S. 55–85.

Pauen, M. (2001): „Freiheit und Verantwortung. Wille, Determinismus und der Begriff der Person“. In: *Allgemeine Zeitschrift für Philosophie* 26, S. 23-44.

Pieper, A. (2017). *Einführung in die Ethik*. Tübingen/ Basel: A. Francke Verlag.

- Precht, P., Burkard, F. (1999). Metzler Philosophie Lexikon. Stuttgart: J. B. Metzler.
- Ritter, J., Gründer, K., Gabriel, G. (2011). Historisches Wörterbuch der Philosophie. Band 11: U-V. Basel: Schwabe & Co. AG.
- Scalet, S., Arthur, J. (2016). Morality and Moral Controversies. Readings in Moral, Social and Political Philosophy. New York: Routledge.
- Schönwälder-Kuntze, T. (2015). Philosophische Methoden zur Einführung. Hamburg: Junius Verlag.
- Singer, P. (1976). „All animals are equal“. In: Regan, T., Singer, P. (Hrsg.): *Animal Rights and Human Obligations*. New Jersey: Prentice Hall, S.148–162.
- Singer, P. (1999). Practical Ethics. Cambridge: Cambridge University Press.
- Sombetzki, J. (2014). Verantwortung als Begriff, Fähigkeit, Aufgabe. Eine Drei-Ebenen-Analyse. Wiesbaden: Springer VS.
- Sparrow, R. (2012). „Can Machines be People? Reflections on the Turing Triage Test“. In: Lin, P., Abney, K., Bekey, G. A. (Hrsg.): *Robot Ethics. The Ethical and Social Implications of Robotics*. Cambridge/ London: MIT Press, S. 301–315.
- Sullins, J. P. (2006). „When is a Robot a Moral Agent?“. In: *International Review of Information Ethics*. 6 (12), S. 23–30.
- Sullins, J. P. (2019). *John P Sullins*. Zugriff am 28.05.2019 unter: <https://johnsullins.academia.edu/>
- Van Wynsberghe, A. (2015). Healthcare Robots. Ethics, Design and Implementation. London/ New York: Routledge.
- Wallace, R. J. (1994): Responsibility and the Moral Sentiments. Cambridge: Harvard University Press.
- Wallach, W., Allen, C. (2009). Moral Machines. Teaching Robots Right from Wrong. New York: Oxford University Press.

Abstract (English)

Robots are increasingly becoming more complex, competent, relevant to social life and therefore the subject of discussions concerning moral agency. In this context, (moral) responsibility often plays a decisive role. Against this background, this master's thesis aims to shed more light on the meaning/ significance of the criterion of (moral) responsibility in the moral agency-discourse of robot ethics.

For this purpose – and after extensive theoretical preparatory work – two selected texts/ papers from the field of robot ethics (*Ethical Acts in Robotics* [Irrgang, 2006] and *When is a Robot a Moral Agent?* [Sullins, 2006]) were examined by means of a philosophically founded, triangulating method with regard to contextual, argumentative and formal aspects of a concept of (moral) responsibility. Subsequently, the results of the analysis were discussed in a comparative manner.

The work provides some profound insights into the meaning of (moral) responsibility in the moral agency-discourse of robot ethics: It was determined that and to what extent the contextual embedding (e.g. understanding of morality, agency, robots etc.) shapes the respective understanding of (moral) responsibility. In addition, different aspects of (moral) responsibility (criteriological-argumentative, conceptual and formal) in the texts were comparatively analysed and critically examined. From the diverse, sometimes contrary and contradictory insights obtained, numerous details, trends, issues, correlations and references in respect of the meaning of the criterion of (moral) responsibility could be identified in the discourse under consideration.

Although the significance of this work should be put into perspective mainly due to its limited scope (for example, only two texts could be analysed/ compared in detail and only some philosophically fundamental problems related to the topic could be addressed), it can nevertheless make at least a small contribution to a debate of increasing moral relevance.

Abstract (Deutsch)

Roboter werden zunehmend komplexer, kompetenter, für das soziale Geschehen relevanter und geraten vermehrt in die Diskussion um moralische Akteurschaft. (Moralische) Verantwortung spielt in diesem Kontext vielfach eine (mit)entscheidende Rolle. Vor diesem Hintergrund setzt sich die vorliegende Masterarbeit das Ziel, die Bedeutung des Kriteriums der (moralischen) Verantwortung im roboterethischen moral agency-Diskurs näher zu beleuchten.

Zu diesem Zweck wurden – nach eingehenden theoretischen Vorarbeiten – zwei ausgewählte Texte aus dem Bereich der Roboterethik (*Ethical Acts in Robotics* [Irrgang, 2006] und *When is a Robot a Moral Agent?* [Sullins, 2006]) mittels einer philosophisch begründeten, triangulierenden Methode hinsichtlich kontextueller, argumentativer und formaler Aspekte eines Verantwortungsbegriffs beleuchtet. Die Analyseergebnisse wurden anschließend vergleichend diskutiert.

Durch die Arbeit konnten einige elaborierte Einsichten in die Bedeutung/ Bestimmung von (moralischer) Verantwortung im roboterethischen moral agency-Diskurs gewonnen werden: So wurde ermittelt, dass bzw. inwiefern die kontextuelle Einbettung (z. B. Verständnis von Moral, Handlungsfähigkeit, Roboter etc.) das jeweilige Verständnis von (moralischer) Verantwortung mitprägt. Außerdem wurden unterschiedliche Aspekte von (moralischer) Verantwortung (kriteriologisch-argumentativer, konzeptueller, formaler Natur) in den Texten vergleichend beleuchtet und kritisch betrachtet: Aus den daraus gewonnenen vielfältigen, z. T. konträren und widersprüchlichen Erkenntnissen konnten zahlreiche Details, Tendenzen, Problempunkte, Zusammenhänge und Hinweise in Bezug auf die Bedeutung des Kriteriums der (moralischen) Verantwortung im betrachteten Diskurs ermittelt werden.

Die Aussagekraft dieser Arbeit ist zwar v.a. auf Grund ihres begrenzten Umfangs zu relativieren, so konnten beispielsweise nur zwei Texte eingehend analysiert/ verglichen und mit der Thematik in Zusammenhang stehende, philosophische Grundsatzprobleme nur z. T. angeschnitten werden, dennoch kann sie zumindest einen kleinen Beitrag für eine Debatte von zunehmender moralischer Relevanz leisten.