Invited chapter to appear 28 Feb 2024 in
Jörg Noller & Georgios Karageorgoudis (eds.)(2024). *Kollektive Personen*. Mentis.
https://brill.com/edcollbook/title/64763?language=de

Soziale Roboter: Schuld ohne Strafe

Verteilte Verantwortung in Mensch-Maschine-Interaktionen

Zusammenfassung
Nicht nur im Arbeitsleben, sondern auch im Sozialen spielen künstliche Systeme immer häufiger eine wichtige Rolle. Angesichts der Vielfalt der Mensch-Maschine-Interaktionen stellt sich in manchen Fällen die Frage, ob solche Interaktionen alle als Werkzeuggebrauch zu beschreiben sind. Wenn man jedoch nicht alle Mensch-Maschine-Interaktionen einfach auf Werkzeuggebrauch reduzieren kann, dann kommt man nicht umhin, Fragen nach der Bewertung solcher Interaktionen zu stellen. Ein Aspekt dieser Fragen bezieht sich auf die Verteilung moralischer Verantwortung.
Im Kontrast zu Positionen, die die bloße Möglichkeit verneinen, künstlichen Systemen moralische Verantwortung zuzuschreiben, stellt dieser Aufsatz die Frage, ob es unter bestimmten Umständen denkbar wäre, auch künstliche Agenten als moralische Agenten zuzulassen. Damit leistet dieser Aufsatz einen Beitrag zu der Diskussion über die Verteilung von Verantwortung zwischen künstlichen Agenten und menschlichen Interaktionspartner*innen und fragt, ob die Zuschreibung von Verantwortung ausschließlich auf der menschlichen Seite verbleiben kann.
Unter Anerkennung kategorialer Unterschiede zwischen lebendigen Menschen und nicht-lebendigen künstlichen Systemen, die sich in einem asymmetrischen Merkmal aller Mensch-Maschine-Interaktionen niederschlagen, werden in diesem Beitrag Kriterien herausgearbeitet, die verteilte Verantwortung in bestimmten Mensch-Maschine-Interaktionen rechtfertigen können. Zu diesem Zweck werden sowohl interaktionsbezogene Kriterien als auch Kriterien, die sich aus sozial konstruierten Verantwortungsbeziehungen ableiten lassen, untersucht. Der Schwerpunkt liegt dabei auf der Bewertung von interaktionsbezogenen Kriterien, bei denen es sich zeigt, dass künstliche Agenten in einigen Aspekten die Fähigkeiten von Menschen übertreffen können.
Diese ‚Überlegenheit' steht in einem Gegensatz zu den üblicherweise sozial konstruierten Verantwortungsbeziehungen, anhand deren eine moralische Verantwortung künstlicher Entitäten prinzipiell ausgeschlossen wird. Ziel dieses Beitrages ist es, Konstellationen zu untersuchen, in denen es plausibel erscheint, dass moralische Verantwortung zwischen künstlichen und menschlichen Agenten verteilt werden kann. Damit wird die Frage aufgeworfen, ob wir unsere etablierten sozial konstruierten Verantwortungsbeziehungen in diesen Fällen nochmals überdenken sollten.

Schlüsselwörter: moralische Verantwortung, soziale Mensch-Maschine-Interaktion, neuer Typus eines sozialen Agenten, verteilte Verantwortung, moralisches Handeln

1. Einleitung

Es ist absehbar, dass schon in naher Zukunft ein nicht zu vernachlässigender Teil unseres sozialen Lebens auch Interaktionen mit künstlichen Systemen beinhalten wird. Das wird auf der einen Seite zu vielen neuartigen Formen des Werkzeuggebrauchs führen, so wie wir zum Beispiel schon jetzt technologische Entwicklungen wie Videotelefonie als Medium für soziale Interaktionen nutzen. Auf der anderen Seite ist es aber nicht auszuschließen, dass sich auch ein neuer Typus sozialer Interaktion etablieren wird, in denen künstliche Systeme nicht nur benutzt werden, sondern eine aktive Rolle in sozialen Interaktionen übernehmen. In einer sozialen Mensch-Maschine-Interaktion würden sich dann ungleiche Interaktionspartner*innen, nämlich lebende und nicht lebende Agenten, gegenüberstehen. Da solche Interaktionen nicht auf bloßen Werkzeuggebrauch reduzierbar sind, sie aber auch nicht alle Bedingungen erfüllen, die wir mit zwischenmenschlichen Interaktionen verbinden, stellt sich die Frage, wie wir unsere Begrifflichkeiten erweitern können, um solche Weder-noch-Fälle zu erfassen.
Auf den ersten Blick mag die Vorstellung, dass künstliche Systeme in den Bereich sozialer Interaktionen eindringen könnten, insbesondere für ein westliches Publikum radikal und absurd klingen. Es scheint ein abwegiges Unternehmen zu sein, das Verständnis von Sozialität auf nicht-lebendige Entitäten auszuweiten. Bevor man jedoch die Möglichkeit von sozialen Mensch-Maschine-Interaktionen prinzipiell ausschließt, sollte man bedenken, dass außerhalb des westlichen Kulturkreises, etwa im Shintoismus und Animismus, auch Objekte als belebt bezeichnet werden, die aus westlicher Sicht als

unbelebt gelten. Außerdem kann man anführen, dass manche Mensch-Maschine-Interaktionen schon jetzt mehr sozialen Mensch-Mensch-Interaktionen ähneln, als dass sie an den Gebrauch von Werkzeugen erinnern. Eine ‚Unterhaltung' mit einem Chatbot würden wohl die Wenigsten als einen Fall von Werkzeuggebrauch beschreiben. Desweiteren hat die Annahme, dass Mensch-Maschine-Interaktionen mit Mensch-Mensch-Interaktionen vergleichbar sind, zumindest schon Eingang in die empirische Forschung gefunden.[1] In vielen Untersuchungen werden experimentelle Studien mit künstlichen Agenten durchgeführt, um Einblicke in menschliche sozio-kognitive Mechanismen zu gewinnen.[2] Gäbe es hier keine belastbaren Ähnlichkeiten, könnten solche Experimente nicht zur Erforschung menschlicher Eigenschaften beitragen. Nicht zuletzt besteht kein Zweifel daran, dass Menschen emotionale Bindungen zu unbelebten Objekten aufbauen können und dass unsere Tendenz zum Anthropomorphisieren schon jetzt zu Fällen führt, in denen wir künstliche Agenten so behandeln, als wären sie soziale Agenten.

In diesem Aufsatz werde ich in einem ersten Schritt mögliche Bedingungen untersuchen, die dafürsprechen, bestimmte Mensch-Maschine-Interaktionen als soziale Interaktionen auszuzeichnen. Dabei werde ich aufzeigen, unter welchen Bedingungen es gerechtfertigt ist, künstliche Agenten als einen neuen Typus sozialer Interaktionspartner*innen zu betrachten. Solche Fälle zeichnen sich dadurch aus, dass hier die künstlichen Agenten in einer relevanten Weise einen Beitrag zu der sozialen Interaktion leisten und eben nicht bloß benutzt werden. So ist es beispielsweise denkbar, dass künstliche Agenten mit ihren sozio-kognitiven Fähigkeiten einen wechselseitigen Austausch sozialer Informationen ermöglichen.

Daran anschließend werde ich die Frage behandeln, ob es darüber hinaus Gründe gibt, künstlichen Systemen, die sich als ein neuer Typus sozialer Interaktionspartner*innen qualifizieren, moralische Verantwortung zuzuschreiben. Unter der Voraussetzung, dass die Zuschreibung einer sozialen Akteurschaft (*social agency*) nicht zwangsläufig mit einer Zuschreibung moralischer Handlungsfähigkeit (*moral agency*) einhergeht, werden hier Fälle diskutiert, in denen es zumindest erwägenswert wäre, von verteilter Verantwortung zu sprechen. Es ist klar, dass nicht alle sozialen Akteure auch moralische Akteure sind, denn ein sozialer Akteur zu sein, ist zwar eine notwendige, aber keine hinreichende Bedingung, um sich als moralischer Akteur zu qualifizieren. Aber selbst wenn man diesem neuen Typus eine moralische Handlungsfähigkeit zuschreibt, können sozial konstruierte Verantwortungsbeziehungen festlegen, dass eben dieser neue Typus keine Verantwortung trägt. Es stellen sich also mehrere Fragen:

1. Kann man manchen künstlichen Systemen eine Form von Handlungsfähigkeit sowie die notwendigen sozio-kognitiven Fähigkeiten zuschreiben, so dass sie sich als ein neuer Typus sozialer Interaktionspartner*innen qualifizieren?
2. Kann man ihnen darüber hinaus auch eine Form moralischer Handlungsfähigkeit zuschreiben?
3. Sollte sich unsere Gesellschaft dafür entscheiden, dass ein Teil der Verantwortung in sozialen Mensch-Maschinen-Interaktionen auch künstlichen Agenten zugeschrieben werden kann?

Vorausgesetzt, dass sich bestimmte künstliche Systeme als ein neuer Typus eines sozialen Agenten erweisen, die sich darüber hinaus auch als moralische Agenten qualifizieren, gilt es, mögliche Kriterien zu untersuchen, die klären, wie Verantwortung zwischen solch ungleichen Interaktionspartner*innen verteilt werden kann oder ob die Verantwortung auf der menschlichen Seite verbleibt.

Um dieser Frage nachzugehen, untersuche ich, inwieweit unsere Praxis der Verantwortungsverteilung in Mensch-Mensch-Interaktionen eine Strategie für die Verteilung von Verantwortung in sozialen Mensch-Maschine-Interaktionen bieten kann. Dabei werde ich zeigen, dass es Gründe gibt, die bis zum heutigen Tage weit verbreitete Intuition, menschlichen Interaktionspartner*innen in allen Mensch-Maschine-Interaktionen prinzipiell die Gesamtverantwortung zuzuschreiben, kritisch hinterfragt werden kann. Trotz der Tatsache, dass menschliche moralische Akteure sowohl Schöpfer der künstlichen Systeme als auch Initiatoren jeglicher Mensch-Maschine-Interaktionen sind, stellt sich aus meiner Sicht die Frage, ob es Fälle gibt, in denen es gerechtfertigt ist, künstlichen Systemen einen Anteil an Verantwortung zuzuweisen. Damit liefert dieser Beitrag eine Alternative zu den extremen

[1] Vgl. Hortensius & Cross 2018.

[2] Siehe Wykowska et al 2016.

Positionen, die die bloße Möglichkeit bestreiten, künstlichen Systemen Verantwortung zuzuschreiben[3] und bietet eine Grundlage für eine Auseinandersetzung mit anderen Ansätzen, die Vorschläge unterbreitet haben, unter welchen Umständen sich künstliche Systeme als moralische Agenten qualifiziert können.[4] In Anerkennung der Tatsache, dass es entscheidende Unterschiede zwischen lebenden Menschen und nicht lebenden künstlichen Systemen gibt, die in einem asymmetrischen Merkmal jeder Mensch-Maschine-Interaktion kulminieren, untersucht dieser Beitrag inwiefern sozio-kognitive Fähigkeiten künstlicher Systeme ein Argument für verteilte Verantwortung in Bezug auf soziale Mensch-Maschine-Interaktionen darstellen können.

Um die Praxis von Verantwortungszuschreibungen zu erfassen, führe ich eine Unterscheidung zwischen interaktionsbezogenen Kriterien und den Kriterien ein, die sich aus sozial konstruierten Verantwortungsbeziehungen ableiten lassen. Beide Arten von Kriterien werden als Rechtfertigung für das Zuschreiben moralischer Verantwortung verwendet. Interaktionsbezogene Kriterien beziehen sich auf die Manifestationen von Fähigkeiten, über die potenzielle soziale und moralische Interaktionspartner*innen verfügen müssen, um sich als solche zu qualifizieren. Denn diese Fähigkeiten sind eine entscheidende Voraussetzung für jede soziale Interaktion. Nur wenn beide Interaktionspartner*innen mit ihren Fähigkeiten das Ergebnis einer Interaktion beeinflussen, sind wir geneigt, von einer sozialen Interaktion zu sprechen. Die bloße Fähigkeit zu handeln ist zwar notwendig, aber nicht hinreichend. Die Qualifikation als soziale Interaktionspartner*innen setzt sozio-kognitive Fähigkeiten voraus, die es zum Beispiel ermöglichen, das Verhalten des Gegenübers in sozialen Interaktionen zu antizipieren, zu kontrollieren und entsprechende Pläne zu entwickeln. So sind etwa zwei Arten der Antizipationsfähigkeit hilfreich, um Teil einer erfolgreichen sozialen Interaktion zu sein; zum einen die Fähigkeit, Handlungen der Interaktionspartner*innen zu antizipieren (*mindreading*) und zum anderen die Fähigkeit, die Folgen einer Interaktion vorherzusehen.

Auf den ersten Blick scheinen unsere sozial konstruierten Verantwortungszuschreibungen zur Bewertung von Interaktionen darauf zu basieren, dass die Gewichtung von Verantwortung in Abhängigkeit von der angenommenen Expertise (im Sinne des Ausmaßes der sozio-kognitiven Fähigkeiten) der beteiligten Interaktionspartner*innen vorgenommen wird. Je ausgeprägter die Expertise, je klarer die interaktionsbezogenen Kriterien erfüllt sind, desto höher scheint der Anteil an Verantwortung zu sein. Beispiele, die diesen Zusammenhang illustrieren, sind Verantwortungsbeziehungen zwischen Erwachsenen (Eltern) und Kindern, Vorgesetzten und Mitarbeiter*innen oder auch Lehrer*innen und Schüler*innen. Die Bewertung einiger dieser Verantwortungsbeziehungen finden sich auch in der Rechtsprechung wieder.

Im Allgemeinen spiegeln die aus sozial konstruierten Verantwortungsbeziehungen abgeleiteten Kriterien die Bewertungen interaktionsbezogener Kriterien wider. Es gibt jedoch auch Fälle, in denen sozial konstruierte Verantwortungsbeziehungen für eine andere Verteilung der Verantwortung sprechen als eine Einzelfallbeurteilung, die man basierend auf interaktionsbezogenen Kriterien vornehmen könnte. Zum Beispiel ist es denkbar, dass ein Kind in einer bestimmten Interaktion verantwortungsbezogene Fähigkeiten zeigt, über die der Erwachsene in diesem Moment nicht verfügt; in solchen Fällen scheint es nichtsdestotrotz vernünftig, Kriterien, die aus sozial konstruierten Verantwortungsbeziehungen abgeleitet werden, ein größeres Gewicht zuzuschreiben.

2. Verantwortung auf der menschlichen Seite

Unsere Praxis der Zuschreibung moralischer Verantwortung bei Menschen ist eng verbunden mit Fragen nach den notwendigen und hinreichenden Bedingungen, mit denen sich handlungsfähige, soziale Akteure als moralische Akteure qualifizieren. Die weitverbreitete Vorstellung vollwertiger moralischer Handlungsfähigkeit (*full-fledged moral agency*) geht mit anspruchsvollen Bedingungen einher: Ein moralischer Akteur muss über Bewusstsein, Autonomie, Intentionalität, freien Willen und Reflexionsfähigkeit verfügen, um eine sogenannte Glaubens-Wunsch-Intentions-Architektur (*belief-desire-intention architecture*)[5] zu realisieren.

[3] Hier sind beispielsweise Nida-Rümelin & Weidenfeld 2018 und Bryson 2010 zu nennen.

[4] Differenzierte Positionen zu Fragen moralischer Handlungsfähigkeit findet man bei Floridi & Sanders 2004; Moor 2006; Misselhorn 2018; Wallach & Allen 2012, 2009 und Verbeek 2006.

[5] Bratman 2014.

Auch wenn wir Menschen als moralisch verantwortlich betrachten, weil sie über bewusste mentale und emotionale Zustände, Intentionalität, Intelligenz, die Fähigkeit zu denken, zu planen, zu urteilen und anders zu handeln (freier Wille) verfügen, können wir das Ausmaß der Verantwortung relativieren, wenn einige dieser Fähigkeiten beeinträchtigt oder noch nicht voll entwickelt sind. An dieser Stelle kommen Konzepte wie die verminderte Schuldfähigkeit ins Spiel. Bei der Erörterung der komplexen Faktoren, die sich auf die Zuweisung von Verantwortung in einem juristischen Sinne und deren Verteilung auswirken, wird deutlich, dass moralische Handlungsfähigkeit sicherlich eine notwendige, aber wahrscheinlich keine hinreichende Bedingung ist. Man kann hier auf Debatten über verminderte Schuldfähigkeit verweisen, in denen die Rolle von Faktoren wie dem Fehlen eines Vorsatzes, eine Einschränkung der allgemeinen Fähigkeit zur Empathie, mangelnde Selbstkontrolle oder verminderte Impulskontrolle thematisiert werden.[6]

Ein interessanter Sonderfall in den Debatten über Verantwortung ist die Zuschreibung von verteilter Verantwortung. Es gibt diverse Mensch-Mensch-Interaktionen, in denen einzelne Akteur*innen nicht allein verantwortlich gemacht wird, da man davon ausgeht, dass hier die Verantwortung mit den Interaktionspartner*innen geteilt wird. Interessanterweise ist der Umfang der Verantwortung zwischen zwei Interaktionspartner*innen nicht immer gleich; wir argumentieren beispielsweise für eine ungleiche Verteilung der Verantwortung in Erwachsenen-Kind-Interaktionen.

Obwohl moralische Handlungsfähigkeit eine notwendige Voraussetzung für die Zuschreibung von Verantwortung darstellt, bleibt der Begriff der Verantwortung selbst ein äußerst komplexes, vielschichtiges und auch umstrittenes Konzept.[7] Insbesondere dann, wenn es um die Frage geht, wie viel Verantwortung einem Akteur zugerechnet werden soll, wenn z.B. mehrere Akteure an einer Handlung beteiligt sind.[8] Es ist oft unklar, welche Kriterien letztendlich eine entscheidende Rolle spielen oder wie eine klare Strategie zur Abwägung möglicher gegensätzlicher Kriterien aussehen könnte. Darüber hinaus ist im Falle einer reduzierten Verantwortungszuschreibung nicht klar, wer der Adressat für die "übrig gebliebene" Verantwortung ist.[9] Solche Verantwortungslücken entstehen beispielsweise, wenn technologische Innovationen – autonome Maschinen, lernende Algorithmen und soziale Roboter – eine wesentliche Rolle in (sozialen) Mensch-Maschine-Interaktionen spielen. Nach der gängigen Praxis werden künstliche Agenten hier als bloße Werkzeuge gedeutet, denen man prinzipiell keine Verantwortung zuschreiben kann, und so eröffnen sie das Feld für Debatten über Verantwortungslücken. In diesem Aufsatz konzentriere ich mich auf verteilte Verantwortung in sozialen Interaktionen und werde einen Vorschlag machen, wie man bei der Konzeptualisierung von Verantwortung auch ein Netzwerk von Menschen und Maschinen berücksichtigen kann.[10]

Selbst wenn man zu dem Schluss kommt, dass es Fälle von verteilter Verantwortung in speziellen sozialen Mensch-Maschine-Interaktionen gibt, ist damit nicht von vornherein klar, ob sich Verantwortung wie eine Pizza aufteilen lässt, d.h. dass man von der Anzahl der Verantwortungsträger auf die Größe des Anteils schließen kann.[11]

Analysiert man unsere Praxis der Verantwortungszuweisung in Mensch-Mensch-Interaktionen, indem man zwischen interaktionsbezogenen Kriterien und Kriterien, die sich aus sozial konstruierten Verantwortungsbeziehungen ergeben, unterscheidet, geben die interaktionsbezogenen Kriterien Auskunft darüber, inwieweit Interaktionspartner*innen aufgrund ihrer Fähigkeiten einen Einfluss auf das Ergebnis der Interaktion haben. Wenn ein Interaktionspartner beispielsweise die Folgen einer Interaktion nicht überblicken kann oder sich in seinen Annahmen über das zukünftige Verhalten einer Interaktionspartnerin irrt, hat er einen verminderten Einfluss. Das Ausmaß, in dem die hier relevanten Fähigkeiten entwickelt sind, lässt einen Rückschluss auf den Grad der Einflussnahme zu. Verfügen die Interaktionspartner*innen über vergleichbar ausgeprägte Fähigkeiten, liegt es nahe, für eine gleichmäßige Aufteilung der Verantwortung zu plädieren. Man kann jedoch auch dafür argumentieren,

[6] Vincent 2010.

[7] Shoemaker 2011; Scanlon 2008.

[8] Die Problematik der Involvierung mehrerer Agenten wird z.B. unter dem Schlagwort ‚*Many-Hands-Problem*' diskutiert (Van de Poel et al. 2012).

[9] Matthias 2004.

[10] Vgl. auch Gunkel 2020.

[11] Brown Coverdale et al. 2021.

dass beiden Akteuren die volle Verantwortung zukommt, oder man begreift die Beteiligten einer sozialen Interaktion als eine kollektive Person, als ein ‚Wir', dem die Verantwortung zukommt. Eine unterschiedliche Ausprägung der Einflussnahme spricht jedoch für eine ungleiche Verteilung der Verantwortung. Die Kriterien, die sich aus gesellschaftlich konstruierten Verantwortungsbeziehungen ableiten lassen, legen hier in vielen Fällen eine von vornherein bestimmte Verteilung von Verantwortung fest. Unabhängig von Einzelfallbeurteilungen, in denen interaktionsbezogene Kriterien evaluiert werden, können sozial konstruierte Verantwortungsbeziehungen festlegen, dass einer Klasse von Akteuren ein größerer Anteil oder sogar die volle Verantwortung zukommt. Jedoch berücksichtigen die sozial konstruierten Verantwortungsbeziehungen in der Regel eine vorangegangene Bewertung der interaktionsbezogenen Kriterien.

3. Verantwortung auf Seiten der künstlichen Akteure

Die Spezifizierungen der Bedingungen für moralisches Handeln spielen in den Debatten über Verantwortung eine entscheidende Rolle, auch wenn die Qualifikation als moralischer Akteur nicht immer ein ausreichender Grund ist, um die Zuschreibung moralischer Verantwortung mit nachfolgenden Sanktionen zu rechtfertigen. Die Zuschreibungsbedingungen des Begriffes einer vollwertigen moralischen Handlungsfähigkeit schließen künstliche Systeme von vornherein aus. Wenn man davon ausgeht, dass die Erfüllung der Kriterien für eine vollwertige moralische Handlungsfähigkeit, wie z.B. Bewusstsein, eine notwendige Voraussetzung für Verantwortung ist, dann können künstliche Systeme nicht verantwortlich (gemacht) werden.[12] Um überhaupt moralisch verantwortliche künstliche Agenten in Erwägung zu ziehen, braucht man eine Alternative zu dieser anspruchsvollen Vorstellung.

An diesem Punkt kommen Vorschläge ins Spiel, die eine graduelle Konzeption des Begriffs der moralischen Handlungsfähigkeit vorschlagen. Solche graduellen Konzeptionen zeichnen sich dadurch aus, dass sie die Notwendigkeit einiger spezifischer Bedingungen vollwertiger moralischer Handlungsfähigkeit in Frage stellen, indem sie davon ausgehen, dass es verschiedene Möglichkeiten gibt, wie moralische Handlungsfähigkeit realisiert werden kann. Indem man multiple Realisierungen annimmt, kann man beispielsweise dafür argumentieren, dass moralische Handlungsfähigkeit sogar unbewussten, nicht lebenden Akteuren zukommt. Die Idee ist hier, dass man nicht ausschließen kann, dass ein und dieselbe Fähigkeit auf unterschiedliche Art und Weise realisiert werden kann. Berücksichtigt man eine Variabilität in der Ausprägung von Bedingungen wie Autonomie und Sensibilität für moralische Werte, kann man nach Wallach & Allen operative moralische Agenten von schwach funktionalen und diese wiederum von vollwertigen funktionalen moralischen Agenten unterscheiden.[13] Alternativ könnte man Moor folgen, der mehrere Arten von moralischen Agenten herausgearbeitet hat, indem er Agenten mit ethischem Einfluss von impliziten ethischen Agenten und expliziten ethischen Agenten unterscheidet. Explizite ethische Agenten entsprechen vollwertigen moralischen Akteuren mit Bewusstsein, Intentionalität und freiem Willen.[14] Nach Floridi & Sanders kann künstlichen Systemen eine moralische Verantwortung zugeschrieben werden, wenn sie Bedingungen hinsichtlich Interaktivität, Autonomie und Anpassungsfähigkeit erfüllen.[15]

Die Diskussion über die verschiedenen Strategien, wie man dafür argumentieren kann, dass auch künstlichen Systemen eine moralische Handlungsfähigkeit zugeschrieben werden kann, werde ich in diesem Beitrag nicht im Detail behandeln. Im Fokus steht hier die Frage, ob soziale Mensch-Maschine-Interaktionen zu einer Form von verteilter Verantwortung führen können. Um diese Frage behandeln zu können, setze ich voraus, dass einige künstliche Agenten sich als eine neue Art von sozialen Interaktionspartner*innen qualifizieren können und dass ein gradueller Begriff moralischer Handlungsfähigkeit auf einige dieser Agenten anwendbar ist. Daher ist die Annahme, dass ein gradueller Begriff von moralischem Handeln auf bestimmte künstliche Akteure anwendbar ist, eine notwendige Voraussetzung für die folgenden Überlegungen.

[12] Siehe auch Coeckelbergh 2020 und Himma 2009.
[13] Wallach & Allen 2009; 2012.
[14] Moor 2006.
[15] Floridi & Sanders 2004.

Die Möglichkeit verteilter Verantwortung in sozialen Mensch-Maschine-Interaktionen basiert auf zwei voraussetzungsstarken Argumenten:

1. *Argument für die Möglichkeit sozialer künstlicher Interaktionspartner*innen*
   P (1): Es gibt Mensch-Maschine-Interaktionen, die sich nicht auf den bloßen Werkzeuggebrauch reduzieren lassen.
   P (2): Es gibt multiple Realisierung sozialer Handlungsfähigkeit.[16]
   C (1): Dies spricht für eine Erweiterung des Begriffs eines sozialen Agenten.
   C (2): Wenn künstliche soziale Agenten in Mensch-Maschine-Interaktionen involviert sind, dann kann man diese als einen neuen Typ sozialer Interaktion betrachten.

Die Behauptung, dass sich bestimmte künstliche Agenten als ein neuer Typ sozialer Agenten qualifizieren können, bedeutet jedoch noch nicht, dass diese Agenten automatisch als moralische Agenten zu betrachten sind. So können beispielsweise Tiere als soziale Agenten betrachtet werden, ohne dass diesen damit eine moralische Handlungsfähigkeit zugeschrieben wird.

2. *Argument für moralische Handlungsfähigkeit sozialer künstlicher Interaktionspartner*innen*
   P (1): Es gibt künstliche Systeme, die sich als soziale Agenten qualifizieren.
   P (2): Auch für moralische Handlungsfähigkeit gilt, dass es multiple Realisierung gibt (graduelle Konzeption / minimale moralische Handlungsfähigkeit).
   C (1): Unter bestimmten Umständen kann einem künstlichen System moralische Handlungsfähigkeit zugeschrieben werden.

Und selbst die Zuschreibung moralischer Handlungsfähigkeit klärt die Frage nach der Zuschreibung von Verantwortung, vor allem wenn dies mit Sanktionen einhergehen soll, noch nicht abschließend. So kann es beispielsweise soziale Akteure geben, denen wir moralische Handlungsfähigkeit zuschreiben, denen wir aber dennoch unter bestimmten Umständen Schuldfähigkeit absprechen. Ziel dieses Beitrags ist es, zu untersuchen, unter welchen Umständen es gerechtfertigt ist, diesem neuen Typus von sozialen Akteuren einen Teil der Verantwortung in sozialen Mensch-Maschine-Interaktionen zuzuschreiben.

4. Verteilte Verantwortung in sozialen Mensch-Maschine-Interaktionen

Nimmt man an, dass bestimmten künstlichen Agenten sowohl soziale als auch moralische Handlungsfähigkeit zugeschrieben werden kann, stellt sich die weitergehende Frage, wie man über die Verteilung von Verantwortung urteilen soll, wenn aus einer sozialen Mensch-Maschine-Interaktion moralisch fragwürdige Konsequenzen resultieren. In Anlehnung an unsere Praxis bei der Zuschreibung von verteilter Verantwortung in Mensch-Mensch-Interaktionen werde ich hier sowohl interaktionsbezogene Kriterien, die auf der Manifestation sozio-kognitiver Fähigkeiten aufbauen und soziale Interaktionen ermöglichen, als auch die Kriterien, die sich aus sozial konstruierten Verantwortungsbeziehungen ergeben, untersuchen. Letztere legen fest, inwieweit dem Menschen als Hersteller künstlicher Systeme und als Initiator sozialer Interaktionen prinzipiell ein gewichtigerer Anteil oder sogar die volle Verantwortung in Mensch-Maschine-Interaktionen zukommt, selbst wenn sich diese als soziale Mensch-Maschine-Interaktionen qualifizieren.

4.1 interaktionsbezogene Kriterien

Aufgrund der Annahme multipler Realisation sozio-kognitiver Fähigkeiten haben wir es bei sozialen Mensch-Maschine-Interaktionen mit einer asymmetrischen Verteilung von Bedingungen zu tun. Man könnte nun geneigt sein zu sagen, dass Menschen künstlichen Systemen in Bezug auf die interaktionsbezogenen Kriterien grundsätzlich überlegen seien, da künstliche Systeme bestimmte

[16] Hier wird der Standpunkt vertreten, dass bestimmte künstliche Systeme als soziale Agenten gelten können, wenn sie sowohl eine Art von Handlungsfähigkeit als auch eine Form von sozialer Kompetenz besitzen, so dass sie sowohl zu einem Austausch sozialer Informationen beitragen als auch einen Einfluss auf das Ergebnis einer sozialen Interaktion haben können (Strasser 2020).

Bedingungen nicht erfüllen. Künstliche Agenten sind nicht lebendig, sie haben weder Bewusstsein noch Emotionen, noch sind sie leidensfähig. Dies alles sind Punkte, die eine berechtigte Skepsis gegenüber der Zuschreibung von Verantwortung nähren,[17] selbst wenn man zubilligt, dass ihnen der Status eines neuen Typus sozialer Agenten zugestanden werden kann. Folgt man diesem Gedankengang, so scheint die Annahme eines neuen Typus sozialer Agenten, der auf multiple Realisation basiert, dazu zu führen, dass die Bewertung interaktionsbezogener Kriterien in sozialen Mensch-Maschine-Interaktionen von vorneherein zu Gunsten der menschlichen Seite ausfällt. Künstlichen Agenten würden dann wesentliche Bedingungen fehlen. Denn wenn man die These vertritt, dass die multiple Realisation keine gleichwertige Form von Handlungsfähigkeit und sozio-kognitiven Fähigkeiten darstellt, dann fallen Bedingungen, die für Menschen notwendig sind und von künstlichen Systemen nicht verlangt werden, auf eine relevantere Art und Weise ins Gewicht.

Ich möchte hier jedoch dafür argumentieren, dass eine andere Form der Realisation als gleichwertig zu betrachten ist, da sie ja zu der Zuschreibung derselben Fähigkeit führt. Auch lässt sich die vermutete Überlegenheit des Menschen in Frage stellen, denn bei genauerer Betrachtung zeigt sich, dass es auch Fähigkeiten gibt, in denen künstliche Agenten den Menschen übertreffen. Sie können zum Beispiel in kürzerer Zeit eine größere Menge an Daten verarbeiten und speichern. Dies kann entscheidende Konsequenzen haben, wenn beurteilt werden soll, inwiefern sie den Ausgang einer Interaktion beeinflussen. Menschliche Reaktionszeiten können einfach zu langsam sein, um wirksam einzugreifen. Hinzu kommt, dass die neuesten technologischen Entwicklungen künstlicher Agenten dazu führen, dass Menschen nicht mehr im Detail verstehen, wie diese künstlichen Systeme funktionieren. Der Mensch ist zwar in der Lage, künstliche Systeme zu konstruieren, sieht sich aber mit dem so genannten Black-Box-Problem konfrontiert und kann oft die internen Prozesse nicht mehr verstehen. So ist beispielsweise nicht ersichtlich, nach welchen Kriterien ein trainiertes neuronales Netz "entscheidet", einen gegebenen Input einer bestimmten Kategorie zuzuordnen. Daher kann der Mensch oft nicht vorhersagen, wie sich künstliche Agenten verhalten werden. Hier stellt sich daher die Frage, inwieweit unsere begrenzte Fähigkeit, das Verhalten künstlicher Agenten zu antizipieren, uns davon entbinden kann, einen größeren Teil der Verantwortung zu übernehmen.

4.2 Interaktionen zwischen Kindern und Erwachsenen

Eine Betrachtung unserer Praxis bezüglich verteilter Verantwortung im Falle von Interaktionen zwischen Kindern und Erwachsenen kann Licht darauf werfen, wie wir mit Verantwortungszuschreibung in asymmetrischen Interaktionen verfahren. Denn in solchen Interaktionen stehen sich auch ungleiche Interaktionspartner*innen gegenüber, da Kinder entwicklungsbedingt noch nicht über eine vollwertige soziale und moralische Handlungsfähigkeit (*full-fledged social and moral agency*) verfügen. Da hier in gewisser Hinsicht ebenso eine asymmetrische Verteilung von Fähigkeiten vorliegt, können solche Interaktionen ein Ausgangspunkt für die Diskussion bezüglich der Verteilung von Verantwortung in sozialen Mensch-Maschine-Interaktionen bieten.

Nach unseren sozial konstruierten Verantwortungszuschreibungen kommt den Erwachsenen in Kind-Erwachsenen-Interaktionen die volle Verantwortung zu. Eine Analyse der interaktionsbezogenen Kriterien liefert hier motivationale Gründe für unsere Praxis. Da es naheliegend ist, den Erwachsenen weiterentwickelte Fähigkeiten zuzuschreiben, wird hier die Asymmetrie in der Ausprägung der sozio-kognitiven Fähigkeiten als eine Begründung für eine ungleiche Verteilung der Verantwortung benützt. Man könnte sagen, dass Erwachsene in Bezug auf viele Aspekte sozialer Interaktionen als Experten gelten, während Kinder als Novizen betrachtet werden. Die Zuschreibung von Expertise (im Sinne von weiterentwickelten sozio-kognitiven Fähigkeiten) in sozialen Interaktionen geht mit mehreren Merkmalen einher: Erwachsene können zum Beispiel aus einem breiteren Spektrum von möglichen Handlungen auswählen, sie sind häufiger in der Lage einzugreifen und können sowohl das zukünftige Verhalten ihrer Interaktionspartner*innen als auch die Folgen von Handlungen besser antizipieren. Erwachsene haben mehr Kontrolle über das Ergebnis einer Interaktion, da sie die Umstände besser einschätzen können und folglich auch mehr über die Folgen von Handlungen wissen. All diese Merkmale gelten als Gründe dafür, dass Erwachsene stärker in die Verantwortung genommen werden. Anhand der interaktionsbezogenen Kriterien, die hier die Entwicklung der Fähigkeiten zur moralischen

[17] Véliz 2021.

Verantwortung in Interaktionen mitspezifizieren, lässt sich eine ungleiche Verteilung der Verantwortung zwischen Kindern und Erwachsenen begründen. Je ausgeprägter die entsprechenden Fähigkeiten eines Gegenübers sind, desto mehr Verantwortung wird diesem Gegenüber zugeschrieben.

5. Verteilte Verantwortung in Mensch-Maschine-Interaktionen

Auch in einer sozialen Mensch-Maschine-Interaktion könnte man auf den ersten Blick geneigt sein, künstlichen Agenten ebenso wie Kindern weniger Verantwortung zuzuschreiben, da die Handlungsfähigkeit und die sozio-kognitiven Fähigkeiten künstlicher Systeme auch auf einer multiplen Realisierung basieren, die aber scheinbar weniger anspruchsvolle Bedingungen fordert.[18] Wenn man jedoch die interaktionsbezogenen Kriterien genauer betrachtet, wird es fraglich, ob eine analoge Rechtfertigung für verteilte Verantwortung auch auf soziale Mensch-Maschine-Interaktionen angewendet werden kann.

Im Folgenden möchte ich zeigen, dass auf Grundlage einer Analyse der interaktionsbezogenen Kriterien künstlichen Agenten nicht notwendigerweise dieselbe Rolle wie Kindern zugewiesen werden kann. Dies liegt daran, dass wir nicht davon ausgehen können, dass menschliche Agenten künstlichen Agenten generell überlegen sind, was die für moralische Verantwortung in sozialen Interaktionen relevanten Fähigkeiten angeht. Die bestehende Asymmetrie sollte anders bewertet werden, da es Aspekte gibt, in denen künstliche Systeme dem Menschen überlegen sind. Wie oben erwähnt, sind künstliche Systeme in der Lage, eine größere Menge an Daten in kürzerer Zeit zu verarbeiten und zu speichern, und dies führt zu einer größeren Kontrolle über das Ergebnis einer sozialen Mensch-Maschine-Interaktion. Fokussiert man allein auf diesen Aspekt, würde die Rolle künstlicher Systeme in dieser Hinsicht eher der Rolle von Erwachsenen entsprechen. Auch bei der Bewertung von Antizipationsfähigkeiten ist es vorstellbar, dass es eher die menschlichen Interaktionspartner*innen sind, denen eine verminderte Antizipationsfähigkeit zuzuschreiben ist. Denn Menschen sind oft nicht in der Lage, das Verhalten künstlicher Systeme und damit auch das Ergebnis der Interaktion mit ihnen zu antizipieren. Daraus folgt jedoch noch nicht, dass künstlichen Agenten mehr Verantwortung als ihren menschlichen Interaktionspartner*innen zukommt. Denn eine weitere Analyse der interaktionsbezogenen Kriterien kann zeigen, dass die multiple Realisation durch künstliche Systeme auch als ein Manko verstanden werden kann. Es gibt Bedingungen, wie Bewusstsein oder Emotionalität, die von den künstlichen Systemen nicht gefordert werden. Man könnte, wie oben beschrieben, nun argumentieren, dass künstliche Systeme unterm Strich weniger anspruchsvolle Bedingungen erfüllen und dies als Rechtfertigung dafür nehmen, künstlichen Agenten weniger oder keine Verantwortung zuzuschreiben. Ich argumentiere jedoch, dass die weniger anspruchsvollen Bedingungen eines graduellen Verständnisses moralischen Handelns nicht der einzige entscheidende Faktor sein können, da wir eben auch Kriterien finden, bei denen künstliche Agenten menschlichen Interaktionspartner*innen überlegen sind. Insofern stellt die Gemengelage der Ausprägung der interaktionsbezogenen Kriterien in sozialen Mensch-Maschine-Interaktionen uns vor die Schwierigkeit, dass wir letztlich abwägen müssen, ob die weniger anspruchsvollen Bedingungen für künstliche Agenten schwerer wiegen sollten als andere Bedingungen, in denen künstliche Agenten dem Menschen überlegen sind.

Ein schwerwiegender Einwand gegen die Zuschreibung von Verantwortung bezüglich künstlicher Akteure resultiert jedoch aus Kriterien, die wir aus unseren derzeitigen sozial konstruierten Verantwortungsbeziehungen ableiten können. Hier scheint es *Common Sense* zu sein, dass nur Menschen als Konstrukteure und als Nutzer künstlicher Systeme moralische Verantwortung für das Ergebnis von Interaktionen haben. Daraus könnte man schließen, dass wir als Konstrukteure von künstlichen Systemen und als Initiatoren von Mensch-Maschine-Interaktionen selbst im Falle von sozialen Mensch-Maschine-Interaktionen die einzigen Adressaten für Verantwortungszuschreibungen wären. Ob es angemessen ist, unsere gesellschaftlich konstruierten Verantwortungsbeziehungen auf den Sonderfall der sozialen Mensch-Maschine-Interaktion anzuwenden, kann man aber in Frage stellen. Denn unsere etablierten Ideen zu sozial konstruierten Verantwortungsbeziehungen beruhen

[18] Vergleiche hierzu Nyholm 2018.

auf der Vorstellung, dass alle Mensch-Maschine-Interaktionen als Fälle von Werkzeuggebrauch behandelt werden können.
Bei der Verwendung von Werkzeugen ist es unstrittig, dass Benutzer*innen (oder auch Hersteller*innen) die Hauptadressaten für jegliche Verantwortung sind. Auch wenn Menschen bei Werkzeugbenutzung nicht immer vollumfänglich verantwortlich gemacht werden, ist es abwegig, Werkzeugen überhaupt Verantwortung in einem moralischen Sinne zuzuschreiben. Um negative Folgen von Werkzeuggebrauch zu vermeiden, verlangt unsere Gesellschaft den Nachweis von Kenntnissen, damit Menschen bestimmte Werkzeuge benutzen dürfen. So braucht man zum Beispiel einen Führerschein, um ein Auto zu fahren. Sowohl äußere als auch innere Gründe können eine eingeschränkte Zurechnung von Verantwortung rechtfertigen. Angenommen, eine Person kann nachteilige Umstände, die das Ergebnis einer Handlung wesentlich bestimmen, weder vorhersehen noch beeinflussen, würden sich diese abschwächend auf die Bewertung der Verantwortung auswirken. Wenn zum Beispiel unvorhersehbare Umweltfaktoren die menschlichen Eingriffsmöglichkeiten einschränken, führt dies zu einer geringeren Verantwortung des menschlichen Akteurs. Darüber hinaus gibt es Situationen, in denen bestimmten menschlichen Akteuren generell eine verminderte Schuldfähigkeit zugeschrieben wird. Nach der deutschen Rechtsprechung sind Personen schuldunfähig, wenn sie zum Zeitpunkt des Vollzuges einer Handlung nicht in der Lage sind, die Rechtswidrigkeit der Handlung zu erkennen oder aufgrund einer krankhaften seelischen Störung, einer tiefgreifenden Bewusstseinsstörung oder einer Intelligenzminderung oder einer anderen schwerwiegenden seelischen Störung nicht in der Lage sind, auf der Grundlage dieser Erkenntnis zu handeln.[19] Ein weiterer Fall betrifft technische Geräte, die trotz sachgerechter Handhabung versagen. Hier wird nicht der Benutzer verantwortlich gemacht, sondern der Hersteller der Geräte kann zur Verantwortung gezogen werden. Es gibt eine umfangreiche Rechtsprechung, die sich mit Haftungsfragen beschäftigt. Interessanterweise gibt es verschiedene Fälle von verminderter Verantwortung, in denen die Frage, wem oder was die verbleibende Restverantwortung zugerechnet werden soll, nicht abschließend geklärt werden kann. Solche Fälle führten in der Philosophie zu Debatten über Verantwortungslücken.[20]
Ich vertrete die Position, dass sich bei sozialen Mensch-Maschine-Interaktionen, die sich nicht auf den bloßen Werkzeuggebrauch reduzieren lassen, erneut die Frage stellt, wie mit verminderter Verantwortung auf der menschlichen Seite umzugehen ist. Es ist zumindest bedenkenswert, hier zu untersuchen, inwieweit eine verminderte Verantwortung auf menschlicher Seite es rechtfertigen könnte, künstlichen Systemen einen Anteil an Verantwortung zuzuschreiben und damit mögliche Verantwortungslücken zu schließen. Das heißt, ich werde im Folgenden kritisch hinterfragen, ob die gegenwärtig angewandten sozial konstruierten Verantwortungsbeziehungen, die Verantwortung ausschließlich Menschen zuschreiben, in Bezug auf soziale Mensch-Maschine-Interaktionen angemessen sind.
Bei der Beschreibung von Kind-Erwachsenen-Interaktionen hat sich gezeigt, dass die zugeschriebene Verantwortung des Erwachsenen durch die Kriterien, die sich aus sozial konstruierten Verantwortungsbeziehungen ableiten lassen, bereits vor jeder Interaktion höher ist – gleichzeitig spiegeln diese Kriterien aber auch interaktionsbezogene Kriterien wider. Im Gegensatz dazu berücksichtigen die derzeit angewandten sozial konstruierten Verantwortungsbeziehungen in Bezug auf Mensch-Maschine-Interaktionen nicht alle interaktionsbezogenen Kriterien solcher Interaktionen. Mit der Infragestellung etablierter sozial konstruierter Verantwortungsbeziehungen sollen keineswegs Gründe ignoriert werden, die dafürsprechen, dass Menschen dennoch dafür verantwortlich sind, gegenüber künstlichen Agenten wachsam und misstrauisch zu sein, bevor sie mit ihnen interagieren.[21] Selbst wenn man, wie in diesem Aufsatz vorgeschlagen wird, aufgrund einer Neubewertung der interaktionsbezogenen Kriterien für verteilte Verantwortung in sozialen Mensch-Maschine-

[19] Siehe dazu § 20 Schuldunfähigkeit wegen seelischer Störungen: „Ohne Schuld handelt, wer bei Begehung der Tat wegen einer krankhaften seelischen Störung, wegen einer tiefgreifenden Bewußtseinsstörung oder wegen einer Intelligenzminderung oder einer schweren anderen seelischen Störung unfähig ist, das Unrecht der Tat einzusehen oder nach dieser Einsicht zu handeln.“ (Strafgesetzbuch (StGB) § 20; https://www.gesetze-im-internet.de/stgb/__20.html)
[20] Matthias 2004.
[21] Deroy 2021; Hauswald 2021.

Interaktionen argumentiert, wird damit keinesfalls die moralische Verpflichtung zum Erwerb von Fachwissen im Umgang mit künstlichen Agenten für verantwortliche Interaktionen negiert. Überlegungen, die sich auf eine besondere Sorgfaltspflicht der menschlichen Interaktionspartner*innen beziehen, können in normative Beschränkungen für die Herstellung bestimmter Systeme, wie etwa Tötungsdrohnen, bei denen der Mensch eine eingeschränkte Eingriffsmöglichkeit hat, münden.[22]

Nichtsdestotrotz sollte kritisch reflektiert werden, dass die derzeitigen gesellschaftlich konstruierten Verantwortungsbeziehungen alle Mensch-Maschine-Interaktionen als Fälle von Werkzeuggebrauch behandeln. Daher vertrete ich die Auffassung, dass der Verweis auf etablierte Verantwortungsbeziehungen kein Freifahrtschein dafür sein sollte, künstliche Agenten a priori von jeglicher Verantwortung freizusprechen. Um zu verdeutlichen, wovon ich spreche, wenn ich die Frage aufwerfe, ob wir unsere sozial konstruierten Verantwortungsbeziehungen überdenken sollten, möchte ich ein fiktives Beispiel diskutieren.

Zu diesem Zwecke stelle man sich eine soziale Interaktion in einer Fahrstunde zwischen einer Fahrlehrerin und einem Fahrschüler vor. Hier interagieren zwei Menschen gemeinsam mit einem Werkzeug (dem Auto). Diese soziale Interaktion ist dadurch gekennzeichnet, dass die Fahrlehrerin in brenzlichen Situationen die Pflicht hat einzugreifen, denn der Fahrschüler kann in der Fahrstunde mit einer Situation konfrontiert werden, die seine Fähigkeiten übersteigt, während die Fahrlehrerin die Situation aufgrund ihrer Expertise bewältigen kann. Insofern ist eine Fahrstunde auch ein Fall einer asymmetrischen sozialen Interaktion und eine Analyse der interaktionsbezogenen Kriterien kann rechtfertigen, dass der Fahrlehrerin ein größerer Anteil der Verantwortung zukommt. Die weiterausgebildete Expertise motoviert hier in bestimmten Konstellationen eine ungleiche Verteilung der Verantwortung.

In einem zweiten Schritt stelle man sich nun vor, dass die Fahrlehrerin durch ein zukünftiges Fahrassistenzsystem ersetzt werden würde. Hier muss man natürlich anmerken, dass es bis zum heutigen Tage noch keine Fahrassistenzsysteme gibt, die sich als ein neuer Typus sozialer Interaktionspartner*innen qualifizieren. Aber angenommen zukünftige technologische Entwicklungen könnten zu solchen Fahrassistenzsystemen führen, würde ich vermuten, dass wir in einer solchen Konstellation auf der Basis interaktionsbezogener Kriterien zu einer ähnlichen Einschätzung wie oben kommen. Jedoch würde diese Einschätzung, nämlich dass ein Fahrassistenzsystem, das in einer brenzlichen Situation nicht eingreift, obwohl es die Möglichkeit dazu hat, einen Anteil an der Verantwortung eines möglicherweise negativen Ausganges dieser Interaktion zukommt, unserer etablierten Verantwortungszuschreibungspraxis widersprechen, die eine Zuschreibung von Verantwortung in Bezug auf künstliche Systeme ausschließt. Vorausgesetzt, dass unsere Gesellschaft in Zukunft mit einem neuen Typus von sozialen Agenten, wie er in diesem Aufsatz beschrieben wird, soziale Interaktionen hat, ist es zumindest denkbar, dass eine Neubewertung der interaktionsbezogenen Kriterien dazu führt, dass etablierte Verantwortungsbeziehungen in Bezug auf diese neue Art sozialer Mensch-Maschine-Interaktionen überdacht werden.

5.1. Mögliche Kriterien für Verantwortungszuschreibungen – Kontrolle

Soziale Interaktionen zeichnen sich dadurch aus, dass beide Interaktionspartner*innen gemeinsam zu dem Ausgang der Interaktion beitragen. Nimmt man gemeinsame Handlungen (*joint actions*) als ein paradigmatisches Beispiel für eine soziale Interaktion, dürfte es unumstritten sein, dass in gemeinsamen Handlungen allen Beteiligten Verantwortung zukommt. Interessant für die Überlegungen in diesem Aufsatz sind asymmetrische Interaktionen, in denen die Ausprägungen der interaktionsbezogenen Kriterien nicht gleich oder schlichtweg durch unterschiedliche Bedingungen erfüllt werden.

Wir haben gesehen, dass eine detaillierte Bewertung der interaktionsbezogenen Kriterien eine Begründung dafür liefern kann, von verteilter Verantwortung zu sprechen. So verfügen Erwachsene beispielsweise über weiterentwickelte Fähigkeiten in Bezug auf interaktionsbezogene Kriterien als Kinder. Sie können aus einer größeren Vielfalt von Handlungen wählen, häufiger eingreifen und das Verhalten der Interaktionspartner*innen vorhersehen. All dies führt zu einem höheren Maß an

[22] Loh 2019; Misselhorn 2018.

Kontrolle. Folglich ist das Ausmaß an Kontrolle ein wichtiger Faktor für die Zuschreibung von Verantwortung.
Bei der Analyse von Interaktionen mit autonomen Fahrzeugen kann die Kontrollierbarkeit unterschiedliche Ausprägungen haben. Nicht alle Mensch-Maschine-Interaktionen können als soziale Mensch-Maschine-Interaktionen bezeichnet werden. Mit der Unterscheidung zwischen *In-the-Loop*-Systemen, *On-the-Loop*-Systemen und *Out-of-Loop*-Systemen kann man mehrere Stufen der Kontrolle unterscheiden.
Wenn ein Fahrzeug vollständig unter menschlicher Kontrolle steht, ist es ein sogenanntes *In-the-Loop*-System. Hier handelt es sich um einen typischen Fall von Werkzeugnutzung, der nicht als soziale Interaktion betrachtet werden kann, da Werkzeuge keine Handlungsfähigkeit besitzen. Bei der Verwendung von Werkzeugen kann eine verminderte Zuweisung moralischer Verantwortung des menschlichen Akteurs nicht rechtfertigen, dass der Rest der moralischen Verantwortung dem Werkzeug zugewiesen wird. Ich folgere daraus, dass *In-the-loop*-Systeme keine Kontrolle über den Ausgang einer Interaktion haben. Dennoch können Werkzeuge eine kausale Rolle für das Ergebnis einer Interaktion haben, wir können ihnen aber keine moralische Verantwortung zuschreiben, da wir ihnen weder soziale noch moralische Handlungsfähigkeit zuschreiben können.
Andere autonome Fahrzeuge stehen jedoch nicht unter unserer vollständigen Kontrolle; sie weisen einen Grad von Autonomie auf und es ist denkbar, dass sie in der Zukunft auch in der Lage sind, sich anzupassen und zu lernen. Als *Out-of-the-Loop*-Systeme bezeichnet man Maschinen, bei denen der Mensch abgesehen von dem Auslösen der Aktion keine weiteren Eingriffsmöglichkeiten hat. In diesen Fällen ist die Frage, ob es sich um eine neue Art der sozialen Interaktion handelt, hinfällig, da der menschliche Agent nicht gemeinsam mit der Maschine interagiert. In Fällen von schwerem Fehlverhalten würde der Mensch natürlich trotzdem für die Einleitung dieser Aktion verantwortlich gemacht werden.
Interessant für die Diskussion über verteilte Verantwortung sind sogenannte *On-the-Loop*-Systeme, die über eine gewisse Autonomie verfügen. Beobachtet man Mensch-Maschine-Interaktionen, in denen *On-the-Loop*-Systeme involviert sind, kann man feststellen, dass sowohl die künstlichen als auch die menschlichen Agenten einen Einfluss und somit Kontrolle über den Ausgang der Interaktion haben. Wenn den involvierten künstlichen Systemen überdies eine Handlungsfähigkeit sowie sozio-kognitive Fähigkeiten zugeschrieben werden kann, dann könnten sich solche Interaktionen als ein neuer Typus sozialer Interaktionen qualifizieren. Und dann wird die hier aufgeworfene Frage, inwieweit alle Beteiligten für das Ergebnis der Interaktion verantwortlich sind, relevant. In der Klasse der *On-the-Loop*-Systeme könnte man Fälle finden, in denen die Zuschreibung von Verantwortung bezüglich der künstlichen Agenten dazu beitragen könnte, auftretende Verantwortungslücke zu schließen. Das ist zum heutigen Zeitpunkt sicherlich noch fiktiv, aber die zukünftige Entwicklung von autonomen, lernenden künstlichen Systemen, die auf neuronalen Netzen, genetischen Algorithmen und komplexen Agentenarchitekturen basieren, hat das Potential, solche Überlegungen relevant werden zu lassen. Negiert man die Möglichkeit, dass sich künstliche Systeme überhaupt als ein neuer Typus sozialer Interaktionspartner*innen qualifizieren können, sieht man sich in solchen Fällen mit einer Verantwortungslücke konfrontiert, die durch traditionelle Konzepte der Verantwortungszuschreibung nicht überbrückt werden kann.[23]
Alternativ ist es denkbar, dass man weiterhin daran festhält, die Hersteller und die solche Systeme benutzenden Menschen in vollem Umfang moralisch verantwortlich und haftbar zu machen, obwohl sie nicht in der Lage sind, das zukünftige Verhalten der künstlichen Systeme vorherzusagen und sie auch nicht die Möglichkeit haben, frühzeitig in die Interaktion einzugreifen. Geht man aber davon aus, dass es Interaktionen geben kann, in denen sowohl menschliche als auch künstliche Agenten in einer relevanten Weise zum Ergebnis der Interaktion beitragen, erscheint es zumindest bedenkenswert, verteilte Verantwortung in Erwägung zu ziehen. Eine Bewertung der interaktionsbezogenen Kriterien, wie z.B. eine eingeschränkte Antizipationsfähigkeit auf menschlicher Seite hinsichtlich der Handlungen des künstlichen Gegenübers, könnte hier für eine verminderte Verantwortung auf der menschlichen

[23] Orginalzitat: “facing a responsibility gap, which cannot be bridged by traditional concepts of responsibility ascription” (Mathias 2004, 175).

Seite sprechen und gleichzeitig dem künstlichen Agenten einen Teil der Verantwortung zuweisen. Ebenso können die eingeschränkten kognitiven Fähigkeiten hinsichtlich der Verarbeitung und Speicherung von Daten, die zu verminderten Mitwirkungs- und Eingriffsmöglichkeiten führen, als Argument dafür dienen, dass dem menschlichen Agenten ein verminderter Anteil an der Verantwortung zukommt. In der Folge könnte man dann einen Konstruktionsfehler bei den etablierten sozial konstruierten Verantwortungsbeziehungen diagnostizieren und hervorheben, dass soziale Mensch-Maschine-Interaktionen nicht ungerechtfertigterweise auf den Gebrauch von Werkzeugen reduziert werden sollten.

5.2 kritische Folgen

Nichtsdestotrotz bleibt das Projekt, künstlichen Agenten moralische Verantwortung zuzuschreiben, aus anderen Gründen kontraintuitiv. Ein gewichtiger Einwand wirft die Frage auf, was es bedeuten soll, wenn solche Agenten als moralisch verantwortlich angesehen werden, und dies gleichzeitig keine weiteren Konsequenzen nach sich zieht. In der menschlichen Sphäre geht die Zuschreibung von Verantwortung in der Regel mit Kritik und Vorwürfen oder auch Sanktionen einher. In Bezug auf künstliche Agenten haben wir keine Vorstellung davon, was überhaupt als Sanktion oder Strafe verstanden werden könnte, da künstliche Agenten keine Strafen erleiden können.[24] Selbst wenn wir also künstlichen Agenten einen Anteil der Verantwortung zuschreiben und ihnen die Möglichkeit einräumen, als neuer Typus sozialer Interaktionspartner*innen den Ausgang gemeinsamer Handlungen auf relevante Weise zu beeinflussen, wird dies von einem unguten Gefühl begleitet, das aus einem kategorischen Unterschied zwischen Menschen und diesem neuen Typus resultiert.

In Anbetracht des Fehlens möglicher Sanktionsmaßnahmen scheint das allgemeine Projekt, künstlichen Agenten Verantwortung zuzuschreiben, nun doch von vornherein zum Scheitern verurteilt. Weder ein gradueller Begriff von moralischer Handlungsfähigkeit noch die Zuschreibung sozio-kognitiver Fähigkeiten, wie etwa die Fähigkeit zum gemeinsamen Handeln, klären, welche Konsequenzen aus der Zuschreibung von Verantwortung folgen sollen.

An dieser Stelle könnte man sich radikalerweise fragen, ob man notwendigerweise einen engen Zusammenhang zwischen der Zuschreibung moralischer Verantwortung und den daraus folgenden Konsequenzen fordern muss. Dies stellt jedoch die Sinnhaftigkeit der Zurechnung moralischer Verantwortung grundsätzlich in Frage. Ähnlich wie bei der Debatte über Verantwortungslücken gibt es keinen Schuldigen, den wir bestrafen könnten. Eine vorübergehende Lösung für dieses Problem könnte darin bestehen, zu fordern, dass soziale künstliche Akteure mit einer Art Haftpflichtversicherung ausgestattet werden, die zumindest möglichen von ihn verantwortlich verursachten Schaden finanziell abdeckt.

Ein weiteres Problem, mit dem sich zukünftige Forschung befassen muss, besteht darin, dass wir künstlichen Agenten nur während einer sozialen Interaktion den Status eines neuen Typus eines sozialen Agenten zuschreiben. Außerhalb der Interaktion fällt der gerade noch sozial interagierende Agent wieder in die Rolle eines Werkzeuges zurück. Spätestens, wenn wir künstliche Systeme ausschalten, entziehen wir ihnen den Status eines sozialen Agenten. Dies widerspricht unserer Intuition, dass soziale Agenten auch außerhalb von Interaktionen einen sozialen Status haben.

6. Ausblick

Die Beantwortung der Frage, ob es Konstellationen gibt, in denen man gerechtfertigterweise künstlichen Systemen einen Anteil der Verantwortung zuschreiben kann, hängt davon ab, ob man der Prämisse, dass sich (zukünftige) künstliche Systeme als ein neuer Typus sozialer Interaktionspartner*innen qualifizieren können, zustimmen kann. Dann kann eine Neubewertung der interaktionsbezogenen Kriterien eine positive Antwort unterstützen.

Gesteht man künstlichen Systemen eine Überlegenheit bezüglich der Antizipationsfähigkeiten in Interaktionen sowie bezüglich der Fähigkeit, größere Datenmenge in kürzerer Zeit zu verarbeiten und zu speichern, zu, kann man dafür argumentieren, dass den menschlichen Interaktionspartner*innen eine geringere Antizipationsfähigkeit und eine weniger ausgeprägte Fähigkeit, Daten zu verarbeiten

[24] Sparrow 2007.

und zu speichern, zukommt. Daraus kann man dann wiederum folgern, dass den künstlichen Systemen ein Anteil an der Verantwortung zuzusprechen ist.
Aus der Neubewertung der interaktionsbezogenen Kriterien kann man dann etablierte sozial konstruierte Verantwortungsbeziehungen insofern hinterfragen, dass man sie um eine neu festzulegende Verantwortungsbeziehung erweitert, die für Interaktionen mit künstlichen Systemen gilt, die sich als soziale Interaktionspartner*innen qualifizieren und die sich nicht auf Werkzeuge reduzieren lassen.
Jedoch hat sich auch gezeigt, dass das Projekt, moralisches Handeln als graduelles Phänomen zu beschreiben, weitreichendere Konsequenzen nach sich ziehen kann, als der vorliegende Aufsatz behandelt hat. Denn selbst wenn sich die Idee künstlicher Agenten als soziale Interaktionspartner*innen in Mensch-Maschine-Interaktionen durchsetzen sollte, steht man vor weiteren grundlegenden Problemen. Dies betrifft den engen Zusammenhang zwischen Verantwortung und möglichen Vorwürfen und Sanktionen und die Frage, welchen Status künstliche Systeme außerhalb sozialer Interaktionen haben. Zukünftige Forschung sollte daher untersuchen, inwieweit künstliche Systeme ihrer Verantwortung gerecht werden könnten, wenn sie beispielsweise mit Haftpflichtversicherungen ausgestattet wären. Darüber hinaus muss detailliert erforscht werden, wie der Status der künstlichen Systeme außerhalb sozialer Interaktionen zu bewerten ist.
Das Besondere an der Verantwortung des Menschen ist ja, dass er sich nicht nur während einer sozialen Interaktion als verantwortlich erweist, sondern dass die Verantwortung darüber hinausgeht. Die kontinuierliche Existenz als sozialer Agent spielt eine wesentliche Rolle in unserer Praxis des Umgangs mit moralischen Verantwortungszuschreibungen. Wie künstliche Agenten, die sich gerade nicht in einer sozialen Interaktion befinden, zu beurteilen sind, bleibt an dieser Stelle eine offene Frage. Die Berücksichtigung dieser Einwände zeigt, dass noch viel Arbeit zu leisten ist, bevor wir in der Lage sind, einen neuen Typus von Interaktionspartner*innen mit all seinen ethischen Konsequenzen zu berücksichtigen.

Literaturverzeichnis

Bratman, Michael 2014, *Shared Agency: A Planning Theory of Acting Together*. Oxford University Press, Oxford.

Brown Coverdale, Helen & Wringe, Bill 2021, „Introduction: Nonparadigmatic punishments", in: *Journal of Applied Philosophy*, 38 (3), 357–365. https://doi.org/10.1111/japp.12499

Bryson, Joanna 2010, „Robots should be slaves", in: Y. Wilks (ed.), *Close engagements with artificial companions: Key social, psychological, ethical and design issues*, 63–74, John Benjamins Publishing, Amsterdam. https://doi.org/10.1075/nlp.8.11bry

Coeckelbergh, Mark 2020, „Artificial intelligence, responsibility attribution, and a relational justification of explainability", in: *Science and Engineering Ethics,* 26, 2051–2068. https://doi.org/10.1007/s11948-019-00146-8

Deroy, Ophelia 2021, „Rechtfertigende Wachsamkeit gegenüber KI", in: A. Strasser, W. Sohst, R. Stapelfeldt, K. Stepec (Hrsg.), *Künstliche Intelligenz - Die große Verheißung*. Reihe: MoMo Berlin Philosophische KonTexte 8, xenomoi Verlag, Berlin, 471–488.

Floridi, Luciando & Sanders, J. W. 2004, „On the Morality of Artificial Agents", in: *Minds and Machines*, 14, 349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

Gunkel, David 2020, „Mind the gap: responsible robotics and the problem of responsibility", in: *Ethics and Information Technology, 22, 307–320.* https://doi.org/10.1007/s10676-017-9428-2

Hauswald, Rico 2021, „Digitale Orakel? Wie künstliche Intelligenz unser System epistemischer Arbeitsteilung verändert", in: A. Strasser, W. Sohst, R. Stapelfeldt, K. Stepec (Hrsg.), *Künstliche Intelligenz - Die große Verheißung*. Reihe: MoMo Berlin Philosophische KonTexte 8, xenomoi Verlag, Berlin, 359–378.

Himma, Kenneth 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, 19–29. https://doi.org/10.1007/s10676-008-9167-5

Hortensius, Ruud & Cross, Emily S. 2018, „From automata to animate beings: the scope and limits of attributing socialness to artificial agents", in: *Annals of the New York Academy of Sciences*, 1426, 93–110.

Loh, Janina 2019, *Roboterethik*. Eine Einführung. Suhrkamp, Frankfurt.

Matthias, Andreas 2004, „The responsibility gap: ascribing responsibility for the actions of learning automata", in: *Ethics and Information Technology, 6, 175–183*. https://doi.org/10.1007/s10676-004-3422-1

Misselhorn, Catrin 2018, *Grundfragen der Maschinenethik*. Reclam, Stuttgart.

Moor, James H. 2006, „The nature, importance, and difficulty of machine ethics“, in: *IEEE Intelligent Systems*, 4, 18–21.
Nida-Rümelin, Julian & Weidenfeld, Nathalie 2018, *Digitaler Humanismus: Eine Ethik für das Zeitalter der Künstlichen Intelligenz.* Piper Verlag, München.
Nyholm, Sven 2018, „Attributing Agency to Automated Systems: Reflections on Human–Robot Collaborations and Responsibility-Loci“, in: *Science and Engineering Ethics*, 24, 1201–1219. https://doi.org/10.1007/s11948-017-9943-x
Scanlon, Thomas 2008, *Moral Dimensions. Permissibility, Meaning, Blame.* Harvard University Press, Cambridge.
Shoemaker, David 2011, „Attributability, answerability, and accountability: toward a wider theory of moral responsibility“, in: *Ethics*, 121(3), 602–632.
Sparrow, Robert 2007, „Killer Robots”, in: *Journal of Applied Philosophy*, 24 (1), 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x
Strasser, Anna 2020, „From tools to social agents“, in: *Rivista Italiana di Filosofia del Linguaggio, 14 (2)*. https://doi.org/10.4396/AISB201907
Van de Poel, Ibo; Nihlén Fahlquist, Jessica; Doorn, Neelke; Zwart; Sjoerd & Royakkers, Lambèr 2012, „The Problem of Many Hands: Climate Change as an Example“, in: *Science and Engineering Ethics,* 18, 49–67. https://doi.org/10.1007/s11948-011-9276-0
Véliz, Carissa 2021, „Moral zombies: why algorithms are not moral agents“, in: *AI & Society*, 36, 487–497. https://doi.org/10.1007/s00146-021-01189-x
Verbeek, Peter-Paul 2006, „Materializing Morality: Design Ethics and Technological Mediation”, in: *Science, Technology, & Human Values*, 31 (3). https://doi.org/10.1177/0162243905285847
Vincent, Nicole 2010, „ On the Relevance of Neuroscience to Criminal Responsibility”, in: *Criminal Law, Philosophy*, 4, 77–98. https://doi.org/10.1007/s11572-009-9087-4
Wallach, Wendell & Allen, Colin 2009, *Moral Machines: Teaching Robots Right from Wrong.* Oxford University Press, Oxford. https://doi.org/10.1093/acprof:oso/9780195374049.001.0001
Wallach, Wendell & Allen, Colin 2012, „Moral machines. Contradiction in terms or abdication of human responsibility?“, in: P. Lin, K. Abney, G. Bekey (eds.), *Robot Ethics. The Ethical and Social Implications of Robotics*, 55–68. MIT-Press, Cambridge.
Wykowska, Agnieszka; Chaminade, Thierry & Cheng, Gordon 2016, „Embodied artificial agents for understanding human social cognition“, in: *Philosophical Transactions of the Royal Society of London*. Series B, Biological Sciences, (371)1693, 20150375