Verantwortung in Zeiten der Künstlichen Intelligenz – Herausforderung für die theologische Ethik

von Kerstin Schlögl-Flierl

Stephan Ernst habe ich stets als denjenigen Kollegen schätzen gelernt, der am Puls der Zeit arbeitet und dabei immer zugleich die Quellen der philosophischen wie theologischen Tradition miteinbezieht und interpretiert. Deswegen widme ich ihm einen Aufsatz, der das Konzept der Verantwortung aufnimmt und auf die Frage nach der Interaktion von Mensch und Maschine zu übertragen versucht.

Ganz grundsätzlich könnte man von Forschung rund um Künstliche Intelligenz (KI) sprechen oder – um technisch genauer zu sein – vom maschinellen Lernen. Der Deutsche Ethikrat hat das folgendermaßen formuliert:

> „Das Gebiet des maschinellen Lernens ist ein rapide gewachsenes Teilgebiet der Informatik, das sich mit Modellen und Verfahren der Datenanalyse beschäftigt und mit einer Vielzahl von Anwendungsgebieten verzahnt ist. Seine historischen Wurzeln hat das maschinelle Lernen in Disziplinen wie Statistik, Physik, Mathematik, Mustererkennung und den Neurowissenschaften. Meist gibt es ein bestimmtes statistisches Modell, das mit Daten ‚trainiert' wird."[1]

Diese Technologien werden zunehmend das individuelle wie gesellschaftliche Leben prägen. Den damit verbundenen Auswirkungen auf die Anthropologie sollte sich auch die Theologie stellen.[2] „Hier wäre innertheologische Interdisziplinarität gefordert, die theologischen Reflexionen stehen allerdings noch am Anfang."[3] – stellt der Medienethiker Alexander Filipović fest. Sowohl von moraltheologischer als auch von sozialethischer Seite können in diesem neuen Forschungsfeld diverse Problemstellungen ausgemacht werden. Exemplarisch seien genannt: Transparenz und Diskriminierungsfreiheit der Systeme, Datenschutz und informationelle Souveränität, die Freiheit des

---

[1] *Deutscher Ethikrat*, Big Data und Gesundheit – Datensouveränität als informationelle Freiheitsgestaltung, Berlin 2017, 72.

[2] Vgl. *K. Schlögl-Flierl*, Gutes Leben zwischen Mensch und Maschine. Utopische und dystopische Anregungen aus dem Roman „Maschinen wie ich und Menschen wie ihr" von Ian McEwan, in: StZ 145 (2020), 529–538.

[3] *A. Filipović*, Nur Science-Fiction? Ethische Problemzonen der Künstlichen Intelligenz, in: HerKorr 75 (2021), 13–15, 15.

Menschen und vieles mehr. Filipović schneidet dezidiert sozialethische Themen an: „Wohlstandsverteilung auch in globaler Hinsicht, Machtfragen und Bedrohungen der demokratischen politischen Systeme sind beispielsweise sozialethische Probleme im Feld des Maschinellen Lernens."[4] In diesem Reigen wird zumeist auch das Konzept der Verantwortung angetippt.

Sie gerät deshalb ins Visier, da sie in Zeiten der Verantwortungsdiffusion immer weniger greifbar ist.

> Der Verantwortungsdiskurs „droht an dem Dilemma, dass seine grenzenlose Ausweitung einerseits moralisch nicht abweisbar erscheint, andererseits jedoch praktisch kaum zu bewältigen ist, zu scheitern. Der Verantwortungsdiskurs wird zum hilflosen Leerlauf. Die Grenzverschiebungen zwischen Schicksal und Verantwortung verkehren sich in eine höchst widersprüchliche Dynamik von zwei gegenläufigen Entwicklungen"[5], umreißt Markus Vogt dieses Spannungsfeld.

Schicksalshafte Widerfahrnisse werden heute nicht selten als Folge von menschlichen oder auch gesellschaftlichen Handlungen und Entscheidungen gedeutet. Andererseits wird der Verantwortungsbegriff ebenso überstrapaziert, denn in einer immer komplexeren Welt ist die Reichweite der möglichen Verantwortungsübernahme durchaus begrenzt. In dreifacher Weise wird der Verantwortungsdiskurs derzeit entgrenzt: in räumlicher Hinsicht durch lokale, aber auch globale Interdependenzen (1), bezogen auf die Zeitlichkeit, beispielsweise im Rahmen des Diskurses um intergenerationelle Gerechtigkeit (2) und drittens durch gesteigerte Eingriffstiefe in diverse Ökosysteme (3), vor allem forciert durch die wachsende Zahl an Möglichkeiten.[6]

In diesem Artikel soll es um die Frage nach der Verantwortung bezüglich maschinellen Lernens gehen und dabei ein Ökosystemblick eingenommen, d. h. KI als komplexes eigenes System betrachtet werden. Die Beleuchtung des Verantwortungsbegriffs stellt ein Desiderat dar, da in der KI-Debatte eine Verantwortungsdiffusion zu beobachten ist. Dennoch wird von der Schaffung eines „verantwortungsvollen, nachhaltigen, gemeinwohlorientier-

---

[4] *Filipović*, Science Fiction (s. Anm. 3), 15.

[5] *M. Vogt*, Verantwortung als Methode: Ethische Erkundigungen im Spannungsfeld zwischen Entgrenzung und Begrenzung, in: A. Seibert-Fohr (Hg.), Entgrenzte Verantwortung, Wiesbaden 2020, 49–71, 53.

[6] Vgl. ebd., 49.

ten und menschenzentrierten europäischen KI-Ökosystems als Wertschöpfungsnetzwerk für Innovationen"[7] gesprochen.

Auch kulturelle Komponenten bei der Verantwortungsübernahme dieser soziotechnischen Systeme sind zu bedenken.[8] Hier soll sich auf das Thema Verantwortung, ein klassisches Thema der Moraltheologie[9], beschränkt werden: Kann die KI Verantwortung übernehmen? Ebenso ist die Frage nach der Verantwortung aller am Prozess Beteiligten zu bedenken. Das Stichwort Dual-Use wäre ebenso anzusprechen, was aber in diesem Aufsatz nicht mehr geleistet werden kann.

1. Verantwortung in den Policy Paper zur KI-Ethik

Einen ersten Eindruck von der schwierigen Problemlage soll ein Blick in die diversen Policy Paper geben. Zum Bereich der KI-Forschung sind schon sehr viele unterschiedliche Policy Paper von Unternehmen, öffentlichen Organisationen und Forschungsinstitutionen geschrieben worden, nicht zuletzt auch von der Päpstlichen Akademie für das Leben 2020 der „Rome Call for AI Ethics"[10] usw. Viele Übersichtsartikel wie auch Homepages[11] bieten Orientierung, wer in welchem Papier welche Prinzipien angelegt hat.

Mit Jobin u. a.[12] soll zwar zum einen die Frage nach der Verantwortung weiterverfolgt werden, mit dem genannten Aufsatz kann man aber ganz kurz den Reigen an Prinzipien, die sich in Stellungnahmen zur AI Ethics finden,

---

[7] *Bundesregierung*, Stellungnahme zum Weißbuch zur Künstlichen Intelligenz – ein europäisches Konzept für Exzellenz und Vertrauen (2020), 4, https://www.bmi.bund.de/SharedDocs/downloads/DE/veroeffentlichungen/2020/stellungnahme-breg-weissbuch-ki.pdf;jsessionid=687249E60266B3CFC0D6B6E961A4BFF6.1_cid287?__blob=publicationFile&v=1 [Zugriff am 18.03.2021].

[8] Vgl. *acatech* (Hg.), Machine Learning in der Medizintechnik. Analyse und Handlungsempfehlungen (acatech Position), München 2020, https://www.acatech.de/publikation/machine-learning-in-der-medizintechnik/ [Zugriff am 18.03.2021].

[9] Vgl. *K.-W. Merks*, Theologische Fundamentalethik, Freiburg i. Br. 2020. Vgl. auch *J. Römelt*, Vom Sinn moralischer Verantwortung. Zu den Grundlagen christlicher Ethik in komplexer Gesellschaft, Regensburg 1996.

[10] Vgl. *Academy for Life*, Rome Call for AI Ethics, Rom 2020, https://www.romecall.org/the-call/ [Zugriff am 18.03.2021].

[11] Für eine Übersicht: Vgl. *T. Hagendorff*, The Ethics of AI Ethics. An Evaluation of Guidelines, in: Minds and Machines 30 (2020), 99–120, 103.

[12] Vgl. *A. Jobin* u. a., The global Landscape of AI Ethics Guidelines, in: Nature Machine Intelligence 1 (2019), 389–399.

deutlich machen: Transparenz, Gerechtigkeit, Fairness, Gleichheit, Nicht-Schädigung, Privatheit.

In den „Ethical guidelines for Trustworthy AI"[13] der Europäischen Kommission kommt beispielsweise der Verantwortungsbegriff recht häufig vor; aber oft als Synonym für Ethik an sich. Die Verantwortung als Konzept wird allerdings nicht näher gezeichnet, jedoch als Haftbarkeit, Nachvollziehbarkeit, Erklärbarkeit mit vertrauenswürdiger KI in Verbindung gebracht und damit operationalisiert.[14]

Jobin u. a. merken zum Verantwortungskonzept in diversen Policy Paper an, dass allgemein ein Definitionsdefizit herrscht.[15] Als mögliche inhaltliche Füllung werden die Herstellung von Verantwortung durch integres Handeln genannt, das Offenlegen von Verantwortlichkeiten (und Haftungsübernahmen), Verantwortung zur Enthüllung (whistleblowing) von potenziell möglichen Schädigungen usw. Als Verantwortliche und Haftende bei Entscheidungen und Handlungen, welche die KI ausführt, treten unterschiedliche Nennungen auf: Entwickler(innen), Designer(innen), Institutionen und die Industrie. Unterschiede sind auch bei den Überlegungen zu finden, ob die KI ebenso verantwortlich gemacht werden kann wie Menschen oder ob der Mensch am Ende die alleinige Verantwortung trägt.[16] All diese Fragen sollen auf den folgenden Seiten näher betrachtet werden.

Als grundlegende Beobachtung ist dabei diejenige von Hagendorff nicht von der Hand zu weisen, dass sich in den diversen Policy Paper eine männerdominierte Gerechtigkeitsethik finden lässt.[17] Mag man diese Vereinfachung, die noch aus der entwicklungspsychologischen Debatte von Carol

---

[13] Vgl. *The European Commission's High Level Expert Group on AI*, Ethical Guidelines for Trustworthy AI, Brüssel 2019; deutsche Fassung: *Unabhängige hochrangige Expertengruppe für Künstliche Intelligenz*, Ethik-Leitlinien für eine Vertrauenswürdige KI, Brüssel 2019, https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai [Zugriff am 18.03.2021].

[14] Im Ethik-Briefing, das Heesen u. a. entwickelt haben, drückt sich der Wert der Selbstbestimmung vor allem im Prinzip der Verantwortungswahrnehmung aus (notwendige Voraussetzungen sind nach Heesen Erklärbarkeit, Nachvollziehbarkeit, Vielfalt des Marktangebots, offene Schnittstellen und Systemoperabilität). All diese Faktoren werden in der technischen Umsetzung angewendet. *J. Heesen* u. a., Ethik-Briefing. Leitfaden für eine verantwortungsvolle Entwicklung und Anwendung von KI-Systemen, Lernende Systeme. Die Plattform für Künstliche Intelligenz, Whitepaper, München 2020, https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG3_Whitepaper_EB_200831.pdf [Zugriff am 18.03.2021].

[15] Vgl. *Jobin* u. a., The global Landscape (s. Anm. 12), 394.

[16] Vgl. ebd., 394f.

[17] Vgl. *Hagendorff*, The Ethics of AI Ethics (s. Anm. 11), 103.

Gilligan mit Lawrence Kohlberg stammt, nicht teilen, so ist doch ersichtlich, dass der Fokus auf Gerechtigkeitsaspekten liegt: „In contrast, almost no guideline talks about AI in contexts of care, nurture, help, welfare, social responsibility or ecological networks."[18] Technische Artefakte werden dabei nach Hagendorff vornehmlich als isolierte Entitäten angesehen, die, um der Findung von Lösung willen, technisch optimiert werden. Er eröffnet die Fehlanzeige, dass die Betrachtung des weiteren Kontextes wie auch die Beziehung(en), in welche die Technik eingebettet ist, stärker berücksichtigt werden sollten. Dies wird sich in der relationalen Aufschlüsselung des Verantwortungsbegriffs zeigen.

2. Grundlegende anthropologische Fragen und Einordnung in die Bereichsethik

Dabei ist die viel tiefergehende anthropologische Dimension, wie bereits angedeutet, zu beachten. Am Begriff der Verantwortung und damit einhergehend am Prozess der Verantwortungszuschreibung zeigt sich paradigmatisch, wie das Verhältnis von Mensch und Maschine in Zukunft gedacht werden soll. Ist die Technik Werkzeug oder Partnerin? Auf dieser Wegstrecke vom Werkzeug hin zur Partnerin könnte man zusätzlich unterschiedliche Pflöcke, als graduelle Abstufungen, einschlagen. In der Beschreibung dieses Interaktionsgeschehens gibt es Positionen, die von ausnahmslosen Asymmetrien (Maschine als reine Technik) über Verantwortungsnetzwerke[19] bis hin zu Symmetrien (Handlungspartnerschaften zwischen Mensch und Maschine) ausgehen.

Als Beispiel für Symmetrien kann Luciano Floridi[20] genannt werden, der von ‚onlife' und ‚Inforgs' spricht und eine völlige Verschmelzung von Mensch und Maschine annimmt, d. h. die Schnittstellen zwischen Mensch und Maschine sind nicht mehr sichtbar. Die Infosphäre wäre dann gleichsam wie die Biosphäre, beides als Selbstzweck zu sehen. Post- und transhumanistische Ansätze, die oft einen Bezug zur KI haben wie bei Ray Kurz-

---

[18] Ebd.

[19] Vgl. *Ch. Neuhäuser*, Roboter und moralische Verantwortung, in: E. Hilgendorf (Hg.), Robotik im Kontext von Recht und Moral, Baden-Baden 2014, 269–286.

[20] Vgl. *L. Floridi*, The Ethics of Information, Oxford 2013.

weil, wären ebenfalls kurz anzureißen.[21] Die Ersetzung des Menschen durch die/eine Maschine würde dann die weitestgehende Option bedeuten.[22]

Dies hat nicht nur die Anpassung der Maschine an den Menschen zur Folge, sondern auch umgekehrt die Anpassung des Menschen an die Maschine. Forciert wird diese Entwicklung, wenn die Schnittstellen zwischen Mensch und Maschine unklarer gestaltet werden oder fast ganz verschwinden.

> „Dadurch entsteht ein Verlust der ‚Spuren' des technischen Agierens mit der Konsequenz, dass die Technik als indisponibel erscheint und keine echte Mensch-Maschine-Interaktion mehr stattfinden kann."[23]

Es ergeben sich infolgedessen sogenannte ‚mixed realities', deren Begreifen der Zusammenhänge nicht mehr einfach möglich ist.[24] Die Interaktion mit den ‚mixed realities' birgt Gefahren und Chancen: Hubig nennt als positive Entwicklungen die Erweiterung der Vorstellungsräume, die Unterstützung und Entlastung bei z. B. gesundheitsbezogenen Sachverhaltsdiagnosen sowie die Entscheidungen in diesem Zusammenhang. Damit können Rationalisierungseffekte bezüglich des Einsatzes von Mitteln einhergehen. Die Möglichkeiten eines risikofreien Probehandelns sind erweiterbar.

> „Andererseits ist in Rechnung zu stellen, dass durch die Konfrontation mit bereits formierten Handlungsumgebungen Einschränkungen bezüglich der Kompetenz, sich zu frei gewählten Aspekten dieser Umgebung in ein Verhältnis zu setzen, stattfinden und aufgrund des Verlustes der Widerständigkeit der Handlungsumgebung auch Kompetenzverluste eintreten können."[25]

Hubig befürchtet eine Routinisierung und Vereinseitigung des Handelns, so dass Handeln immer stärker zu einem dekontextualisierten Handeln wird, d. h. sich die Zusammenhänge zwischen Mensch und Maschine nicht mehr nachvollziehen lassen. Werden die Schnittstellen also indisponibler, stellt sich die Forderung nach einer neuen Parallelkommunikation ein, um die

---

[21] Einen guten Überblick bietet *J. G. Michel*, Vom Transhumanismus zur These des erweiterten Geistes: Ethische Implikationen?, in: B. P. Göcke/F. Meier-Hamidi (Hg.), Designobjekt Mensch. Die Agenda des Transhumanismus auf dem Prüfstand, Freiburg i. Br. 2018, 199–224.

[22] Vgl. *K. Klöcker*, Zur ethischen Diskussion um Enhancement. Eine kritische Anmerkung zum Transhumanismus aus theologisch-ethischer Perspektive, in: B. P. Göcke/F. Meier-Hamidi (Hg.), Designobjekt Mensch (s. Anm. 21), 309–338, 316.

[23] *Ch. Hubig*, Der technisch aufgerüstete Mensch – Auswirkungen auf unser Menschenbild, in: A. Roßnagel u. a. (Hg.), Digitale Visionen. Zur Gestaltung allgegenwärtiger Informationstechnologien, Berlin u. a. 2008, 165–175, 165.

[24] Vgl. ebd., 168.

[25] Ebd., 172.

Interaktion neu bzw. auf einem anderen Kanal sichtbar zu machen. Damit geht auch die Forderung einher, den Bias[26], der in den Trainingsdaten evtl. vorliegt (Bevorzugung von weißhäutigen Proband(inn)en aufgrund der verwendeten Trainingsdaten), offenzulegen. Während der Nutzung wird die Frage nach einer Systemtransparenz ‚on demand' laut. Dabei darf nicht vergessen werden, dass auch eine gesellschaftliche Metakommunikation über die Mensch-Maschine-Vernetzung stattfinden muss.

Die Interaktion ist hierbei vor allem das zu betrachtende Element: Ist diese auf individueller Ebene des Menschen anzusiedeln? Welches sind die gesellschaftlichen Veränderungsprozesse, wenn sich Mensch und Maschine über einen längeren Zeitraum gegenseitig wechselseitig beeinflussen? Bevor die Debatte aber gleich in zu grundsätzliche Fragestellungen abdriftet, gilt es in einem nächsten Schritt, den Begriff und das Konzept der Verantwortung zu skizzieren.

Ganz kurz sei auch noch das Thema der Bereichsethik angetippt. Man spricht im weiten Feld von Maschinen-[27], Roboter- oder Informationsethik. Wichtig ist diesbezüglich, dass es sich um KI handelt, und zwar um maschinelles Lernen. Ob die Ethik nun von außen am Schluss einprogrammiert wird[28] oder die Ethikforschung eingebettet ist, kann als unterschiedliche Zugangsweisen zu diesem Thema aufgefasst werden. „Den Niederungen einer angewandten Ethik im Sinne einer Anleitung für Ingenieure oder Software-Designer ist die Maschinenethik konsequent entwachsen, weil Maschinen jetzt selbst als Teil unserer Gesellschaft verstanden werden müssen."[29]

Dies lässt sich an Einsatzbereichen des maschinellen Lernens in der Medizintechnik veranschaulichen: von der Prävention bis zur Nachsorge, von Anwendungen für Laien ohne Einbeziehung einer Ärztin oder eines Arztes, über die Einteilung der Patient(inn)en in Gruppen unterschiedlicher Priorität, Anwendungen zur Unterstützung der Ärztin oder des Arztes bei Diagno-

---

[26] Automatische Textanalyse reproduziert Stereotypen. Das ist bspw. problematisch, wenn damit Klassifikationen beim Arbeitsamt durch Algorithmen vorbereitet werden. Vgl. *J. Simon* u. a., Algorithmic bias and the Value Sensitive Design approach, in: Internet Policy Review 9 (2020), https://doi.org/10.14763/2020.4.1534 [Zugriff am 30.03.2021].

[27] Vgl. *M. Gessmann*, Maschinenethik. Vom philosophischen Schmuddelkind zum Hoffnungsträger. Teil I, in: PhR 66 (2019), 89–117, 89.

[28] Vgl. *C. Misselhorn*, Grundfragen der Maschinenethik, Stuttgart ^{3}2019.

[29] *Gessmann*, Maschinenethik (s. Anm. 27), 94.

sen und Therapieauswahl, bis hin zu autonomen Anwendungen ohne Eingreifen einer Ärztin oder eines Arztes.[30]

Als wie aktuell sich diese Fragen erweisen, wird neben den klassischen Anwendungsfeldern wie automatisiertes Fahren, medizinischer Diagnose, automatischer Übersetzung z. B. im Feld der Psychotherapie und KI sichtbar.[31] Bei der Diagnose und Überprüfung von Therapieoptionen wird in der Psychotherapie mittlerweile ohne Einbindung von Therapeut(inn)en gearbeitet, indem Avatare oder Virtual Reality Szenarien genutzt werden. Bei Depressionen und Angsterkrankungen kommen Chatbots zum Einsatz, von psychotherapeutischer Seite Elemente der Kognitiven Verhaltenstherapie, Akzeptanztherapie sowie Psychodynamik, wobei deutlich festgehalten werden muss, dass groß angelegte randomisierte Studien sowie Untersuchungen zu Patient(inn)enakzeptanz, Behandlungserfolg und klinischem Nutzen noch ausstehen. Kuhn gibt zu bedenken:

> „Bezogen auf die vulnerable Gruppe der psychisch Erkrankten ist weiterhin der Umgang mit Eigen- und Fremdgefährdung, folglich das Erkennen von und Reagieren auf Risiken zentral, um Schäden zu vermeiden. Menschliche Psychotherapeutinnen bzw. -therapeuten haben die ethische Verantwortung, geeignete Maßnahmen zu ergreifen, wenn eine Patientin bzw. ein Patient eine Gefahr für sich selbst oder andere darstellt."[32]

Das Fazit kann lauten, KI generiert Chancen, denn schwer erreichbare (Krankheitsbild oder räumliche Entfernung) bzw. behandelbare Gruppen können Zugang zur Psychotherapie erhalten. Risiken sind vor allem bei Selbstbestimmung, Zugangsgerechtigkeit und epistemischer Gerechtigkeit zu sehen.

3. Verantwortungsrelata

Angesichts dieser neuen Herausforderungen braucht es auch eine vertiefte Klärung des Konzeptes von Verantwortung: Verantwortung lässt sich als Entwurf von einer zweistelligen Relation (Rede-und-Antwort-Stehen) bis

---

[30] Vgl. *acatech*, Machine Learning (s. Anm. 8).

[31] Vgl. *E. Kuhn* u. a., Psychotherapie mit einer autonomen Künstlichen Intelligenz. Ethische Chancen und Herausforderungen, in: Psychiat Prax 48/Suppl. 1 (2021), 26–30. *A. Fiske* u. a., Your Robot Therapist Will See You Now: Ethical Implications of Embodied Artificial Intelligence in Psychiatry, Psychology, and Psychotherapy, in: J Med Internet Res 21 (2019), e13216.

[32] *Kuhn*, Psychotherapie (s. Anm. 31), 28.

hin zu einer mehrfachen Relation fassen.[33] Ganz grundsätzlich kann man beim Verantwortungssubjekt ansetzen (jemand als Träger – sei es eine Person oder eine Kooperation), welche die nötige Freiheit zur Verantwortungsübernahme besitzt. Davon zu unterscheiden ist das Verantwortungsobjekt (für etwas verantwortlich zu sein), seien es Handlungen und deren Folgen, Aufgaben, Zustände usw. Als dritte Relation wird der Adressat, die Adressatin benannt: beispielsweise die vom Handeln des Verantwortungssubjektes Betroffenen. Die vierte Relation bildet die Instanz, vor der die Verantwortung übernommen wird. Das Gewissen, andere Personen oder auch eine staatliche Rechtsgemeinschaft sind hier denkbar. Für eine Normwissenschaftlerin ist ebenso bedeutsam, dass der Bezug auf ein präskriptives Kriterium offengelegt wird: z. B. andere nicht zu schädigen. Das wäre dann die fünfte Relation.[34]

Sogleich soll der Verantwortungsbegriff auch technikspezifisch gerahmt werden, um der Verantwortung der KI auf die Spur zu kommen. Ausgangspunkt ist hier, dass es sich um Personen als verantwortliche Wesen handelt, denen eine Verantwortungsübernahme möglich ist: Ingenieurinnen und Ingenieure, Entwickler(innen), Informatiker(innen), Planer(innen), User(innen) und auch die Unternehmen und der Staat. Das Verantwortungsobjekt ist dann je nach Rolle zu beschreiben: Planen, Erfinden, Entwickeln, Anwenden.

Gegenüber einem Adressaten ist man in Bezug auf ein präskriptives Kriterium (z. B. Wohltun der Technik für den Menschen) im Rahmen eines Verantwortungsbereichs vor einer Instanz (seien es ebenso Gerichte) verantwortlich.

Heidbrink hat hier vier verschiedene Modelle der Verwendung des Verantwortungskonzepts im heutigen Diskurs ausgemacht. Zum einen ist in seinen Augen die Zurechnungsfähigkeit und Zuständigkeit als Verantwortung zu beschreiben. Neben dieser Handlungs(ergebnis)verantwortung ist es vor allem auch die Verantwortungskonzeption als folgenbasierte Legitimation, die retrospektiv oder prospektiv erfolgen kann. Bei der ex-post-Verantwortung sind es vollzogene Handlungsvorgänge, bei der ex-ante-Verantwortung vorausliegende Handlungsfolgen. Die dritte Form der Verantwortung als kontextualistisches Reflexionsprinzip (Meta-Verantwortung): Hier erfolgt die

---

[33] Als sehr guter Überblicksartikel hat sich hier erwiesen: *L. Heidbrink*, Definitionen und Voraussetzungen von Verantwortung, in: Ders. u. a. (Hg.), Handbuch Verantwortung, Wiesbaden 2017, 3–28.

[34] Vgl. noch viel ausführlicher *J. Sombetzki*, Strukturen und Relata der Verantwortung, in: L. Heidbrink, u. a. (Hg.), Handbuch Verantwortung (s. Anm. 33), 35–56.

Eruierung der Angemessenheit von Entscheidungsgründen in komplexen Einzelgebieten.[35] Bei der Metaverantwortung geht es nicht um die Bewertung der Handlungsfolgen, sondern um die Relation zwischen Handlungsregeln und -folgen. Bereits existierende Regeln und Prinzipien sind auf ihre jeweilige Anwendbarkeit zu überprüfen. Nicht zuletzt weist Heidbrink auf die Verantwortung als Struktur- und Steuerungselement hin. Darunter falle die Übernahme gesellschaftlicher Verantwortung durch privatwirtschaftliche Unternehmen, ausgedrückt als Corporate Social Responsibility: Rahmenregeln, Kontextgestaltung, Formen der Selbstbindung komplexer Systemprozesse usw. Konditionale und generische Voraussetzungen sind dabei möglich.

Je nachdem welche Moraltheorie angesetzt wird, formatiert sich der Verantwortungsbegriff. Ist man mit Nida-Rümelin in der kantischen Tradition verortet, so gehören Verantwortung, Rationalität und Freiheit zusammen und bilden ein sehr voraussetzungsreiches Konzept. Verantwortung heißt hier, Gründe benennen zu können. Sobald diese Handlung des Gründenehmens und -gebens auch auf Maschinen zutreffen würde, wären auch Maschinen verantwortlich. D. h., Maschinen würde der Personstatus zugeschrieben, was jedoch (noch) nicht der Qualität dieser Entität entspricht.[36] Aber da sich auch die Art von Handlungen für Maschinen verändert, scheint ein gradualistisches Konzept von Verantwortung angesichts des technologischen Fortschritts angezeigt. Die fast schon zu nennende ‚Letztverantwortung' würde beim Menschen liegen. Aber dies führt letztendlich zur spezifischen Frage der Verantwortung in der KI.

Für Verantwortlichkeit und Rechenschaftspflicht in Anwendungsfeldern von (KI-) Algorithmen wird die „Algorithmic Accountability" in Anschlag gebracht, die Fragen der Offenheit und Transparenz beinhaltet. Dabei beschreibt der Begriff Accountability die Bereitschaft oder Verpflichtung, Verantwortung zu übernehmen, verantwortungsvoll zu handeln und Rechenschaft abzulegen.

> „So erleichtert eine Offenlegung von Algorithmen Manipulationen und Imitationen und kann zu nachteiligen Effekten führen. Daher dreht sich die Diskussion zur Governance von Algorithmen auch um die Frage, wie Kontrolle und Verantwortlichkeit hergestellt werden können, ohne Codes offenzulegen."[37]

---

[35] Vgl. *Heidbrink*, Definitionen und Voraussetzungen (s. Anm. 33), 17.

[36] Vgl. *J. Nida-Rümelin/F. Battaglia*, Mensch, Maschine und Verantwortung, in: O. Bendel (Hg.), Handbuch Maschinenethik, Wiesbaden 2019, 57–71.

[37] Vgl. *F. Saurwein*, Automatisierung, Algorithmen, Accountability. Eine Governance Perspektive, in: M. Rath u. a. (Hg.), Maschinenethik. Ethik in mediatisierten Welten, Wiesbaden 2019, 35–56, 40.

Von direkter Verantwortung kann bei Algorithmen nicht gesprochen werden, sondern die Algorithmen sollen so in Verantwortungszusammenhänge eingebunden werden, dass bei der Programmierung auf ihre Accountability geachtet wird.

4. Verantwortung in der KI – Überblick über die angedachten Verantwortungskonzepte

Verantwortung bedeutet, dass eine Person, der man eine spezifische psychomotivationale Verfasstheit zuschreibt, in der Lage ist, im normativen und nicht nur rein deskriptiven Sinn für etwas Rede und Antwort zu stehen. Der Verantwortungsdiskurs hat in verschiedene Richtungen eine Ausweitung erfahren. Galt die Verantwortung lange als klassisches Handlungsprinzip (man denke an Hans Jonas, Prinzip Verantwortung[38]), so hat es sich zu einem nachklassischen Systemprinzip gewandelt.[39]

Auch aus der KI-Forschung selbst wird die Frage nach der Verantwortung heraus ventiliert und, wie ebenfalls in der Philosophie, unterschiedlich beantwortet[40], von Robotern und KI als moral agents bis hin zu moral patients.[41]

Besonders hier ist diese Frage virulent, denn je nach Verantwortungszuschreibung besitzt die Maschine eine eigene *agency* (Handlungsurheberschaft). *Agency* wird in der philosophischen und theologischen Tradition sehr eng gefasst, nämlich als die Durchführung von intentionalen, auf Gründen basierenden Handlungen. Ich würde aber diesen Punkt an dieser Stelle mit Fritz/Brand zurücknehmen:

> „Schlussendlich kommt man zu dem Ergebnis, dass die Risiken einer *agency*-Attribution auf maschinelles Verhalten den Nutzen überwiegen. Daher sollte eine sachgerechte Differenzierung zwischen Mensch und Maschine auch begrifflich erkennbar sein. So kann die Mensch-Maschine-Interaktion nicht nur präziser beschrieben, sondern auch die normative Struktur deutlicher herausgestellt werden. Dadurch wird angesichts zunehmend komplexer und sich ausdifferenzierender

---

38 Vgl. *H. Jonas*, Das Prinzip Verantwortung, Frankfurt a. M. [12]1995.

39 Vgl. *Heidbrink*, Definitionen und Voraussetzungen (s. Anm. 33).

40 Vgl. *J. Loh*, Roboterethik, Berlin 2019. Vgl. auch: *J. Loh*, Verantwortung und Roboterethik, in: K. Berr/J. H. Franz (Hg.), Zukunft gestalten – Digitalisierung, Künstliche Intelligenz (KI) und Philosophie, Berlin 2019, 17–28.

41 Vgl. *J. Loh*, Verantwortung – alt oder neu? Überlegungen zu dem Für und Wider einer Transformation der Verantwortung, in: Conc(D) 55 (2019), 317–324.

> Mensch-Maschine-Interaktion auch das nicht vergessen, worum es der Ethik ab der zweiten Hälfte des letzten Jahrhunderts mit dem Blick auf den technischen Fortschritt vornehmlich ging: die Verantwortung von Menschen."[42]

Anders ist der Ansatz von Wallach und Allen, denn sie formulieren in „Moral Machines. Teaching Robots Right from Wrong"[43] den Ansatz der funktionalen Äquivalenz, mit dem dieses Problem mangelnder Kompetenz bei artifiziellen Systemen umgangen werden soll. Roboter sind bei ihnen moral agents, also Handlungssubjekte. Moral agency bestimmen die beiden als graduelles Konzept mit zwei Bedingungen: Autonomie und Empfänglichkeiten. Menschen gelten als moralische Akteure, einige Maschinen sind operationale moralische Akteure. Nur besonders artifizielle Systeme haben bereits den Status funktionaler moralischer Akteursfähigkeit (wie z. B. kognitive, emotionale und andere Kompetenzen).

> „Obwohl Wallach und Allen den Übergang von operationaler über funktionale bis hin zu voller bzw. genuiner Moralzuschreibung in Abhängigkeit von den vorliegenden Kompetenzen (Autonomie und moralische Sensitivität) graduell denken, bleibt es doch schwer vorstellbar, wie (zumindest in der nahen Zukunft) ein artifizielles System zu der genuin menschlichen Fähigkeit ‚second order volitions' zu bilden entwickeln könnte bzw. zu der Fähigkeit, als self-authenticating sources of valid claims [...] über selbst gesetzte moralische Prämissen und Prinzipien zu reflektieren."[44]

Je höher die (moralischen) Ansprüche gedacht werden, umso schwieriger ist die Vorstellung von Robotern als moral agents.

Um dem Relationalitätsaspekt[45] dieser Interaktion von Mensch und Maschine gerecht zu werden, wird vom Konzept der geteilten Verantwortung ausgegangen. In einer ersten Position wird dabei von einer extrem eingeschränkten, aber vorhandenen Verantwortungsfähigkeit auch bei der Maschine (weder Freiheit noch Bewusstsein) gesprochen. Aber es handelt sich dabei nicht um moralische Verantwortung, also keine starke Verantwortung, sondern eine schwache, die dann nicht mehr den Begriff Verantwortung, verstanden als personale Instanz, verdient. Denn im Falle eines Fehlverhal-

---

[42] *A. Fritz/W. Brandt*, Verteilte Moral in Zeiten von KI? Über die moralische Bedeutung technischer Artefakte in der Mensch-Maschine-Interaktion, in: ThPh 94 (2019), 526–555, 554.

[43] Vgl. *W. Wallach/C. Allen*, Moral Machines. Teaching Robots Right from Wrong, Oxford 2009.

[44] *Loh*, Verantwortung und Roboterethik (s. Anm. 40), 22f.

[45] Vgl. *M. Coeckelbergh*, Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability, in: Science and Engineering Ethics 26 (2020), 2051–2068.

tens der Maschinen, die zu Schädigungen einer Person führen und wenn keine maschinellen-technischen Ursachen vorliegen, würde man nicht vom moralischen Versagen der Maschine sprechen, sondern vom Versagen des- oder derjenigen, der bzw. die die Maschine falsch, unvollständig etc. programmiert hat.

Beim Konzept der geteilten Verantwortung in einer zweiten Position werden nur Menschen als Akteure betrachtet und besonders in ihrer jeweiligen Funktion bedacht. Es geht also nicht um das Teilen der Verantwortung zwischen Mensch und Maschine, sondern zwischen den in unterschiedlichen Funktionen involvierten Menschen: Konstrukteurinnen und Konstrukteure, Anwender(innen), politische Entscheidungsträger(innen) usw. Die Verantwortung bleibt beim Menschen, wie es auch die Datenethikkommission vertreten hat:

> „Selbstbestimmung bedeutet auch, dass Menschen nicht nur Verantwortung übernehmen dürfen, vielmehr müssen sie sie übernehmen und müssen ihr auch gerecht werden. Verantwortung liegt nie bei einer Maschine, sondern immer beim Menschen, gegebenenfalls im Rahmen von institutioneller Verantwortung. Auch wenn ein technisches System eingesetzt wird, um im Rahmen einer automatisierten Datenauswertung Schlussfolgerungen wie die Gewährung eines Kredites anzuwenden, ist es die Verantwortung des Menschen, dieses System in einer ethisch vertretbaren Weise zu entwickeln und einzusetzen."[46]

Auf der Ebene der Menschen ist von einer Multiakteursverantwortung[47] die Rede. Der Ethikrat unterscheidet beispielsweise drei Ebenen möglicher Verantwortungszuschreibung im Bereich gesundheitsbezogener Big-Data-Anwendungen: Individuen, Organisationen, Staat. Dies kann meines Erachtens auch auf Fragen zur KI angewendet werden.

Bleibt man eher bei der ersten Position und sieht auch die Maschine als mögliche Akteurin an, so sind Anschlussfragen, wie diejenigen von Janina Loh (oft im Verbund mit Mark Coeckelbergh), zu adressieren. Sie fragen, ob wir mit einem Denken in der zweiten Richtung allen Herausforderungen, die uns aktuell in Situationen, in denen der traditionelle Verantwortungsbegriff an seine Grenzen zu stoßen scheint, begegnen können. Vielleicht müsste auch der Verantwortungsbegriff radikal transformiert werden oder

---

[46] *Datenethikkommission*, Gutachten der Datenethikkommission, Berlin 2019, 44, https://www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/it-digitalpolitik/gutachten-datenethikkommission.pdf?__blob=publicationFile&v=6 [Zugriff am 18.03.2021].

[47] Vgl. *Deutscher Ethikrat*, Big Data (s. Anm. 1), 249f.

man könnte „doch auf gemäßigtere Modulationen des traditionellen Verständnisses von Verantwortung bauen wollen“[48]. Je nachdem von wie vielen Stellen der Verantwortungsrelata man ausgeht, wird die Zuschreibung von Verantwortung an Maschinen umso hürdenreicher. Statt von Handlungen bei KI, könnte man von maschinellen Vollzugsweisen sprechen, um Maschinen und Menschen in ihren Vollzügen voneinander zu unterscheiden.

5. Beitrag der Theologischen Ethik[49]

Nach diesem Einblick in die Debatte um Verantwortung oder Nichtverantwortung der KI soll in einem letzten Schritt der mögliche Beitrag einer Theologischen Ethik in diesem Feld eruiert werden. Zuerst wird der aktuelle lehramtliche Zugang zur KI-Frage vorgestellt, dann die anthropologische Dimension und schließlich ein bestehender ethischer Ansatzpunkt erörtert.

5.1 Rome Call for AI Ethics

Bereits im Februar 2020, noch vor Ausbruch der Corona-Pandemie, veranstaltete die Päpstliche Akademie für das Leben eine Tagung zur KI, auf der der „Rome Call for AI Ethics“ entstand, der auch von Microsoft und IBM u. a. unterzeichnet wurde. Damit wurde die Absicht verfolgt, eine geteilte Verantwortung unter internationalen Organisationen, Regierungen, Institutionen und Technologien zu etablieren. Jedem Menschen und der gesamten Menschheit solle die KI dienen. Nicht größerer Profit oder die Ersetzung der Arbeitskraft des Menschen, sondern der Nutzen für jeden stehen bei diesem Aufruf im Mittelpunkt.

Sechs Prinzipien des ethischen Gebrauchs von KI werden von den unterzeichnenden Akteuren postuliert: Transparenz, d. h. die prinzipielle Erklärbarkeit wird gefordert; Inklusion („the needs of all human beings must be taken into consideration so that everyone can benefit and all individuals can be offered the best possible conditions to express themselves and develop“[50]); keine Voreingenommenheit zu schaffen oder danach zu handeln,

---

[48] *Loh*, Verantwortung – alt und neu? (s. Anm. 41), 323.

[49] Für ein Beispiel aus islamischer Perspektive sei hier genannt: *M. Y. Chaudhary*, Initial Considerations for Islamic Digital Ethics, in: Philosophy & Technology 33 (2020), 639–657.

[50] *Academy for Life*, Rome Call for AI Ethics (s. Anm. 10).

um Fairness und Menschenwürde zu wahren; Verlässlichkeit der KI-Systeme und nicht zuletzt Sicherheit und Datenschutz. Auch Verantwortung kommt vor, und zwar als die Verantwortlichkeit von jeder und jedem, die und der den Einsatz von KI konzipieren und einsetzen. Das entspräche dann dem Konzept der geteilten Verantwortung.

Drei große Felder werden in diesem Call benannt: neben der Perspektive der Ethik (mit Verweis auf Menschenwürde und Diskriminierungsfreiheit) auch dasjenige der Ausbildung, sprich die Teilhabe an der Gestaltung des Lebens mit KI durch entsprechende Kompetenzen, und die rechtliche Sphäre, die einerseits den Grundrechtsbezug herstellen, aber auch mit gesetzlichen Novellierungen den neuen Herausforderungen begegnen muss.[51]

Verantwortung wird in diesem „Call for AI Ethics" vornehmlich auf den Menschen bezogen, gar nicht auf die Maschine, was im Diskurs um die KI ein doch, wie oben dargelegt, bedeutsames Thema darstellt.

5.2 Anthropologische Herausforderung mit KI

Nach diesem lehramtlichen Zugang sollen auch grundsätzliche Herausforderungen benannt werden, gesteigert von fundamentalen hin zu speziellen. In der Fähigkeit der Unterscheidung zwischen Gott und Mensch zeigt sich theologisches Gespür. Dies gilt auch im Zeichen von maschinellem Lernen. Nicht so sehr die Verschmelzung zwischen Mensch und Maschine, sondern ebenso die noch mögliche Distinktion muss im Mittelpunkt stehen. Keine kybernetischen und posthumanistischen Ansätze (Abschaffung jeglicher kategorialer Unterschiede zwischen Mensch und Maschine), keine Akteur-Netzwerk-Theorien (welche den Unterschied zwischen Aktanten und Akteur(innen) vermindern), sondern eine dem maschinellen Lernen entsprechende Anthropologie muss entworfen werden, die Pluralität, kulturelle Relativität, Historizität des Menschseins mit ihrer Erforschung menschlicher digitaler Lebenswelten (global, inter- und transkulturell und divers) in den Mittelpunkt stellt. Nida-Rümelin fordert:

> „Der einzig plausible Gegenentwurf zur starken KI-Ideologie und ihrem impliziten mechanistischen Denken ist der digitale Humanismus. Ein Humanismus, der

[51] Eine für den deutschen Sprachraum spannende Veröffentlichung stellt das Thesenpapier „Digitalität und Künstliche Intelligenz: Technik im Dienst des Geist-begabten und Selbst-bewussten Menschen" der Clearingstelle Medienkompetenz an der KH Mainz dar. Abrufbar unter: https://medienkompetenz.katholisch.de/thesen-digitalitaet-ki/ [Zugriff am 18.03.2021].

> die menschliche Autorschaft nicht bezweifelt und nicht gefährdet, sondern diese vielmehr durch den Einsatz digitaler Techniken erweitert."[52]

Dabei wird sich vehement die Frage nach der menschlichen Souveränität stellen. Mit Filipović sind diese Transformationen des menschlichen Selbstverständnisses auf den Punkt gebracht:

> „Was ist das entscheidend Menschliche, wenn KI-Systeme besser fahren, Diagnosen treffen oder die geduldigeren Pfleger sind? Wie verhält sich die zumindest potenzielle All-Macht eines KI-Systems zu unserer Gottesvorstellung?"[53]

Wie im Bereich der Tier- oder Pflanzenethik wird es auch in der KI-Ethik darum gehen, nicht dem Anthropozentrismus zu verfallen, sondern auch die Bedürfnisse und Bedarfe des Menschen nicht aus dem Blick zu verlieren. Es sind die technischen Geräte auf den Menschen hin zuzuschneiden, zudem sollte eine anthropologische Hybris nicht darin bestehen, die Erklärbarkeit maschineller Systeme zu überstrapazieren. Wie auch menschliche Entscheidungen zumeist nicht bis ins Letzte perfekt sind, so ist auch die Erklärbarkeit von KI ebenso wenig perfekt.

> „Wer einen Computer zu Hilfe nimmt, um sich bei seinen Entscheidungen unterstützen zu lassen, wird bald merken, dass er an einen Punkt gelangen kann, wo der Computer faktisch vorentscheidet, weil uns als menschliche Entscheider Kriterien fehlen, anhand derer es uns vernünftig erscheinen kann, die Vorschläge des Computer noch einmal umzustürzen."[54]

Hier muss der Mensch weiterhin in Verantwortung gedacht werden, vor allem als Verantwortung fünfstelliger Relation, wie oben angedeutet. Verantwortung zeigt sich hier als mehrschichtiges komplexes Ganzes, wo es manchmal um direkte moralische Forderungen geht. Mit Korff kann hier die Gestaltungsverantwortung in Erinnerung gerufen werden.[55]

Menschen wägen Gründe ab. Vor allem diese Praxis des Abwägens lässt sich nicht algorithmisieren, besonders deutlich in Situationen wie moralischen Dilemmata.[56] Veranschaulicht wird der kategoriale Unterschied an

---

[52] *J. Nida-Rümelin/N. Weidenfeld*, Digitaler Humanismus. Eine Ethik für das Zeitalter der Künstlichen Intelligenz, München 2018, 60.

[53] *Filipović*, Science Fiction (s. Anm. 3), 13.

[54] *B. Koch*, Ethische und anthropologische Herausforderungen durch KI-basierte militärische Kampfroboter, in: US 74 (2019), 228–238, 235.

[55] Vgl. *Merks*, Theologische Fundamentalethik (s. Anm. 9), 155 legt den Ansatz von Korff zugrunde.

[56] Vgl. *Nida-Rümelin/Weidenfeld*, Digitaler Humanismus (s. Anm. 52), 103.

der Frage, ob wir einem Roboter vergeben können bzw. ein Roboter uns vergeben kann.[57] Oft ist von Strafe (auch im rechtlichen Sinne) für Roboter die Rede, im theologischen Kontext sollte es um Vergebung gehen. Wenn wir die Rede von Verantwortlichkeiten auch bei Robotern einplanen, müsste man auch von Vergebung reden.[58] So kann die Vergebung als Lackmustest für die Mensch-Maschine-Interaktion formatiert werden.

Eine theologische Fragestellung steht somit unter folgenden Vorzeichen: Wenn sich das menschliche Selbstverständnis in dieser Weise verändert, wie ist dann unter den genannten Bedingungen von Gott zu sprechen?[59] Auf der Ebene der Maschine wird immer wieder die Zuschreibung einer Seele an dieselbige angefragt. Nida-Rümelin bringt es auf den Punkt:

> „Es spricht viel dafür, dass das, was im Digitalisierungsdiskurs als ‚starke KI' bezeichnet wird – also die These, dass Softwaresysteme über Bewusstsein verfügen, Entscheidungen treffen, Ziele verfolgen, dass ihre Leistungen nicht lediglich Simulationen personaler Kompetenzen sind, sondern diese realisieren [...] eines Tages als eine Form des modernen Animismus, also der Beseelung von Nicht-Beseelung, gelten wird."[60]

Es bleibt abzuwarten, ob sich eine ‚starke KI' entwickelt, die Fragestellungen sind jedenfalls schon da und müssen bearbeitet werden.

5.3 Theologisch-ethische Agenda: Unterstützung der Vertrauenswürdigkeit

Um gleichfalls auf die konkrete Ebene zu kommen, soll kurz die Vertrauenswürdigkeit der KI besprochen werden. Vertrauenswürdigkeit ist nicht zwangsläufig ein ethisches Kriterium, sondern kann auch rein funktional verstanden werden, i. S. einer psychologischen Hemmschwelle. Es muss sich genähert werden, möchte man, dass die KI auch verwendet und akzeptiert wird.

Worin besteht das ethische Problem? Nicht wenige Black-Box-KI Ansätze widerstreiten dem ethischen Ideal des Menschen als autonomem Akteur, als autonomer Akteurin. Denn für autonome Entscheidungen ist eine hinrei-

---

[57] Vgl. *M. Nagenborg*, Can We Forgive a Robot?, in: B. Beck/M. Kühler (Hg.), Technology, Anthropology, and Dimensions of Responsibility (Techno:Phil 1), Stuttgart 2020, 147–155.
[58] Vgl. ebd., 149.
[59] Vgl. *Filipović*, Science Fiction (s. Anm. 3), 14.
[60] *Nida-Rümelin/Weidenfeld*, Digitaler Humanismus (s. Anm. 52), 19.

chende Informiertheit erforderlich, wie es sich im in der medizinischen Praxis etablierten und zugleich durch Big-Data und KI-Anwendung neu herausgeforderten Kriterium des „Informed Consent" ausdrückt. KI-Algorithmen sollten vertrauenswürdig, d. h. nachvollziehbar und infolgedessen erklärbar sein. Z. B. müssen sowohl Pflegepersonal als auch Ärzt(inn)en in der Lage sein, einen für die Diagnostik verwendeten KI-Algorithmus grundsätzlich zu verstehen und den Patient(inn)en gegenüber erklären zu können.

Auf drei Komponenten weist die Unabhängige Expertengruppe für Künstliche Intelligenz, eingesetzt von der Europäischen Kommission im Juni 2018, in ihren Ethik-Leitlinien für eine Vertrauenswürdige KI (veröffentlicht 2019)[61] hin: rechtmäßig im Sinne der Einhaltung der rechtlichen Vorgaben, ethisch im Sinne der Einhaltung ethischer Grundsätze und Werte und zudem robust, also die technische und soziale Perspektive aufnehmend. „Idealerweise wirken alle drei Komponenten harmonisch zusammen und überlappen sich in ihrer Funktionsweise."[62] Genau diese Vertrauenswürdigkeit muss ebenso auf der Agenda der Moraltheologie stehen, vor allem dadurch, dass das Verantwortungskonzept geklärt wird.

Um die Vertrauenswürdigkeit zu steigern, ist insbesondere die zeitliche Dimension der Verantwortung zu bedenken: Neben retrospektiver Verantwortung (wer ist schuld?), soll es gleichwohl um die vorausschauende gehen, wie es die Verfasser(innen) des Gutachtens der Datenethikkommission fordern:

> „Bei der Sammlung, Verarbeitung und Weitergabe von Daten müssen mögliche Auswirkungen auf Einzelne oder die Allgemeinheit unter Berücksichtigung künftiger Akkumulations-, Netzwerk- und Skaleneffekte, technologischer Möglichkeiten und Akteurskonstellationen abgeschätzt werden."[63]

An all diesen Stellen gilt es weiterzuarbeiten.

Schluss

Beim Thema Verantwortung in der Mensch-Maschine-Interaktion bekommt man einen ersten Einblick in zukünftige Inhalte der erweiterten Gemeinschaft von Mensch und Maschine. Dabei sollte es in den nächsten Jahren

---

61 Vgl. *The European Commission's*, Ethical Guidelines for Trustworthy AI (s. Anm. 13).

62 *The European Commission's*, Ethical Guidelines for Trustworthy AI (s. Anm. 13), Nr. 16.

63 *Datenethikkommission*, Gutachten (s. Anm. 46), 16.

weniger um kleinteilige Moraldebatten wie um automatisiertes Fahren und die Anwendung der utilitaristischen oder deontologischen Ethiktheorie auf Unterstützungsmaßnahmen medizinischer Entscheidung gehen, sondern um Gesamtkonzepte, die z. B. auch die Nachhaltigkeit der maschinellen Lernfähigkeit noch stärker berücksichtigen.[64] Zwar wird die ethische Reflexion als der technologischen Entwicklung nachhinkend empfunden und deswegen sollte die „Ethics by, in and for Design"[65] angegangen werden, jedoch bleibt der Bedarf an Ethik und Anthropologie bestehen. Steht weiterhin der Nutzen für den Menschen im Mittelpunkt der KI-Forschung, so müssten nicht nur bei Kampfrobotern[66], sondern auch bei Entscheidungstools für Richtersprüche oder Klassifikationsschemata bei der Vermittlung im Arbeitsamt klare Grenzziehungen vorgenommen werden.

---

[64] Vgl. *Gessmann*, Maschinenethik (s. Anm. 27), 101. https://bits-und-baeume.org/de [Zugriff am 30.03.2021] wäre als ein Beispiel anzuführen.

[65] Vgl. *P.-H. Wong/J. Simon*, Thinking About ‚Ethics' in the Ethics of AI, in: IDEES (20.02.2020), revistaidees.cat/en/thinking-about-ethics-in-the-ethics-of-ai/ [Zugriff am 18.03.2021].

[66] Vgl. *Koch*, Ethische und anthropologische Herausforderungen (s. Anm. 54).