bidt Ein Institut der Bayerischen Akademie der Wissenschaften

WORKING PAPER

Philosophische Überlegungen zur Verantwortung von KI

Eine Ablehnung des Konzepts der E-Person

Autoren

Klaus Staudacher, bidt
Julian Nida-Rümelin, Ludwig-Maximilians-Universität und bidt

Herausgeber

bidt – Bayerisches Forschungsinstitut für Digitale Transformation
www.bidt.digital

Impressum

Working Paper Nr. 2

Die vom bidt veröffentlichten Working Paper geben die Ansichten der Autorinnen und Autoren wieder; sie spiegeln nicht die Haltung des Instituts als Ganzes wider.

bidt – Bayerisches Forschungsinstitut für Digitale Transformation
Gabelsbergerstraße 4
80333 München
www.bidt.digital

Koordination
Margret Hornsteiner, Nicola Holzapfel
Dialog bidt
dialog@bidt.digital

Gestaltung
made in – Design und Strategieberatung
www.madein.io

Veröffentlichung: Dezember 2020
ISSN: 2701-2409
DOI: 10.35067/bv16-2z28

Das Bayerische Forschungsinstitut für Digitale Transformation (bidt) trägt als Institut der Bayerischen Akademie der Wissenschaften dazu bei, die Entwicklungen und Herausforderungen der digitalen Transformation besser zu verstehen. Damit liefert es die Grundlagen, um die digitale Zukunft der Gesellschaft verantwortungsvoll und gemeinwohlorientiert zu gestalten.

Das Projekt „Zum Verhältnis von Ethik und Rechtspolitik in der Digitalisierung“ bewertet rechtspolitische Vorhaben im Feld der Digitalisierung ethisch und juristisch. Neben der Erarbeitung konkreter ethischer und rechtlicher Empfehlungen geht es darum, etwaige Argumentationsdefizite aufzuzeigen, um so zu einer Rationalisierung der Debatte beizutragen. Darüber hinaus sollen die Voraussetzungen für eine gelungene interdisziplinäre Kooperation zwischen Rechtspolitik, Rechtswissenschaft und Ethik auf dem Gebiet der Digitalisierung geklärt werden. Das Projekt beschäftigt sich schwerpunktmäßig mit den Themenbereichen digitalisierte (internetgestützte) Kommunikation und Medizin sowie mit Verantwortungs- und Haftungsfragen bei der Kooperation von Mensch und Maschine.

Die Autoren

Klaus Staudacher, M.A., ist wissenschaftlicher Mitarbeiter am bidt.
E-Mail: klaus.staudacher@bidt.digital

Prof. Dr. Dr. h.c. Julian Nida-Rümelin lehrt Philosophie und politische Theorie an der Ludwig-Maximilians-Universität München und ist Mitglied im bidt-Direktorium.
E-Mail: nida-ruemelin.sekretariat@nida-ruemelin.de

Abstract

Beim Einsatz von Maschinen, deren Verhalten nicht umfassend vorhersehbar und kontrollierbar ist, kann es dazu kommen, dass im Schadensfall nicht zu klären ist, welcher der beteiligten menschlichen Akteure (ForscherInnen, KonstrukteurInnen, ProgrammiererInnen, TrainerInnen, BetreiberInnen) einen Fehler gemacht hat. Es erhebt sich daher die Frage, ob in solchen Fällen der Schaden nicht auch der Maschine selbst zugerechnet werden sollte, zum Beispiel, wie vom Europäischen Parlament gefordert, durch die Einführung der rechtlichen Kategorie einer „elektronischen Person". Gegen das Konzept einer E-Person und ganz generell gegen eine solche Art der Verantwortungszuschreibung sprechen jedoch grundsätzliche Bedenken. Verantwortung setzt nämlich, wie wir in diesem Working Paper zeigen wollen, ein Maß an Vernunft und Freiheit und, damit zusammenhängend, an Autonomie voraus, über das auch komplexe KI-Systeme auf absehbare Zeit nicht verfügen werden.

When using machines whose behaviour cannot be comprehensively predicted and controlled, it may occur that it is not possible to clarify which of the human actors involved (researchers, designers, programmers, trainers or operators) has made a mistake in the event of damage. In such cases, the question therefore arises as to whether the damage should not also be attributed to the machine itself, for example, by introducing the legal category of an „electronic person" as requested by the European Parliament. Fundamental concerns exist about the concept of an e-person, however, and, more generally, against such a form of responsibility. As we want to show in this working paper, responsibility presupposes a degree of reason and freedom and, in relation to that, a degree of autonomy that even complex AI systems will not have in the foreseeable future.

Philosophische Überlegungen zur Verantwortung von KI*

A. Einleitung

Vor dem Hintergrund immer größerer Fortschritte bei der Konzipierung und Entwicklung komplexer KI-Systeme wird seit einiger Zeit sowohl von Juristinnen und Juristen wie von Philosophinnen und Philosophen die Frage aufgeworfen, ab welchem Grad von Autonomie oder zumindest Eigenständigkeit Maschinen für ihr Verhalten moralisch und vor allem auch juristisch verantwortlich gemacht werden können bzw. sollten.[1] Beziehen sich manche Beiträge zu dieser Thematik ausdrücklich auf (aus Literatur oder Film bekannte) Science-Fiction-Szenarien,[2] erörtern andere AutorInnen den rechtlichen Status von bereits jetzt oder in Zukunft real existierender KI im Hinblick auf deren zivil- und sogar strafrechtliche Haftbarkeit.[3] Darüber hinaus wird im Hinblick auf Maschinen, deren Verhalten nicht umfassend vorhersehbar und kontrollierbar ist, auch das Problem der sog. „Verantwortungslücke“[4] diskutiert. Zu einer solchen Lücke kann es beim Einsatz komplexer KI-Systeme kommen, wenn bei einem Schadensfall nicht zu klären ist, welcher der beteiligten menschlichen Akteure (ForscherInnen, KonstrukteurInnen, ProgrammiererInnen, TrainerInnen, BetreiberInnen) einen Fehler gemacht hat. Es erhebt sich daher die Frage, ob in solchen Fällen der Schaden nicht auch der Maschine selbst zugerechnet werden sollte. Ganz in diesem Sinne hat das Europäische Parlament gefordert, zumindest in Bezug auf die „ausgeklügeltsten autonomen Roboter“ die Einführung der rechtlichen Kategorie einer „elektronischen Person“ zu prüfen, die „für den Ausgleich sämtlicher von ihr verursachter Schäden verantwortlich wäre“, wobei die Schadensregulierung durch einen von „Herstellern, Programmierern, Eigentümern und Nutzern“[5] zu stiftenden Haftungsfond ermöglicht werden soll. Selbst wenn durch ein geeignetes

* Der folgende Text erscheint – mit einigen Abänderungen und ohne den Teil B sowie ergänzt um einen Beitrag von *Nikolaus Bauer* zur Haftungslückenproblematik bei KI – mit dem Titel: „Verantwortungsteilung zwischen Mensch und Maschine?“ im Dezember 2020 in der Festgabe zum 10-jährigen Jubiläum der Reihe Robotik und Recht.

---

1 Vgl. exemplarisch *Beck, S.,* Über Sinn und Unsinn von Statusfragen – zu Vor- und Nachteilen der Einführung einer elektronischen Person, in: Günther, J.-P./Hilgendorf, E. (Hrsg.), Robotik und Gesetzgebung. Robotik und Recht, Band 2, Baden-Baden 2013[1], S. 239–262: „Je selbständiger die Entscheidungen der Maschinen werden, je menschenähnlicher sie aussehen und sich verhalten, je bedeutsamer sie für die Funktionsfähigkeit der Gesellschaft werden, desto größer scheint das Bedürfnis zu werden, ihnen Verantwortung, Pflichten, möglicherweise sogar Rechte zuzuschreiben.“ (S. 243)

2 Vgl. z. B. *Dennett, D.,* When HAL Kills, Who Is to Blame? Computer Ethics, in: Battaglia, F./Mukerji, N./Nida-Rümelin, J. (Hrsg.), Rethinking Responsibility in Science and Technology, Pisa 2014, S. 203–214.

3 Vgl. z. B. *Beck, S.,* Brauchen wir ein Roboterrecht? Ausgewählte juristische Fragen zum Zusammenleben von Menschen und Robotern, in: Japanisch-Deutsches Zentrum (Hrsg.), Mensch-Roboter-Interaktionen aus interkultureller Perspektive. Japan und Deutschland im Vergleich, Berlin 2012, S. 124–146; *dies.,* Dealing with the Diffusion of Legal Responsibility: The Case of Robotics, in: Battaglia/Mukerji/Nida-Rümelin (Fn.2), S. 167–181; *Erhardt, J./Mona, M.,* Rechtsperson Roboter – Philosophische Grundlagen für den rechtlichen Umgang mit künstlicher Intelligenz, in: Gless, S./Seelmann, K. (Hrsg.), Intelligente Agenten und das Recht. Robotik und Recht, Band 9, Baden-Baden 2016, S. 61–94; Gaede, K., Künstliche Intelligenz – Rechte und Strafen für Roboter?. Robotik und Recht, Band 18, Baden-Baden 2019; *Matthias, A.,* Automaten als Träger von Rechten, Berlin 2010[2].

4 Vgl. zu diesem Begriff *Matthias* (Fn.3), S. 33 ff.; vgl. zu diesem Problem auch *Beck* (Fn.3), S. 126 f.

5 Vgl. Europäisches Parlament, Parlamentarisches Dokument P8_TA (2017) 0051.

Versicherungssystem gewährleistet wäre, dass Opfer auf diese Weise adäquat entschädigt werden könnten, sprechen gegen das Konstrukt einer E-Person und ganz generell gegen eine solche Art der Verantwortungszuschreibung doch grundsätzliche Bedenken. Wie wir in diesem Paper zeigen wollen, setzt Verantwortung nämlich ein Maß an Vernunft und Freiheit und, damit zusammenhängend, auch an Autonomie voraus, über das auch komplexe KI-Systeme auf absehbare Zeit nicht verfügen werden.[6] Gleichzeitig wollen wir damit auch deutlich machen, dass düster-pessimistische ‚Terminator-Prognosen' bezüglich der Gefahren, die der Menschheit durch den Einsatz von KI drohen,[7] unbegründet sind. Damit soll nicht gesagt werden, dass die Verwendung von KI und ganz allgemein der voranschreitende Prozess der Digitalisierung keine Risiken birgt; aber diese bestehen nicht darin, dass Maschinen anstreben könnten, die Menschheit zu beherrschen oder gar zu vernichten.

Bevor wir unsere eigene Konzeption vorstellen, wollen wir uns zunächst mit einigen Aspekten der Position von *Matthias* auseinandersetzen, der in seinem Buch „Automaten als Träger von Rechten"[8] für die Verantwortlichkeit bestimmter Formen von KI argumentiert.

B. Die zentralen Kriterien der Verantwortungszuschreibung bei *Matthias*

Während viele AutorInnen eine Verantwortungszuschreibung bei KI erst in (allerdings nicht unbedingt mehr weit entfernter) Zukunft für gerechtfertigt halten, plädiert *Matthias* vor dem Hintergrund der von ihm diagnostizierten „Verantwortungslücke" bereits heute für eine Gesetzesänderung, die es ermöglichen soll, bestimmte lernfähige und ohne menschliche Intervention agierende Maschinen zivil- und strafrechtlich zur Verantwortung zu ziehen.[9] Da diese Maschinen noch in der Anwendung lernen müssen, käme es „unausweichlich" zu einem ‚suboptimalen und fehlerhaften Verhalten',[10] das weder von HerstellerInnen noch ProgrammiererInnen oder BetreiberInnen kontrolliert werden könne[11] und für das diese daher auch nicht

---

6 Auch wenn dabei Ausgangspunkt unserer Argumentation die bestehende Praxis der *menschlichen* Verantwortungszuschreibung ist – und damit die Frage nach den Bedingungen, die erfüllt sein müssen, damit wir andere Menschen (oder auch uns selbst) für etwas verantwortlich machen –, sind unsere Überlegungen doch nicht „speziesistisch" motiviert. Unsere Ablehnung der Einführung einer E-Person beruht also nicht auf der Annahme, nur Menschen könnten Personen (im Sinne von Verantwortungssubjekten) sein, und wir schließen keineswegs aus, dass es *irgendwann in ferner Zukunft* KI-Systeme geben könnte, denen Vernunft, Freiheit und Autonomie in dem für Verantwortungszuschreibung erforderlichen Umfang zukommt.

7 Vgl. insbesondere und prominent die Warnungen und Befürchtungen von Elon Musk: „There have been movies about this, you know, like Terminator [...]. There are some scary outcomes. And we should try to make sure the outcomes are good, not bad." (The Guardian vom 18.06.2014: → **https://www.theguardian.com/technology/2014/jun/18/elon-musk-deepmind-ai-tesla-motors** (zuletzt abgerufen am 27.09.2020)); „AI is a fundamental risk to the existence of human civilisation." (The Guardian vom 17.07.2017: → **www.theguardian.com/technology/2017/jul/17/elon-musk-regulation-ai-combat-existential-threat-tesla-spacex-ceo** (zuletzt abgerufen am 27.09.2020)).

8 Vgl. *Matthias* (Fn.3).

9 Vgl. *Matthias* (Fn.3), S. 239–254.

10 *Matthias* (Fn.3), S. 36.

11 Vgl.dazu *Matthias* (Fn.3), S. 11–38. Als Beispiel für ein derartiges noch in der Anwendung lernendes System führt Matthias unter anderem lernfähige Aufzüge an, die mithilfe neuronaler Netze die Verkehrsmuster im Gebäude analysieren und so die Warte- und Fahrtzeiten der Benutzer minimieren sollen. Da der der Aufzug „*selber* [...] im Laufe seines Betriebes die verwendeten Algorithmen verändere" (ebd., S. 18; Hervorhebung von *Matthias)*, sei sein Verhalten nicht mehr vorhersehbar. Für weitere Beispiele siehe ebd., S. 17–20.

verantwortlich gemacht werden dürften.[12] *Matthias* nennt fünf Bedingungen, die seiner Ansicht nach alle zusammengenommen gegeben sein müssen, um eine Entität *juristisch* zur Verantwortung ziehen zu können, und von denen die folgenden drei die wichtigsten sind: *Intentionalität; Rezeptivität und Responsivität für Gründe; Wünsche zweiter Ordnung*.[13] All diese Kriterien erläutert er an der Funktions- und Spielweise eines Schachprogramms und sieht sie bereits bei diesem als erfüllt an.[14]

Ob einem System *Intentionalität* zukommt, hängt für *Matthias* in Anschluss an Dennett[15] davon ab, ob es „Einsichten" in das Verhalten eines Systems gibt, zu denen wir *nur dann* gelangen können, wenn wir ihm gegenüber eine *intentionale Einstellung (intentional stance)* einnehmen,[16] d. h. wenn wir annehmen, dass das System Überzeugungen und Ziele hat und die geeigneten Mittel ergreift, um diese Ziele zu erreichen.[17] Ob das System dabei tatsächlich so beschaffen ist, wie von der *intentionalen Einstellung* unterstellt, ist dabei für *Matthias* gar nicht relevant: „wenn der intentionale Beschreibungsstandpunkt es uns erlaubt, es so zu beschreiben, dass wir auf möglichst einfache Weise zutreffende Voraussagen über sein Verhalten machen können, dann ist er gerechtfertigt".[18] *Matthias* ist nun ähnlich wie *Dennett* der Meinung, dass schon ein Schachprogramm von den meisten Menschen nur als ein intentionales System begriffen werden könne. Angesichts der Komplexität und Geschwindigkeit der in dem Programm ablaufenden Vorgänge der Zugauswahl und -bewertungen könnten wir nämlich gar nicht anders, als intentionale Formulierungen wie „‚Das Programm will das Feld e6 besetzen'" oder „‚es will seinen König aus der Schusslinie nehmen'" zu verwenden

12 Diese Konsequenz ist allerdings zumindest in rechtlicher Hinsicht nicht zwingend. Denn auch bei Schadensfällen, bei denen keinem der beteiligten menschlichen Akteure ein Fehler nachgewiesen werden kann, kennt das Recht verschiedene Formen *verschuldensunabhängiger* Haftung.

13 Vgl. *Matthias* (Fn.3), S. 44–70. Die anderen beiden Voraussetzungen sind: *die Fähigkeit zur Unterscheidung zwischen intendierten und bloß vorsehbaren Konsequenzen von Handlungen* und die *juristische Sanität.* Letztere bestehe darin, „dass der Akteur seine Wünsche zweiter Ordnung so wählt, dass sie von ähnlicher Art sind wie die Wünsche der anderen Akteure in seiner speziellen Umwelt" (S. 86). Da beide auf zwei der drei oben genannten Bedingungen Bezug nehmen, hängt ihre Plausibilität von der (von uns im Folgenden verneinten) Adäquatheit dieser Bedingungen ab. Sie müssen daher hier nicht extra behandelt werden. Alle fünf Kriterien sollen auch bei *moralischer* Verantwortung gelten. Darüber hinaus müsse eine *„moralische Person* […] Gegenstand moralischer Erwägungen und reaktiver Attitüden" wie Lob, Tadel oder Bewunderung sein können *(ders.,* S. 42 sowie S. 123–128 unter Berufung auf *Strawson,* P., Freedom and Resentment, in: Proceedings of the British Academy 48 (1962), S. 1–25.

14 Vgl. *Matthias* (Fn.3), S. 84 ff. Auch wenn er zugesteht, dass es sich bei ‚der Welt des Schachbretts' um eine ‚stark eingeschränkte Modellwelt' (S. 86) handelt, wird damit bereits deutlich, wie gering die Anforderungen sind, die *Matthias* an die Zuschreibung von Verantwortung knüpft.

15 Er bezieht sich vor allem auf *Dennett, D.,* The Intentional Stance. MIT Press, Cambridge/Mass., London 1987.

16 Vgl. *Matthias* (Fn.3), S. 48.

17 Vgl. *Matthias* (Fn.3), S. 46 ff. Oder in *Dennetts* Worten: „The intentional stance is the strategy of interpreting the behaviour of an entity (person, animal, artifact, whatever) by treating it as if it were a rational agent who governed its ‚choice' of ‚action' by a ‚consideration' of its ‚beliefs' and ‚desires'" *(Dennett, D.,* Intentional Stance, in: Wilson, R. A./Keil, F. C. (Hrsg.), The MIT Encyclopedia of the Cognitive Sciences, Cambridge/Mass. 2001, S. 412). Neben der *intentionalen Einstellung* unterscheidet *Dennett* noch eine *physikalische* und eine *funktionale Einstellung,* bei denen Systeme jeweils in physikalischer Hinsicht bzw. im Hinblick auf ihre Funktionsweise beschrieben werden. *Dennetts* Ansatz lässt sich als Versuch deuten, die Position eines intentionalen Realismus (d. h. die These, dass intentionale Zustände existieren) mit der Auffassung zu vereinbaren, dass sich alles Verhalten durch neuronale oder funktionale Zustände erklären lässt, für die es auf intentionaler Ebene keine Entsprechungen gibt. Siehe dazu und für eine kritische Darstellung von *Dennetts* Theorie intentionaler Systeme *Beckermann, A.,* Analytische Einführung in die Philosophie des Geistes, Berlin 2008³, S. 329–357.

18 *Matthias* (Fn.3), S. 46.

und „das Spiel als eine Folge von Zügen [zu] betrachten, die gemacht werden, um bestimmte strategische Ziele zu *erreichen,* Gefahren *abzuwenden* und so fort".[19]

Gegen diese Position lässt sich zunächst mal geltend machen, dass daraus, dass wir als Menschen ein System nur dann richtig verstehen können, wenn wir ihm intentionale Zustände zuschreiben, noch nicht folgt, dass dieses System tatsächlich intentionale Zustände hat; zumal dieses mangelnde Verständnis ja möglicherweise auch nur an der Begrenztheit unserer intellektuellen Fähigkeiten liegt und intelligentere Wesen als wir das Verhalten eines solchen Systems begreifen könnten, ohne dass sie unterstellen müssten, es verfüge über *Intentionalität.*[20] *Sofern* wir das Verhalten eines Systems aber tatsächlich nur dann vollständig einordnen können, wenn wir ihm intentionale Zustände zuschreiben, könnte es gleichwohl dennoch sinnvoll sein, es so zu behandeln, als ob es intentionale Zustände hätte. Dies ist im Hinblick auf Schachprogramme und ganz allgemein in Bezug auf heutige KI aber überaus fraglich. Zwar hat Matthias sicher recht, dass wir den Spielverlauf nicht durch „technische" Detailbeschreibungen wiedergeben von der Art: „‚Die Figur auf b7 ist nach b6 gewandert, weil der Ast 1372 im untersuchten Spielbaum 6.Halbzug ab dieser Position um 2.735 Punkte höher ausfällt als die alternativ betrachteten Äste 611, 79, 428 und 3881'".[21] Daraus dass wir Vorgänge oder das Verhalten eines System mit *intentionalistischem* Vokabular darstellen, folgt allerdings noch nicht, dass wir ihm tatsächlich intentionale Zustände zuschreiben; denn dieser Sprachgebrauch kann auch rein metaphorisch gemeint sein – so etwa, wenn wir die (sicher komplexe und regelgeleitete, aber doch nicht Absichten mitteilende) Interaktion zwischen Bienen als „Bienen-Sprache" bezeichnen oder wenn wir die Wirkungsweise eines Thermostats erläutern, indem wir sagen: *Das Thermostat „stellt fest", wenn eine bestimmte Raumtemperatur überschritten wird, und aufgrund dieser „Information" wird die weitere Zufuhr von Heißwasser in die Heizkörper abgestellt.*[22] Eine solche metaphorisch-figurative Ausdrucksweise liegt natürlich immer dann nahe, wenn die entsprechenden ‚wortwörtlichen' Formulierungen komplizierter sind. *Matthias* müsste allerdings konsequenterweise bestreiten, dass wir das Schachprogramm lediglich aus Gründen der Sprachökonomie *intentionalistisch* beschreiben; seine These ist ja vielmehr, dass wir (oder zumindest die meisten von uns) die Spielweise des Programms gar nicht richtig wiedergeben können, wenn wir ihm gegenüber keine *intentionalistische Einstellung* einnehmen. Letzteres ist nun aber durchaus zweifelhaft, denn wir können auch extrem leistungsfähige Schachcomputer wie Deep Blue oder Hydra als Systeme auffassen, die – solange sie bestimmungsgemäß funktionieren (d. h. die Zwecke erfüllen, zu denen sie konstruiert, programmiert und trainiert wurden) – auf Grundlage der ihnen zu Verfügung gestellten Daten stets diejenigen Züge ausführen werden, die geeignet sind, das ihnen *von außen vorgegebene* Ziel zu erreichen, das Spiel zu gewinnen.[23] Das *intentionale* Element kommt bei dieser Beschreibung vollständig von außen, ist also kein interner Bestandteil des Systems. *Matthias* selbst räumt ein, dass sogar beliebig komplexe

---

19 *Matthias* (Fn.3), S. 84 (Hervorhebung bei *Matthias).*

20 Vgl. zu diesem Einwand auch *Beckermann* (Fn.17), S. 335.

21 Ebd.

22 Vgl. zu diesen beiden Beispielen und zur konstitutiven Rolle von *Intentionalität* für sprachliche Bedeutung und Verständigung *Nida-Rümelin, J.,* Verantwortung, Stuttgart 2011, S. 56 ff.

23 Und das gilt ungeachtet dessen, dass diejenigen Personen, die sie konstruiert, programmiert und trainiert haben, nicht in der Lage gewesen wären, Großmeister wie Garri Kasparow oder Michael Adams zu schlagen. Es waren fraglos jeweils Deep Blue und Hydra, die gegen sie gewonnen haben. Aber auch das macht sie eben, wie gerade erläutert, noch nicht zu Systemen, deren Verhalten wir nur adäquat verstehen können, wenn wir ihnen Überzeugungen und Absichten zuschreiben.

und lernfähige Maschinen intentionales Verhalten nicht aus eigenem Antrieb zeigen können und sich in ihnen lediglich die Intentionen ihrer KonstrukteurInnen und ProgrammiererInnen widerspiegeln.[24] Seine von *Dennett* übernommene Verteidigungsstrategie besteht nun in der Position, dass auch Menschen letztlich über keine „originäre", sondern lediglich über eine „abgeleitete" Intentionalität verfügen würden; unsere Intentionalität sei nämlich abgeleitet von der ‚Intentionalität unserer eigennützigen Gene'; wir seien nichts anderes als „‚Überlebens-Maschinen', durch den evolutionären Prozess von Variation und Selektion für die Aufgabe optimiert, die Zukunft unserer Gene zu sichern [...]. Und wenn es irgendwo eine originäre Intentionalität gibt, welche von keiner anderen Quelle abgeleitet ist, dann ist es die Intentionalität der natürlichen Evolutionsprozesse selbst".[25]

Was hier gesagt werden soll, ist offensichtlich: So wie das Schachprogramm seine, metaphorisch gesprochen, ‚Gewinnabsicht' ausschließlich seinen KonstrukteurInnen und ProgrammiererInnen verdankt, so verdanken auch wir unsere Intentionen ausschließlich der Evolution; folglich verfügen weder das Schachprogramm noch wir Menschen über ‚eigene', d. h. nicht von außen vorgegebene Intentionen. Da wir uns aber, so ist der Gedankengang wohl zu verstehen, gleichwohl als Menschen gegenseitig Intentionen zuschreiben, spricht auch nichts dagegen, Maschinen als intentionale Systeme anzusehen. Eine solche Schlussfolgerung ist allerdings alles andere als zwingend, denn wenn uns unsere Intentionen tatsächlich in derselben vorprogrammierten Weise vorgegeben worden wären, wie die Gewinnabsicht dem Schachprogramm, sollten auch wir selbst sofort aufhören, uns als intentionale Wesen zu betrachten. Wie wir im nächsten Abschnitt nämlich zeigen wollen, sind Intentionen, die unseren Handlungen vorausgehen, nicht gleichzusetzen mit bloßen Neigungen, die sich nach der Art eines Hungergefühls ohne unser Zutun einfach in uns einstellen, sondern sie sind das Ergebnis eines internen Abwägungsprozesses. Darüber hinaus ist aber auch die Analogie zwischen KonstrukteurInnen/ProgrammiererInnen und Evolution fragwürdig. Denn während Erstere dem Schachcomputer eine Gewinnabsicht erst von außen durch entsprechende Programmierung ‚einsetzen', sind uns diejenigen Intentionen, die sich im Prozess der Evolution als vorteilhaft erwiesen haben, eben nicht erst von der Evolution gegeben worden; sie waren vielmehr schon in uns als *unsere eigenen* ‚un-abgeleiteten' Intentionen (oder entsprechende Vorformen) vorhanden und stellten sich dann, da ihre Ausprägung sich gegenüber anderen Intentionen als vorteilhaft erwiesen hat, im Prozess der Evolution als für das Überleben besonders nützlich heraus.[26]

Matthias' Verständnis von *Intentionalität* vermag also nicht zu begründen, warum wir Maschinen in vergleichbarer Weise intentionale Zustände zuschreiben sollten wie Menschen.

---

24 Vgl. *Matthias* (Fn.3), S. 49.

25 *Matthias* (Fn.3), S. 50 f. unter Berufung auf *Dennett* (Fn.15), S. 297 f., und auf *Dawkins, R.,* The Selfish Gene, Oxford 1976.

26 Und ebenso wurde auch unsere Fähigkeit, überhaupt intentionale Zustände ausbilden zu können, nicht erst von der Evolution geschaffen, sondern sie (bzw. zunächst rudimentäre Formen davon) existierte bereits und musste dann noch die Herausforderungen des evolutionären Prozesses ‚überstehen'. Damit soll nicht bestritten werden, dass die im Prozess der Evolution jeweils wirksamen Außenbedingungen für die Entwicklung unserer intentionalen Zustände relevant sind. Aber eben nicht in der Weise, dass uns die entsprechenden Fähigkeiten von außen durch den evolutionären Prozess von Variation und Selektion ‚implantiert' worden wären.

Mit dem zweiten Kriterium, der *Rezeptivität und Responsivität für Gründe,* schließt *Matthias* an Überlegungen von *Fischer* und *Ravizza* an.[27] Ihm zufolge müssen (juristische) Verantwortungsträger in der Lage sein, „auf *Gründe* für oder gegen eine bestimmte Handlung [zu] reagieren [...]. Sie müssen empfänglich (rezeptiv) für Gründe sein, und sie müssen auf überzeugende Gründe mit einer Änderung ihrer Handlungspläne reagieren können."[28] Auch diese Voraussetzungen würden von einem Schachcomputer erfüllt werden, denn dieser sei „sowohl empfänglich für Gründe, die für bestimmte Züge und gegen andere sprechen, als auch responsiv, d. h. die von ihm durchgeführten Züge stellen sinnvolle Antworten auf diese Gründe dar"[29]. Ein derartiges Verständnis von Gründe-geleitetem Verhalten geht jedoch völlig fehl. Denn ob und inwieweit ein Schachcomputer Züge ausführt, für die gute (oder auch schlechte) Gründe sprechen, hängt bei vorschriftsmäßiger Anwen-dung ausschließlich davon ab, wie er konstruiert, programmiert und gegebenenfalls trainiert wurde (und Spielverhalten, für das keine guten Gründe sprechen, würden wir daher auch nicht ihm, sondern, wenn überhaupt, eben dem/der KonstrukteurIn, ProgrammiererIn und/oder TrainerIn zurechnen). Die guten Gründe, an denen er sein Verhalten ausrichtet, sind ihm also von außen vorgegeben worden. Für gute Gründe empfänglich zu sein und angemessen auf sie zu reagieren, bedeutet aber mehr, als ein Verhalten zu zeigen, für das, rein äußerlich betrachtet, gute Gründe sprechen. Man muss sich dafür vielmehr diese guten Gründe in einem Abwägungsprozess selbst zu eigen gemacht haben; es müssen also die ‚eigenen' guten Gründe sein.[30]

Ein weiterer Punkt, der sich gegen die Möglichkeit eines Gründe-geleiteten Verhaltens zumindest derjenigen KI-Systeme anführen lässt, die rein auf der Ebene der Symbolverarbeitung funktionieren, ist die Nicht-Algorithmizität von Gründen. Es lässt sich nämlich beweisen, dass bereits für die Prädikatenlogik erster Stufe (und damit erst recht für ausdrucksstärkere Logiksysteme) kein mechanisches Verfahren, also kein Algorithmus existiert, mit dem sich seinerseits sämtliche in ihr vorkommenden logisch wahren Aussageformen beweisen lassen könnten,[31] wenn nun aber Logik ein essenzieller Bestandteil unserer Praxis des Begründens ist, dann ist damit auch gezeigt, dass Deliberation, also die Abwägung von Gründen, kein rein mechanischer Vorgang sein kann.

Die dritte Bedingung der Verantwortungszuschreibung, die *Wünsche zweiter Ordnung,* übernimmt *Matthias* von *H. G. Frankfurt.*[32] *Frankfurt* differenziert zwischen *Wünschen erster Ordnung (first-order desires), Willen (will),* Wünschen *zweiter Ordnung (second-order desires)* und *Volitionen zweiter Ordnung (second-order volitions).* Unter *Wünschen erster Ordnung* versteht er Wünsche, die sich auf Handlungen beziehen, die aber im Vergleich zu den anderen Wünschen einer Person nicht unbedingt stark genug sind, um tatsächlich

---

27 Vgl. dazu *Fischer, J. M./Ravizza, M.,* Responsibility and Control. A Theory of Moral Responsibility. Cambridge Studies in Philosophy and Law, Cambridge University Press, New York 1998.

28 *Matthias* (Fn.3), S. 52 (Hervorhebung bei Matthias).

29 *Matthias* (Fn.3), S. 84.

30 Vgl. *Nida-Rümelin, J.,* Eine Theorie Praktischer Vernunft, Berlin 2020, S. 384 f.; *ders.,* Über menschliche Freiheit, Stuttgart 2005, S. 89 f.

31 Dieser Befund ergibt sich aus meta-mathematischen Resultaten zur Berechenbarkeit und Entscheidbarkeit von Kurt Gödel (vgl. dazu *Gödel, K.,* Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I, in: Monatshefte für Mathematik und Physik 38 (1931), S. 173–198). Siehe dazu genauer *Nida-Rümelin* 2020 (Fn.30), S. 349 f.

32 Vgl. dazu *Frankfurt, H. G.,* Freedom of the Will and the Concept of a Person, in: Journal of Philosophy 68 (1971) S. 5–20.

handlungsleitend zu sein; einen *Wunsch erster Ordnung,* der die entsprechende Intensität aufweist, um ihr Handeln zu leiten, bezeichnet *Frankfurt* als *Willen;* einen *Wunsch zweiter Ordnung* hat jemand, der einen bestimmten *Wunsch erster Ordnung* haben will; will er dabei auch, dass dieser *Wunsch erster Ordnung ein Wille* (also ein handlungsleitender Wunsch) wird, dann ist der entsprechende *Wunsch zweiter Ordnung* zugleich eine *Volition zweiter Ordnung*. Für *Frankfurt* sind diese Unterscheidungen für die Bestimmung des *Personenbegriffs* und des Begriffs der *Willensfreiheit* von Bedeutung. So sei es für Personen charakteristisch, dass sie *Volitionen zweiter Ordnung* haben, also dass es für sie wichtig ist, welche *Wünsche erster Ordnung* bei ihnen handlungsleitend werden. *Frankfurt* grenzt Personen dabei ab vom Typus des sog. *Wanton,* ein triebhaftes Wesen, das zwar *Wünsche erster Ordnung* hat, das es aber nicht kümmert, welche dieser Wünsche sein Handeln leiten; der *Wanton* wählt einfach immer diejenige Handlung, die seiner stärksten Neigung entspricht, wobei er durchaus Überlegungen anstellen kann, wie er das, was er tun möchte, am besten erreicht. Während *Wantons* über Handlungsfreiheit verfügen (können) – sofern sie nicht von außen daran gehindert werden, können sie das tun, was sie (im Sinne ihrer *Wünsche erster Ordnung)* wollen –, kommt Willensfreiheit nur Personen zu; sie sei gegeben, wenn die Person frei ist, den *Willen* zu haben, den sie haben möchte.[33]

Wie *Matthias* ganz richtig feststellt, setzt Verantwortung nicht nur die Freiheit des Akteurs voraus, „nach vorgegebenen Wünschen *[erster Ordnung]* zwischen alternativen Handlungssequenzen zu wählen, sondern er muss in einem übergeordneten Sinn über die Freiheit verfügen, auch diese Wünsche selber wählen zu können".[34] Seiner Meinung nach kann man dabei den „Wunsch" des Schachprogramms, einen bestimmten Zug auszuführen (also z. B. den Zug eines Bauern von e2 nach e4), als *Wunsch erster Ordnung* auffassen, der selbst wiederum Teil einer aus einer Züge-Kette bestehenden Strategie sei; und genau die Entscheidung zwischen unterschiedlichen Züge-Ketten ließe sich als „Wahl der konkreten Mittel, um aus einer bestimmten Spielsituation ein Maximum an Bewertungspunkten herauszuholen" mit einem *Wunsch zweiter Ordnung* vergleichen.[35] Dagegen sei das dem Programm vorgegebene Ziel, das Spiel zu gewinnen, nicht lediglich ein „‚Wunsch' *zweiter,* sondern noch höherrangiger Ordnung, und sofern man dieses Ziel als ‚letzten Zweck eines Schachprogramms' ansehe, ein ‚Wunsch letzter (n-ter) Ordnung'".[36] *Matthias* hat hier offensichtlich ein Verhältnis der Über- und Unterordnung vor Augen, bei dem die Erfüllung des jeweils untergeordneten Wunsches – er spricht auch von ‚(Sub)-Zielen'[37] – ein geeignetes Mittel zur bestmöglichen Erfüllung des jeweils übergeordneten darstellt; hat jemand z. B. den Wunsch, an einem bestimmten Tag schnellstmöglich von München nach Hamburg zu gelangen, und ein Direktflug ist unter den gegebenen Bedingungen schneller als die Alternativen Auto- oder Zugfahrt, dann wäre dieser Wunsch dieser Sichtweise zufolge ein *Wunsch zweiter Ordnung,* der durch das von einem *Wunsch erster Ordnung* zu wünschende Mittel *(einen Direktflug nehmen)* bestmöglich erfüllt wird. Zwischen *Wünschen zweiter* und Wünschen

33 In *Frankfurts* Worten: „It is in securing the conformity of his will to his second-order volitions, then, that a person exercises freedom of the will" *(Frankfurt* (Fn.32), S.15). Für *Frankfurt* steht es der Verantwortung eines Akteurs dabei nicht entgegen, wenn seine *Wünsche zweiter Ordnung* determiniert sind. Siehe zu Kritik an dieser Position den folgenden Abschnitt.

34 *Matthias* (Fn.3), S. 55 f.

35 *Matthias* (Fn.3), S. 85. Ganz ähnlich argumentieren *Erhardt/Mona* (Fn.3), S. 75 f. im Hinblick auf ein Go-Programm.

36 Ebd.

37 *Matthias* (Fn.3), S. 56.

erster Ordnung besteht aber keine derartige Zweck-Mittel-Relation. *Wünsche zweiter Ordnung* haben *Wünsche erster Stufe* zum Inhalt und bewerten sie und sind daher von ganz anderer Art als diese. Mit *Wünschen zweiter Stufe* bringen wir zum Ausdruck, mit welchen unserer *Wünsche erster Ordnung* wir uns identifizieren und von welchen von ihnen wir uns distanzieren.[38] Unsere Kritik zielt hier wohlgemerkt nicht in erster Linie darauf ab, dass *Matthias* eine inadäquate Interpretation von *Frankfurts* Konzeption von *Wünschen zweiter Ordnung* zugrunde legt. Wir wollen vielmehr vor allem deutlich machen, dass *Matthias'* Verständnis von *Wünschen zweiter Ordnung* nicht geeignet ist, um Verantwortung zuzuschreiben. Denn Maschinen, die in seinem Sinne *Wünsche zweiter Ordnung* haben, können allenfalls – d. h. sofern man ihnen überhaupt intentionale Zustände zuschreiben will, was, wie wir bereits gesehen haben, überaus problematisch ist – Frankfurt'sche *Wantons* sein, also kritiklos nur denjenigen „Wünschen" folgen, die sich am stärksten in ihnen manifestieren.

Wir kommen damit zu dem Ergebnis, dass keines der drei hier vorgestellten Kriterien eine Grundlage dafür bietet, heutige KI im Schadensfall verantwortlich zu machen. Denn versteht man diese Kriterien so, wie *Matthias* sie interpretiert, vermögen sie keine Verantwortlichkeit zu begründen; fasst man sie dagegen so auf, dass sie für die Zuschreibung von Verantwortung geeignet sind, werden sie von KI in absehbarer Zeit nicht erfüllt werden können.

C. Der Zusammenhang zwischen Verantwortung[39], Freiheit und Vernunft

Der Begriff der *Verantwortung* ist kein isoliert zu betrachtendes Konzept, sondern steht in enger Beziehung zu den Begriffen *Freiheit* und *Vernunft* und, damit zusammenhängend, auch zu dem Begriff der *Handlung*.[40] Um zu klären, welche Bedingungen erfüllt sein müssen, um Verantwortung zuzuschreiben, sollen daher zunächst diese Begriffe näher erläutert werden.

Es ist naheliegend, genau diejenigen Handlungen (aber auch die ihnen zugrunde liegenden Entscheidungen) als vernünftig bzw. rational anzusehen, für die alles in allem betrachtet *gute*

---

38 Vgl. dazu *Frankfurt* (Fn.32), S. 13, der dort diesen Zusammenhang am Beispiel eines ‚widerwilligen Drogensüchtigen' *(unwilling addict)* erläutert, der einerseits zwei konfligierende *Wünsche erster Ordnung* hat – nämlich den Wunsch, die Droge zu konsumieren, und den Wunsch, sich davon zurückzuhalten, sie zu nehmen –, andererseits zugleich aber auch den *Wunsch zweiter Ordnung,* dass der zweite und nicht der erste dieser beiden Wünsche handlungsleitend, also sein Wille werde. Man kann sich allerdings fragen, ob es sich bei dieser Art von Identifikation und Distanzierung tatsächlich um einen *Wunsch* handelt. Uns erscheint es jedenfalls angebrachter zu sein, hier von einer Stellungnahme zu sprechen, die wir bezüglich konativer Einstellungen erster Ordnung (unsere Neigungen, Bedürfnisse und Interessen) abgeben und die als Teil eines Deliberationsprozesses (in dem die Gründe, die für oder gegen eine Handlung sprechen, gegeneinander abgewogen werden) letztlich unser Handeln bestimmt. Vgl. dazu ausführlich *Nida-Rümelin* 2005 (Fn.30), S. 87–92.

39 Die folgenden Überlegungen beziehen sich ausschließlich auf *persönliche* Verantwortung. *Politische* Verantwortung kann dagegen auch ohne persönliches Fehlverhalten zugeschrieben werden. Um eine effektive öffentliche Kontrolle zu gewährleisten, hat dabei eine Ministerin oder ein Minister in letzter Instanz für alle Entscheidungen des von ihr oder ihm geleiteten Ministeriums einzustehen. Diese Art von Verantwortlichkeit beruht dabei weitgehend auf einer Fiktion, denn angesichts der großen Anzahl von täglich innerhalb eines Ministeriums zu verzeichnenden Einzelvorgängen ist eine echte Einzelfallüberprüfung von der Ministerin praktisch gar nicht zu leisten. Vgl. dazu genauer *Nida-Rümelin* (Fn.22), S. 147 ff.

40 Vgl. *Nida-Rümelin* (Fn.22), S. 19–33 und S. 53.

Gründe sprechen;[41] denn Sätze wie: ‚Es ist vernünftig/rational, die Handlung h zu vollziehen, aber alles in allem betrachtet sprechen gute Gründe dagegen, h zu tun' bzw. ‚Es ist unvernünftig/irrational, die Handlung h zu vollziehen, aber alles in allem betrachtet sprechen gute Gründe dafür, h zu tun', sind schon rein sprachlich gesehen überaus irritierend. Vernunft lässt sich vor diesem Hintergrund als die Fähigkeit charakterisieren, die Gründe, an denen sich unsere Handlungen, Überzeugungen und Einstellungen orientieren, angemessen zu deliberieren.[42] Freiheit ist dann die Möglichkeit, genau den in einem solchen Abwägungsprozess für besser befundenen Gründen auch folgen zu können; wenn ich frei bin, sind es also jeweils *meine* durch Deliberation ermittelten Gründe, die mich leiten, so oder so zu urteilen und zu handeln.[43]

Aber was sind *Gründe?*[44] Liegt ein Unfallopfer schwer verletzt und ohne Hilfe am Straßenrand, dann hat man einen Grund, diesem zu helfen (z. B. durch Erste Hilfe oder indem man einen Krankenwagen ruft). Oder wenn Peter Fritz verspricht, ihm am kommenden Wochenende beim Umzug zu helfen, dann hat Peter einen Grund, dies am kommenden Wochenende auch zu tun. Es mag Umstände geben, die dagegensprechen; diese Umstände sind dann aber selbst wiederum Gründe, nämlich eben gewichtigere Gründe, wie z. B. der Grund, dass Peters Mutter am Wochenende seiner Hilfe bedarf, weil sie schwer erkrankt ist. Aber ein Versprechen gegeben zu haben ist zunächst einmal – zumindest in aller Regel – für sich genommen ein Grund, gemäß dem Versprechen zu handeln.[45] An diesen beiden Beispielen kann man zwei wesentliche Merkmale von Gründen erkennen. Zum einen, dass Gründe *normativ* sind, denn wenn ein Grund für eine Handlung spricht, dann *soll* man diese Handlung auch

41 Vgl. *Nida-Rümelin* 2020 (Fn.30), S. 22 f. und S. 325.

42 Die hier vertretene Gründe-basierte Vernunftkonzeption ist zu unterscheiden von einem rein instrumentellen Vernunftverständnis im Sinne von „Zweckrationalität", demzufolge eine Handlung genau dann rational ist, wenn sie geeignet ist, die mit der Handlung verfolgten Ziele zu erreichen. Denn es gibt zahlreiche Handlungen, die die Ziele der handelnden Personen zwar optimal verwirklichen, gegen deren Ausführung aber die besten Gründe sprechen und die wir daher als unvernünftig bezeichnen können. So wird man z. B. insbesondere die von den Nazis begangenen Verbrechen auch dann als verwerflich brandmarken wollen, wenn ihre Präferenzen durch diese Taten optimal erfüllt worden sind; auch kann kein Zweifel bestehen, dass die besten Gründe dagegen sprechen, das zu tun, was die Nazis getan haben. Der denkbare Einwand, hier würden Vernunft/Rationalität und Moral unzulässigerweise in eins gesetzt, da die Taten der Nazis zwar klarerweise *moralisch* falsch waren, aber im Sinne der Präferenzen der Täter eben doch möglicherweise rational, geht hier fehl. Rationale Handlungen *soll* man ausführen, irrationale hingegen nicht (eine Aussage wie: ‚Deine Vorgehensweise ist völlig irrational' ist klar als Vorwurf formuliert). Dieser Sollens-Charakter (un)vernüftiger/(ir)rationaler Handlungen spricht gegen eine Trennung zwischen Gründe-basierter Vernunft und Moral, denn eine moralisch gebotene Handlung kann zwar „zweckirrational", aber nicht unvernünftig sein (vgl. dazu auch *Nida-Rümelin* 2020 (Fn.30), S. 325).

43 Vgl. *Nida-Rümelin* 2020 (Fn.30), S. 384 f. Durch den Zusammenhang mit dem Gründe-geleiteten Abwägungsprozess wird deutlich, dass mit *Freiheit* hier nicht lediglich Handlungs(vollzugs)freiheit gemeint ist. Letztere liegt bereits dann vor, wenn der Akteur nicht durch äußere Hindernisse daran gehindert wird, das zu tun, was er will, und kann auch bei klar als unfrei zu qualifizierenden Zwangshandlungen von Geisteskranken oder schwer Süchtigen gegeben sein (vgl. dazu auch *Merkel, R.,* Willensfreiheit und rechtliche Schuld. Eine strafrechtsphilosophische Untersuchung, Baden-Baden 2008, S. 11 f.).

44 Aus Platzgründen beschränken wir uns im Folgenden auf praktische Gründe. Die hier vorgenommenen Charakterisierungen lassen sich jedoch auf theoretische Gründe (d. h. Gründe für theoretische Überzeugungen) übertragen. (vgl. dazu *Nida-Rümelin* 2020 (Fn.30), S. 332 ff. und S. 352 ff.).

45 Eine Ausnahme sind lediglich Versprechen, deren Erfüllung moralisch fragwürdig oder gar verboten ist (wie etwa das Versprechen, eine andere Person grausam umzubringen). Auch hier ist es dann aber wieder ein Grund, der gegen die Einhaltung des Versprechens spricht, nämlich genau sein moralisch fragwürdiger oder verbotener Inhalt.

ausführen, sofern nicht gewichtigere Gründe dagegen sprechen.[46] Und zum anderen, dass sie *objektiv* sind; damit ist hier gemeint, dass die Aussage, dass etwas ein *guter* Grund ist, sich nicht übersetzen lässt in Aussagen über mentale Zustände. So besteht der Grund, Fritz bei dem versprochenen Umzug zu helfen, für Peter auch dann noch, wenn er dazu auf einmal keine Lust mehr hat. Und der Grund, dem Unfallopfer zu helfen, verschwindet ebenfalls nicht dadurch, dass man gerade andere Präferenzen hat, oder weil man z. B. der kruden Überzeugung ist, dass Unfallopfer verdiene keine Hilfe. Subjektive Gründe gibt es genauso wenig wie subjektive Tatsachen![47]

Was bedeutet dieses Gründe-basierte Vernunft- und Freiheitsverständnis nun für den Verantwortungsbegriff? Verantwortung setzt sowohl auf Handlungs- wie auf Willens- bzw. Entscheidungsebene zumindest die Freiheit voraus, die fragliche Handlung und die ihr zugrunde liegende Entscheidung auch unterlassen zu können.[48] Der sog. Semi-Kompatibilismus bestreitet dies und vertritt demgegenüber die These, dass Verantwortung auch ohne Freiheit – und damit ist hier gemeint auch unter der Annahme eines umfassenden Determinismus – möglich ist.[49] Diese Position geht maßgeblich auf zwei Ende der 1960er- bzw. Anfang der 1970er-Jahre veröffentlichte Aufsätze des amerikanischen Philosophen *Harry G. Frankfurt* zurück,[50] die die Debatte bis heute prägen. Die *Frankfurt-Type-Examples,* die im Anschluss an die von *Frankfurt* dort angeführten Szenarien entwickelt wurden, sollen zeigen, dass eine Person auch dann für ihre Entscheidung moralisch verantwortlich ist, wenn sie faktisch gar keine andere Möglichkeit hatte, als sich so zu entscheiden, wie sie sich entschieden hat. Bei diesen Gedankenexperimenten gewährleistet jeweils eine andere Person – z. B. ein Neurochirurg, der über eine spezielle Computervorrichtung die Entwicklung der Absichten der Versuchsperson, die sich in entsprechenden Bereitschaftspotenzialen niederschlagen, verfolgen und beeinflussen kann –, dass die Entscheidung nur im Sinne einer von ihr vorher festgelegten Alternative (Tun oder Unterlassen) gefällt und umgesetzt werden kann. Entscheidet die Versuchsperson sich dann für diese Alternative, dann ist sie für diese Entscheidung verantwortlich, obwohl ihr aufgrund der Interventionsmöglichkeit der anderen Person gar keine andere Entscheidungsalternative offenstand; dass die fehlende Möglichkeit, sich

---

46 Mit der Normativität von Gründen hängt eng ihre Inferenzialität zusammen, die es ermöglicht, von empirischen Tatsachen auf normative Tatsachen zu schließen: Die empirische Tatsache, dass ein schwer verletztes Opfer hilflos am Straßenrand liegt, spricht dafür, der Person zu helfen (normative Tatsache), weil sie sonst bleibende körperliche Schäden davontragen oder sogar sterben könnte (Inferenz). Siehe zur Inferenzialität von Gründen genauer *Nida-Rümelin* 2020 (Fn.30), S. 335 f.

47 Vgl. dazu *Nida-Rümelin* 2020 (Fn.30), S. 344 ff. Das heißt nicht, dass subjektive Elemente wie Wünsche, Vorlieben oder Entscheidungen für die Beurteilung, ob etwas ein guter Grund ist, irrelevant wären. Und was für die eine Person in einer bestimmten Situation ein guter Grund ist, muss für eine andere Person, die sich in derselben Situation befindet, aber eben andere Präferenzen hat, nicht unbedingt ein guter Grund sein. Aus dem bloßen Umstand, dass eine Person etwas wünscht oder entscheidet, folgt aber nicht, dass sie einen guten Grund hat, den Wunsch oder die Entscheidung auch umzusetzen. Ob dazu ein guter Grund besteht, hängt nämlich vom Inhalt dieses Wunsches bzw. dieser Entscheidung ab, und die Beurteilung dieses Inhalts erfolgt selbst eben nicht nach subjektiven Kriterien.

48 Vgl. dazu ausführlich *Nida-Rümelin* 2005 (Fn.30), S. 79–105. In Bezug auf Handlungen spricht man hier vom sog. „principle of alternat(iv)e possibilities (PAP)": „a person is morally responsible for what she has done only if she could have done otherwise" (so *Frankfurts* Charakterisierung des von ihm abgelehnten Prinzips in: *Frankfurt, H. G.,* Alternate Possibilities and Moral Responsibility, in: Journal of Philosophy 66 (1969), S. 829).

49 Der Semi-Kompatibilismus steht also für die Vereinbarkeit von *Verantwortung* und *Determinismus*. Anders als der Kompatibilismus, der die Vereinbarkeit von *Freiheit* und *Determinismus* vertritt, nimmt der Semi-Kompatibilismus bezüglich dieser Frage eine agnostische Haltung ein.

50 Vgl. *Frankfurt, H. G.,* Alternate Possibilities and Moral Responsibility, in: Journal of Philosophy 66 (1969), S. 829–839; *ders.,* Freedom of the Will and the Concept of a Person, in: Journal of Philosophy 68 (1971) S. 5–20.

anders entscheiden zu können, für die Frage nach der Verantwortlichkeit irrelevant sei, sieht man aus semi-kompatibilistischer Perspektive auch daran, dass die Versuchsperson sich im Falle der Entscheidungsfreiheit (also ohne Interventionsmöglichkeit von außen) genauso entschieden hätte. Dies zeige, dass Verantwortung weder Handlungs- noch Willensfreiheit voraussetze. Diese Argumentation übersieht jedoch, dass wir der Versuchsperson in dem hier geschilderten Szenario eben doch nur deswegen die Verantwortung zuschreiben, weil sie sich von sich aus zwischen zwei Möglichkeiten (Tun oder Unterlassen) für eine Alternative entschieden hat, sie also über Entscheidungsfreiheit verfügte. Maßgeblich für die Frage nach der Verantwortung ist offensichtlich, zu welchem Zeitpunkt der Neurochirurg eingreift: Erfolgt die Intervention erst nach der Entscheidung der Versuchsperson, dann verfügte sie über Entscheidungsfreiheit zwischen zwei Alternativen und ist genau deswegen auch verantwortlich; erfolgt sie hingegen zu einem Zeitpunkt, zu dem sich die Versuchsperson noch in einem Abwägungsprozess befindet, und damit also bevor sie sich entschieden hat, dann ist sie nicht verantwortlich, denn die am Ende stehende Entscheidung wurde dann gar nicht von ihr herbeigeführt, sondern beruht auf einer Manipulation durch den Neurochirurgen.[51] Die *Frankfurt-Type-Examples* widerlegen also nicht, dass Freiheit eine Voraussetzung für Verantwortlichkeit darstellt.

Aber sind denn unsere Entscheidungen und Handlungen wirklich frei? Handlungen unterscheiden sich von bloßem Verhalten in mehrerlei Hinsicht. Verliert bei einer Busfahrt die Insassin I infolge einer Vollbremsung derart das Gleichgewicht, dass sie auf die Person P fällt und diese sich dabei verletzt, beschreibt und bewertet man das anders, als wenn I sich von sich aus auf P fallen lässt und P dabei die gleiche Art von Verletzung davonträgt. Nur im zweiten Fall schreiben wir I Intentionen zu und nur hier würden wir von einer Handlung sprechen; im ersten Fall dagegen von einem unabsichtlichen, unwillkürlichen und damit gerade nicht von ihren Intentionen gesteuerten Verhalten. Handlungen weisen also offensichtlich neben einer rein raum-zeitlichen Verhaltenskomponente das Merkmal der *Intentionalität* auf.[52]

Eine weitere Eigenschaft von Handlungen besteht darin, dass sie Gründe-geleitet sind, d. h. dass die handelnde Person immer einen *Grund* bzw. *Gründe* für ihre Handlung hat.[53] Handlungen werden in dieser Hinsicht konstituiert durch Gründe; nicht unbedingt durch gute Gründe, aber sie erfolgen eben nie völlig grundlos. Damit eignet ihnen aber auch zwingend immer ein Element der Rationalität, zumindest in dem Sinne, dass man – anders als bei bloßem Verhalten, bei dem sich diese Frage gar nicht stellt – immer beurteilen kann, ob eine

---

51 Siehe zu diesem Einwand im Detail *Nida-Rümelin* 2005 (Fn.30), S. 102 f.; vgl. zur fehlenden Überzeugungskraft einiger weiterer Varianten von *Frankfurt-Type-Examples* sowie außerdem zu deren strafrechtlicher Einordnung auch *Merkel* (Fn.43), S. 96–104.

52 Dabei ist zu differenzieren zwischen der Handlung *vorausgehenden* und *handlungsbegleitenden* Intentionen. Erstere werden durch die Handlung erfüllt und bringen einen (wenn auch oft nur rudimentären) Deliberationsprozess in Form einer Entscheidung zum Abschluss (siehe dazu gleich die anschließenden Erläuterungen im Text); Letztere stehen für die Absichtlichkeit des die Handlung realisierenden Verhaltens. Daneben gibt es auch noch *handlungsmotivierende* Absichten, die sich auf die kausalen Folgen und die strukturelle Rolle der Handlung beziehen (vgl. hierzu *Nida-Rümelin, J.,* Kritik des Konsequentialismus, München 1993, §§ 6–7 (S. 29–35); *ders.* 2005 (Fn.30), S. 51–60). Wenn im Alltag die Begriffe „Handlung" und „Verhalten" in der Weise synonym verwendet werden, dass auch *Handlungen* als „Verhalten" bezeichnet werden (z. B. in Formulierungen wie ‚Jetzt erkläre mir mal dein seltsames Verhalten von gestern Abend! '), dann ist darunter – korrekterweise – *intentionales Verhalten* zu verstehen.

53 In aller Regel kann die handelnde Person den Grund auf Nachfrage auch angeben. Selbst wenn der Grund ihr aber entfallen sein sollte – z. B. aufgrund eines unfallbedingten Gedächtnisverlusts –, ändert das nichts an seinem Vorhandensein zum Zeitpunkt der Handlung.

Handlung rational ist oder nicht; Handlungen sind, so könnte man sagen, „rationalitätsfähig“; denn wie wir gesehen haben, ist eine Handlung genau dann vernünftig, wenn insgesamt gute Gründe für sie, und unvernünftig, wenn insgesamt gute Gründe gegen sie sprechen. Von welchen Gründen wir uns leiten lassen, ist dabei das Ergebnis eines (manchmal auch sehr kurzen) Deliberationsprozesses, bei dem die unterschiedlichen Gründe gegeneinander abgewogen werden und der, wenn er abgeschlossen ist (und nur dann!), in eine Entscheidung mündet, die dann durch eine Handlung realisiert wird. Verkürzt lässt sich daher sagen: ‚Keine Handlung ohne Entscheidung‘[54]. Die jeweilige Entscheidung ist dabei in dem Sinne notwendig frei, dass schon rein begrifflich ausgeschlossen ist, dass sie vor dem Abschluss des Abwägungsprozesses bereits feststeht, denn es gehört schlicht zum Wesen von Entscheidungen, dass es, bevor die Entscheidung getroffen wurde, tatsächlich etwas zu entscheiden gab. Eine Entscheidung, deren Inhalt schon feststeht, bevor sie getroffen wurde, ist von vornherein keine Entscheidung![55]

Es ist nun aber genau diese Fähigkeit, Gründe abzuwägen, d. h. die Fähigkeit zu Deliberation, die uns nicht nur zu rationalen Wesen, sondern auch verantwortlich macht.[56] Letzteres wird deutlich, wenn man sich vergegenwärtigt, dass zwar Handlungen, nicht aber bloßes Verhalten Gegenstand eines Vorwurfs sein können. Während wir bei einem durch bloßes Verhalten verursachten Schaden keine Vorwürfe erheben und uns mit einer rein *kausalen Erklärung* zufriedengeben (im obigen Beispiel: ‚Aufgrund der durch Vollbremsung auf sie wirkenden Kräfte ist I auf P gefallen, was zu einer Verletzung von P geführt hat‘), erwarten wir bei einer Handlung, die andere schädigt oder beeinträchtigt, eine Begründung und, sofern dies möglich ist, eine *Rechtfertigung,* und d. h. Gründe, die diese Handlung rechtfertigen. Rechtfertigen muss und kann man sich aber nur für etwas, für das man auch verantwortlich gemacht werden kann.

Es ist also der Umstand, dass Handlungen Gründe-geleitet sind, der uns für sie verantwortlich sein lässt. Das führt uns nun aber zu der allgemeineren Formulierung und zugleich der zentralen Aussage des hier vertretenen Verantwortungskonzepts, nämlich, *dass wir für genau das Verantwortung tragen, für das wir Gründe haben.*[57] Neben Handlungen sind dies unsere

---

54 Vgl. dazu ausführlich *Nida-Rümelin* 2005 (Fn.30), S. 45–60. Mit dieser deliberativen Handlungskonzeption geht eine Ablehnung des sog. *belief-desire*-Modells einher, das man auch als Standardtheorie der Handlungsmotivation bezeichnen kann. Ihm zufolge motivieren uns ausschließlich Wünsche zum Handeln. Überzeugungen spielen nur rein *instrumentell,* d. h. hinsichtlich der Wahl des geeigneten Mittels eine Rolle, das eingesetzt werden muss, um den jeweiligen Wunsch zu erfüllen. Die Wünsche sind uns dabei vorgegeben oder allenfalls hervorgegangen aus anderen grundlegenderen Wünschen und entziehen sich daher jeder Kritik. Abgesehen von seiner rein „zweckrationalen“ Ausrichtung (vgl. dazu oben die Kritik in Fn.9) spricht gegen dieses (von *D. Hume* zumindest inspirierte) Modell vor allem, dass es die handlungsmotivierende Rolle *normativer Überzeugungen* verkennt. Insbesondere vermag das *belief-desire*-Modell nicht zu erklären, warum wir manchmal unseren momentanen Neigungen zugunsten längerfristiger (aber gerade nicht in Form eines Wunsches aktualisierter) Interessen *nicht* folgen. Vgl. zu diesem „Argument der intertemporalen Koordination“ und den weiteren hier vorgebrachten Einwänden *Nida-Rümelin, J.,* Strukturelle Rationalität. Ein philosophischer Essay über praktische Vernunft, Stuttgart 2001, S. 32–38.

55 Vgl. dazu *Nida-Rümelin* 2005 (Fn.30), S. 49–51. Es ist daher völlig abwegig, wenn *Erhardt/Mona* in: *dies.* (Fn.3), S. 69, schreiben, ein Spamfilter treffe aufgrund seiner Überzeugungen Entscheidungen.

56 Vgl. *Nida-Rümelin* (Fn.22), S. 53.

57 Vgl. *Nida-Rümelin* (Fn.22), S. 17 und S. 53 sowie passim.

theoretischen und normativen Überzeugungen, aber auch unsere emotiven Einstellungen,[58] denn auch für sie können wir Gründe angeben und auch für sie müssen wir uns gegebenenfalls rechtfertigen.[59] Der begriffliche Zusammenhang zwischen Verantwortung, Freiheit und Vernunft lässt sich vor diesem Hintergrund folgendermaßen wiedergeben: Weil bzw. sofern wir vernünftig sind, d. h. über die Fähigkeit zu Deliberation verfügen, sind wir, indem wir diese Fähigkeit ausüben und unser Handeln und Urteilen danach ausrichten, frei, und nur weil und nur in dem Maße, in dem wir frei sind, können wir verantwortlich sein.

Aus dem Befund, dass unsere Praxis der Verantwortungszuschreibung ein bestimmtes Verständnis von Freiheit voraussetzt, folgt als solches selbstverständlich noch nicht, dass wir tatsächlich über diese Art von Freiheit verfügen. Es sei allerdings angemerkt, dass zumindest das Argument, gegen die Annahme menschlicher Freiheit spräche das auf einem allumfassenden Determinismus und einem universal geltenden Kausalprinzip basierende Weltbild der Naturwissenschaften, so nicht haltbar ist. So ist das Konzept umfassender kausaler Klärung, demzufolge alles, was geschieht, eine Ursache hat und durch einen gesetzmäßigen Ursache-Wirkung-Zusammenhang beschrieben werden kann, in der modernen Physik längst aufgegeben worden; und bereits die klassische Newton'sche Physik war aufgrund der in ihr auftretenden Singularitäten keineswegs deterministisch; erst recht gilt dies für die moderne irreduzibel probabilistische Physik und mehr noch für die mit noch weit komplexeren Systemen befassten Disziplinen Biologie und Neurophysiologie.[60]

D. Verantwortung und Autonomie

Über was für eine Art von Autonomie muss nun eine Entität verfügen, um in dem im vorigen Abschnitt erläuterten Sinne verantwortlich sein zu können? Dabei ist zunächst zu berücksichtigen, dass der Begriff der Autonomie von Disziplin zu Disziplin sehr unterschiedlich verwendet wird. So versteht man in technischer Hinsicht unter der Autonomie einer Maschine häufig ihre Fähigkeit zu lernen oder ihre vollständige Automatisiertheit.[61] Aber auch innerhalb der Disziplinen finden sich divergierende Auffassungen darüber, was unter Autonomie zu verstehen sei.[62] Eine Auseinandersetzung mit dieser Vielzahl an unterschiedlichen Konzepten kann hier schon aus Platzgründen nicht geleistet werden. Stattdessen bietet es sich an, zwischen

---

58 Die Verantwortlichkeit für unsere emotiven Einstellungen mag vielleicht zunächst überraschen. Auch sie entziehen sich jedoch nicht gänzlich einer Begründungspflicht. So zeigen wir uns befremdet, wenn eine Person für die negativen Gefühle (wie z. B. Hass), die sie gegenüber einer anderen Person hegt, keine nachvollziehbaren Gründe anführen kann. Vgl. dazu *Nida-Rümelin* (Fn.22), S.48–52; *ders.* 2020 (Fn.30), S. 372–375.

59 Vgl. dazu *Nida-Rümelin* (Fn.22), S. 33–52.

60 Siehe dazu sowie zur Frage nach der Vereinbarkeit von menschlicher Freiheit und naturwissenschaftlicher Erklärbarkeit *Nida-Rümelin* 2005 (Fn.30), S. 69–78; *ders.* 2020 (Fn.30), S. 402 ff.

61 Vgl. dazu *Zech, H.,* Risiken digitaler Systeme: Robotik, Lernfähigkeit und Vernetzung als aktuelle Herausforderungen für das Recht, in: Weizenbaum Series #2, Aufsatz Februar 2020, S. 1 (S. 27 und S. 37 f.), → **www.weizenbaum-institut.de/media/Publikationen/Weizenbaum_Series/Weizenbaum_Series_2_Zech_070220.pdf** (zuletzt abgerufen am 07.05.2020).

62 Vgl. für den juristischen Bereich die Übersicht bei *Zech* (Fn.61), S. 38 f./Fn.83.

zwei Formen von Autonomie zu differenzieren: *Autonomie im starken Sinn* und *Autonomie im schwachen Sinn*.[63]

Autonomie im starken Sinn besteht in der Fähigkeit des Akteurs, sich *selbst* Ziele zu setzen und seine Handlungen im Hinblick auf diese Ziele auszurichten. Die Ziele sind dabei also nicht von außen vorgegeben. Sie entsprechen aber auch nicht einfach Wünschen oder Neigungen, sondern sind das Ergebnis eines Entscheidungsprozesses. *Autonomie im schwachen* Sinn liegt hingegen vor, wenn ein konkretes Verhalten zwar nicht durch die Intervention eines externen Akteurs bestimmt wird, dieser jedoch das übergeordnete und zu verfolgende Ziel festlegt. Schwache Autonomie manifestiert sich also nicht in der Wahl selbst gesetzter (übergeordneter) Ziele, sondern allenfalls in der Wahl der geeigneten Mittel, wie die extern vorgegebenen Ziele erreicht werden können. Man könnte daher auch von ‚heteronomer Autonomie'[64] sprechen.

Es ist nun offensichtlich, dass die hier vertretene Gründe-basierte Konzeption starke Autonomie voraussetzt, denn nur sie gewährleistet die für die Zuschreibung von Verantwortung erforderliche Selbstbestimmtheit des eigenen Handelns. Es stellt sich damit die Frage, ob Maschinen bereits heute oder in absehbarer Zukunft im starken Sinn autonom sind. Von Bedeutung sind dabei vor allem solche Anwendungen, bei denen es zu unvorhersehbarem Verhalten von Maschinen kommen kann. Das gilt insbesondere für den Bereich des sog. unüberwachten Lernens.[65] Die Unvorhersehbarkeit eines Verhaltens ist nun allerdings weder eine notwendige noch eine hinreichende Bedingung für Autonomie (und zwar weder für starke noch für schwache Autonomie!). So können etwa bei Paaren, die schon länger zusammenleben, die PartnerInnen häufig mit großer Treffsicherheit vorhersagen, wie die jeweils andere Person in bestimmten Situationen reagieren wird, ohne dass man daher dieser Person Autonomie absprechen würde. Und daraus, dass leider nicht genau vorhersehbar ist, wann der Vesuv das nächste Mal ausbrechen wird, lässt sich nicht ableiten, dass er im schwachen oder gar starken Sinn autonom wäre. Nun könnte man in Bezug auf den Vesuv möglicherweise einwenden, dass hier die Unvorhersehbarkeit nur auf der mangelnden Verfügbarkeit relevanter Daten beruht und dass bei vollständiger Kenntnis sämtlicher dieser Daten der nächste Ausbruch sehr wohl vorhersehbar ist. Aber auch eine „echte" Unvorhersehbarkeit, die nicht auf unvollständige Informationen zurückzuführen ist, impliziert noch keine Autonomie. So folgt z. B. daraus, dass es nicht möglich ist, den Zeitpunkt vorherzusagen, zu dem das nächste radioaktive Atom aus einer Stoffmenge zerfällt, nicht dessen Autonomie; wir haben es hier vielmehr mit einem objektiven (d. h. nicht auf epistemischen Mängeln, sondern auf Indeterminiertheit beruhenden) Zufall zu tun.

Unvorhersehbarkeit maschinellen Verhaltens spricht also nicht für die Zuschreibung von Autonomie im starken Sinn. Soweit eine KI über die Fähigkeit verfügt, die für die Erreichung eines vorgegebenen Ziels geeignetste Verhaltensalternative auszuwählen, könnte man dies

---

63 Vgl. dazu *Bertolini, A.,* Robots and Liability – Justifying a Change in Perspective, in: Battaglia/Mukerji/Nida-Rümelin (Fn.2), S. 150 f. in Anschluss an *Gutmann, M./Rathgeber, B./Syed, T.,* Action and Autonomy: A Hidden Dilemma in Artificial Autonomous Systems, in: Decker, M./Gutmann, M. (Hrsg.), Robo- and Informationethics. Some Fundamentals, Zürich, Berlin 2012, S. 245 ff.

64 Vgl. dazu *Bertolini* (Fn.63), S. 151 in Anschluss an *Gutmann/Rathgeber/Syed* (Fn.63), S. 246 f.

65 Vgl. dazu *Zech* (Fn.61), S. 36 f. und S. 46 f. Ganz allgemein spricht man bei Risiken, die mit der Lernfähigkeit von KI verbunden sind, von einem „Autonomierisiko" (vgl. *ders.* (Fn.61), S. 44).

dagegen als Autonomie im schwachen Sinn interpretieren. Der fundamentale Unterschied zwischen starker und schwacher Autonomie sei hier noch mal anhand eines Beispiels verdeutlicht: Georg übt seit mehreren Jahren zufrieden den Beruf des Krankenpflegers aus. Mit der Zeit stören ihn aber bestimmte Unzulänglichkeiten des Gesundheitssystems und er beschließt, sein Hobby zu seinem neuen Beruf zu machen und sich zum Gärtner umschulen zu lassen. Auch dieser Tätigkeit widmet er sich dann wieder mit viel Engagement. Eine solche Geschichte wird man über eine Maschine wohl auf absehbare Zeit nicht erzählen können. Denn auch ein noch so versierter Pflegeroboter mag zwar die Tätigkeit, für die er konzipiert und trainiert wurde, mit hoher Qualität und Zuverlässigkeit ausführen, aber er wird sich nicht von sich aus *entscheiden,* einen neuen Beruf zu erlernen. Vielleicht kann man ihn zum Gartenroboter umprogrammieren oder ihn die neue Tätigkeit sogar unbeaufsichtigt erlernen lassen; er kann sich aber dieses neue übergeordnete Ziel nicht im Sinne starker Autonomie *selbst* geben. Genau diese Fähigkeit ist aber, wie wir gesehen haben, für die Zuschreibung von Verantwortung erforderlich!

Aber, so ein denkbarer Einwand, verfügen wir denn tatsächlich über diese Fähigkeit? Ist sie nicht bloß eine Illusion? Sind wir nicht einfach nur komplexere Roboter, die sich lediglich einbilden, sie seien frei und zu einer Gründe-basierten Vernunft fähig? Meinen wir vielleicht nur, dass wir in dem im vorigen Abschnitt charakterisierten Sinn *handeln* und *Entscheidungen* treffen? Da wir uns selbst als Akteure begreifen, die zumindest über ein gewisses Maß an Autonomie im starken Sinn und Freiheit verfügen und dieses Selbstbild in uns sehr tief verankert ist, liegt die Beweislast auf der Seite desjenigen, der einen solchen Einwand formuliert. Er muss also Gründe dafür anführen, warum unsere Selbstwahrnehmung falsch ist. Dabei genügt es nicht, dass er darauf verweist, es sei eben evolutionär vorteilhaft gewesen, ein solches Selbstbild auch unabhängig von dessen Richtigkeit auszuprägen. Denn allein daraus, dass es auch dann evolutionär vorteilhaft sein kann, eine bestimmte Vorstellung zu haben, wenn diese falsch ist, folgt nicht, dass sie tatsächlich falsch ist. Um zu zeigen, dass wir uns in unser Selbstwahrnehmung als freie und autonome Akteure täuschen, bedarf es also weiterer Argumente. Wie wir bereits gesehen haben, vermag zumindest das in diesem Zusammenhang häufig vorgebrachte Argument eines allumfassenden Determinismus schon allein in naturwissenschaftlicher Hinsicht nicht zu überzeugen. Selbst wenn sich solche Argumente aber finden lassen, würde dies nichts an der hier von uns vertretenen Kernthese ändern, dass die Zuschreibung von Verantwortung ein Gründe-basiertes Vernunft- und Freiheitsverständnis voraussetzt. Sollte sich also herausstellen, dass wir selbst gar nicht über diese Art von Vernunft und Freiheit verfügen, dann müssten wir eben aufhören, uns im moralischen Sinn als Verantwortungssubjekte zu betrachten.[66] An unserer Haltung gegenüber Maschinen bräuchten wir hingegen nichts zu ändern. Es könnten dann eben weder Maschinen noch Menschen für ihr jeweiliges Verhalten verantwortlich gemacht werden.

66 Das bedeutet allerdings nicht unbedingt, dass dann im Bereich des Strafrechts das Schuldprinzip zugunsten eines reinen Maßregelrechts aufgegeben werden müsste. Siehe dazu und ganz allgemein zu den Konsequenzen und Problemen, die sich für das Strafrecht aus einer solchen mangelnden Verantwortungsfähigkeit ergeben würden, die überaus differenzierte Darstellung bei *Merkel* (Fn.43), S. 110–136.

E. Zusammenfassung

Unsere Untersuchung hat ergeben, dass Verantwortung nur Entitäten zugeschrieben werden kann, die über Vernunftfähigkeit und Freiheit verfügen. Vernunft lässt sich dabei als Vermögen auffassen, die Gründe, an denen sich unsere Handlungen, Überzeugungen und Einstellungen orientieren, angemessen zu deliberieren. Freiheit zeigt sich dann in der Möglichkeit, genau den in einem solchen Abwägungsprozess für besser befundenen Gründen auch folgen zu können. Verantwortung tragen wir dementsprechend nur für das, wofür wir Gründe haben; das sind unsere Handlungen, Überzeugungen und unsere emotiven Einstellungen. Diese Verantwortungskonzeption setzt eine Autonomie im starken Sinn voraus, d. h. die Fähigkeit, sich selbst Ziele zu setzen und entsprechend zu handeln. Auch komplexe KI-Systeme sind dazu auf absehbare Zeit nicht in der Lage. Gegen *Matthias* sind daher Forderungen nach einer Gesetzesänderung abzulehnen, lernfähige und ohne menschliche Intervention agierende Maschinen bei von ihnen verursachten Schäden bereits heute zivil- und strafrechtlich verantwortlich zu machen. Ebenso ist aus philosophischer Perspektive auch die Forderung des EU-Parlaments nach Einführung einer E-Person zurückzuweisen.[67]

67 Denkbar wäre allenfalls die Einführung einer E-Person aus rein pragmatischen Erwägungen. Wenn es aber gar nicht darum geht, Kriterien der Verantwortlichkeit zu formulieren und andererseits die Maschine selbst ja ohnehin keinen eigenen finanziellen Beitrag zu einem Hilfsfond leisten kann, stellt sich allerdings die Frage, warum es überhaupt einer pragmatisch begründeten Verantwortlichkeit von KI bedarf. Entsprechende Fondlösungen ließen sich schließlich grundsätzlich auch ohne Konstrukte wie die einer E-Person konzipieren. Vgl. gegen die Einführung einer E-Person aus juristischer Sicht (und dabei dezidiert auch gegen die These einer Analogie zur juristischen Person) auch *Kreutz, P.,* Autonomes Fahren: Produkt und Produzentenhaftung, in: Oppermann, B. H./Stender-Vorwachs, J. (Hrsg.), Autonomes Fahren. Technische Grundlagen. Rechtsprobleme. Rechtsfolgen, München 2020², S. 195–199.

bidt – Bayerisches Forschungsinstitut
für Digitale Transformation
Gabelsbergerstraße 4
80333 München
www.bidt.digital