Daniela Tafani

Dilemmata der Maschinen. Künstliche Intelligenz, Ethik und Recht*

> Von zentraler Bedeutung für das Projekt der Maschinenethik ist die Überzeugung (bzw. die Hoffnung), dass Ethik berechenbar gemacht werden könne, dass sie hinreichend verfeinert werden könne, um in eine Maschine programmiert werden zu können
>
> *Susan Leigh Anderson*, 2011

> Hier ist eine sehr menschliche Eigenschaft des einzelnen Wissenschaftlers am Werk. Wir nennen sie das Gesetz des Werkzeugs, und es lässt sich folgendermaßen formulieren: Gib einem Kind einen Hammer, und es wird entdecken, dass alles, was ihm begegnet, irgend eines Schlages bedarf. Es erstaunt daher überhaupt nicht, zu entdecken, dass ein Wissenschaftler die Probleme auf eine Weise formuliert, die für deren Lösung nur jene Techniken verlangt, in denen er selbst besonders befähigt ist.
>
> *Abraham Kaplan*, 1964

1. Das Dilemma des Eisenbahnwagens

Gedankenexperimente bilden für die philosophische Reflexion ein nützliches und starkes begriffliches Instrument, denn aufgrund ihres narrativen Charakters ermöglichen sie es, eine Lehre auf die Probe zu stellen, indem sie feststellen, ob deren extremste Ergebnisse als intuitiv akzeptabel erscheinen, und denjenigen herausfordern, der diese Lehre unterstützt, auch die Ergebnisse explizit zu unterschreiben[1]. Werden freilich Gedankenexperimente schlecht formuliert oder von ihrem ursprünglichen Kontext abgehoben und in einen ungeeigneten Zusammenhang verpflanzt, verwandeln sie sich in lächerliche Instrumente der Überredung, d.h. in abwegige und gefährliche Werkzeuge[2]. Dies ist, wie im Folgenden nachzuweisen versucht werden soll, der Fall beim

---

* Übersetzung aus dem Italienischen von *Thomas Vormbaum.*

1 Übersicht über die alternativen Auffassungen darüber, was ein Gedankenexperiment sei und wozu es dienen könne, b. *J. R. Brown / Y. Fehige / M. Stuart* (Hrsg.), Routledge Companion to Thought Experiments. Abingdon / New York (Routledge) 2018, insb. der Beitrag von *G. Brun*, Thought Experiments in Ethics, S. 195–210.

2 Vgl. *D. C. Dennett*, Intuition pumps and other tools for thinking. New York, London (Norton & Company) 2013, S. 2–3; it. Übers. von *S. Frediani* u.d.T. Strumenti per pensare. Mailand (Cortina) 2014, S. 3.

https://doi.org/10.1515/jajuz-2019-0017

Dilemma des Eisenbahnwagens (*trolley problem*) als ein Problem, das sich unumgänglich der neu entstandenen Disziplin der Maschinenethik stellt[3].

1967 hatte Herbert L. A. Hart in einem Artikel über die Verbindung zwischen Absicht, rechtlicher Verantwortlichkeit und Strafe bemerkt, dass die zivilisierten Strafrechtssysteme die Strafbarkeit daran knüpfen, dass eine äußerliche kriminelle Handlung absichtlich begangen worden ist, und dass sie diese Absicht mit dem Bewusstsein des Handelnden gleichsetzen, dass die schädliche Folge als das Ergebnis der eigenen willentlichen Handlung eingetreten ist (unabhängig davon, ob er tatsächlich diese Konsequenz als Mittel oder Ziel der eigenen Handlung verfolgt hat oder ob er sie nur als unerwünschte Nebenfolge vorausgesehen hat)[4]. Hart bemerkt, dass damit das Gesetz sich von dem unterscheidet, was man gewöhnlich meint, wenn man davon spricht, dass ein Mensch eine Handlung „absichtlich" begangen habe. Der Unterschied zwischen dem, was man zu tun beabsichtigte, und dem, was man als Folge der eigenen Handlung vorhergesehen hat, wird nämlich vom allgemeinen Empfinden und – mitunter in Übereinstimmung mit diesem – von der Philosophie erfasst; dies belegen beispielsweise die von Bentham eingeführte[5] Unterscheidung zwischen *direkter* Absicht und *obliquer* Absicht und – in der katholischen Moral – die Lehre von der „doppelten Wirkung"[6].

---

3 Die Ethik der Maschinen „befasst sich damit, den *Maschinen* ethische Prinzipien zu geben bzw. ein Verfahren zur Entdeckung eines Weges zur Lösung der ethischen Dilemmata, in die sie geraten kann, wobei ihnen ermöglicht wird, mittels eines eigenen Entscheidungsprozesses ethisch verantwortlich zu handeln" (*M. Anderson / S. L. Anderson* (Hrsg.) Machine Ethics. Cambridge (Cambridge University Press) 2011, S. 1). S. auch *W. Wallach / C. Allen*, Moral machines: Teaching Robots Right from Wrong. Oxford (Oxford University Press) 2009.

4 *H. L. A. Hart*, Intention and Punishment, in: Oxford Review IV, 1967, S. 5–22.

5 *J. Bentham*, An Introduction to the Principles of Morals and Legislation. 1781, Kap. VIII, 6, S. 86; it. Übers. von E. Lecaldano u.d.T. „Introduzione ai principi della morale e della legislazione". Turin (UTET) 1998, S. 180.

6 Die ursprüngliche Formulierung der Lehre vom doppelten Effekt wird allgemein auf Thomas von Aquin zurückgeführt, der seinerseits auf Augustinus verweist: „Sankt Augustinus sagt: 'Nie soll uns als Schuld angerechnet werden, wenn bei dem, was wir in guter Absicht und erlaubterweise tun, gegen unseren Willen etwas Schlechtes entsteht'. Nun kommt es tatsächlich vor, daß denen, die etwas Gutes tun, durch Zufall die Tötung eines Menschen unterläuft. Also wird ihnen dies nicht als Schuld angerechnet. […] Wer sich daher – [so das Recht] – mit erlaubten Dingen abgibt und dabei trotz aufgewandter Sorgfalt der Tod eintritt, entgeht dem Vorwurf schuldhafter Tötung. Beschäftigt er sich jedoch mit etwas Unerlaubtem oder auch mit etwas Erlaubtem, jedoch ohne die nötige Sorgfalt, dann ist er von Schuld nicht frei, falls sein Tun den Tod eines Menschen zur Folge hat". *Thomas von Aquino,* Recht und Gerechtigkeit Theologische Summe II /II, Fragen 57–79 Nachfolgefassung von Band 18 der Deutschen Thomasausgabe Neue Übersetzung von Josef F. Groner […] Quaestio 64, Artikel 8, S. 102.

Im selben Jahr hat Philippa Foot in der darauffolgenden Nummer derselben Zeitschrift, ausdrücklich von den Überlegungen Harts ausgehend, die „Lehre der doppelten Wirkung" untersucht, um die Plausibilität der These, dass es mitunter erlaubt sei, mit obliquer Absicht – d.h. bloß voraussehend – das herbeizuführen, was direkt anzustreben nicht erlaubt ist, als Unterscheidungskriterium dafür, was in Fällen, in denen die Interessen verschiedener Personen konfligieren, moralisch erlaubt sei, zu ermitteln[7]. Neben den von Hart formulierten Beispielen fand Foot es zweckmäßig, weitere klare Beispiele zu präsentieren – darunter die drei folgenden:

> Nehmen wir an, dass ein Richter oder Magistrat sich Antragstellern gegenüber sieht, die verlangen, dass eine Person, die eines bestimmten Verbrechens schuldig sei, gefunden werde, und dass sie drohen, andernfalls ihre Blutrache an einem bestimmten Viertel der Stadt zu vollziehen. Da der wahre Schuldige unbekannt ist, meint der Richter, er könne dem Blutvergießen nur dadurch begegnen, dass er eine unschuldige Person festnehmen und hinrichten lässt. Neben dieses Beispiel stellt sie ein weiteres, in dem ein Pilot, dessen Flugzeug vor dem Absturz steht, vor der Entscheidung steht, ob er vom Kurs über ein stärker besiedeltes Gegend auf ein weniger besiedeltes Gebiet ausweichen soll. Um die Parallele möglichst eng zu führen, kann man auch unterstellen, dass er der Führer einer außer Kontrolle geratenen Straßenbahn ist, der nur von einem geradeaus führenden Gleis auf ein anderes ausweichen kann; fünf Arbeiter stehen auf dem einen Gleis, einer auf dem anderen; in jedem Fall wird jeder, der sich auf dem von ihm befahrenen Gleis befindet, getötet werden.[8]

Aus der Analyse dieser und zahlreicher weiterer Beispiele zog Foot den Schluss, dass die Unterscheidung zwischen direkter Absicht und obliquer Absicht nicht entscheidend sei. Entscheidend dafür, die moralisch bedeutsamen Unterschiede in den verschiedenen Fällen erkennen zu können, sei vielmehr die Unterscheidung zwischen negativen Pflichten und positiven Pflichten, d.h. zwischen der Pflicht, bestimmte Handlungen zu unterlassen – z.B. Tötung und Diebstahl –, und den Pflichten zur Fürsorge bzw. zur Hilfe[9], sowie der Vorrang der Ersteren vor den Letzteren: im Falle des Richters

---

7 *P. Foot*, The Problem of Abortion and the Doctrine of the Double Effect, in: Oxford Review V, 1967, S. 5–15.

8 Ebd., S. 8.

9 Foot bezog sich auf die Formulierung von *J. Salmond*, Jurisprudence. London (Sweet & Maxwell), 11. Aufl. 1957, ohne zu erwähnen, dass die Auffassung vom Vorrang der negativen Pflichten vor den positiven Pflichten z.B. von Kant vertreten worden ist, und zwar als Vorrang der Pflichten der Gerechtigkeit, auch vollkommene Pflichten oder essentielle Pflichten genannt, vor den Pflichten des Gutseins (unvollkommene oder akzidentielle Pflichten); s. *W. Kersting*, Artikel „Pflichten, unvollkommene/ vollkommene", in: Historisches Wörterbuch der Philosophie, Bd. 7, hrsg. von K. Gründer. Basel, Stuttgart 1989, Sp. 433–439.

beispielsweise müsste die Pflicht, nicht einem Unschuldigen Schaden zuzufügen, den Vorrang vor der Pflicht haben, Hilfe zu leisten. Der Umfang des herbeigeführten Schadens bleibe nur in Fällen wie dem der außer Kontrolle geratenen Tram entscheidend, in dem der Konflikt ein solcher zwischen negativen Pflichten sei: „Wenn die Entscheidung darin besteht, ob man ein Unrecht gegenüber *einer* Person oder *mehreren* Personen begehen soll, scheint es für uns nur einen einzigen rationalen Handlungsverlauf zu geben“[10].

Das Dilemma der außer Kontrolle geratenen Straßenbahn wurde wenige Jahre später von Judith J. Johnson diskutiert und neu formuliert als „das Problem des Eisenbahnwagens [trolley] zu Ehren des Beispiels von Frau Foot“[11]. Als solches hat es in den folgenden Jahrzehnten reüssiert: zusammen mit den weiteren Gedankenexperimenten von Foot ist es nämlich zum Gegenstand einer dauerhaften Diskussion und Neudefinition im Bereich der Moralphilosophie geworden[12], und es ist als Test – mit dem Anspruch, es gelöst zu haben – in den Untersuchungen von Joshua Greene zur empirischen Moralpsychologie, die sich auf die Neurowissenschaften der Ethik stützen, verwendet worden[13].

Was Foot sich im Jahre 1967 nicht vorgestellt hätte, ist, dass ihr Tram-Dilemma einige Jahrzehnte später nicht nur für ein Dilemma gehalten würde, das ein Fahrzeug zum Gegenstand hat, sondern auch als ein Dilemma für das Fahrzeug selbst in Fällen, in denen dieses selbst in Notstandssituationen konkrete Entscheidungen treffen soll.

---

10 *P. Foot*, The Problem of Abortion and the Doctrine of the Double Effect, a.a.O., S. 12.

11 *J. J. Thomson*, Killing, letting die and the trolley problem, in: The Monist LIX, 1976, S. 204–217.

12 Vgl. *F. M. Kamm*, The Trolley Problem Mysteries, with commentaries by J. J. Thomson, T. Hurka, S. Kagan; hrsg. von E. Rakowski. New York (Oxford University Press) 2016; populäre Einführung b. *D. Edmonds*, Would you kill the fat man? The trolley problem and what your answer tells us about right and wrong. Princeton (Princeton University Press) 2013; it. Üb. von *G. Guerrerio* u.d.T. „Uccideresti l’uomo grasso? Il dilemma etico del male minore“. Mailand (Cortina) 2014.

13 *J. D. Greene*, Solving the trolley problem, in: J. Sytsma / W. Buckwalter (Hrsg.), A Companion to Experimental Philosophy. Chichester (Wiley-Blackwell) 2016, S. 175–189; *Ders*., Beyond Point-and-Shoot Morality: Why Cognitive (Neuro) Science Matters for Ethics, in: Ethics CXXIV, 4, 2014, S. 695–726. Zu den normativen Folgen, die er aus seinen empirischen Beobachtungen zu ziehen beansprucht – die aber in Wirklichkeit nur aus deren Verbindung mit den axiologischen und normativen Positionen, die er dort vertritt, folgen – s. *A. S. Berker*, The normative insignificance of neuroscience, in: Philosophy & Public Affairs XXXVII, 2009, S. 293–329. Vgl. *A. Kauppinen*, Ethics and Empirical Psychology – Critical Remarks to Empirically Informed Ethics, in: M. Christen / C. van Schaik / J. Fischer / M. Huppenbauer / C. Tanner (Hrsg.), Empirically Informed Ethics: Morality between Facts and Norms. Cham (Springer) 2014, S. 279–305.

2. Automobile, die darüber entscheiden, wer getötet werden soll

Der Mythos der Sicherheit von selbstfahrenden Automobilen ist zusammen mit der massenhaften Verbreitung von Kraftfahrzeugen entstanden, nämlich 1935 in *The safest place*[14] – einem von der Abteilung Chevrolet der General Motors in Auftrag gegebenen Kurzfilm zur Straßenverkehrserziehung – wo ein ruhiger Herr, der zur Hause singt, mehrfach in Gefahr gerät, sich in einer Reihe von häuslichen Unfällen, denen er nur wie durch ein Wunder entgeht, das Genick zu brechen, bis er dann am Ende das Haus verlässt („gehen wir von hier fort" – sagt eine Stimme aus dem Off – „bevor jemand sich noch einen Schaden zufügt"). Dieselbe Stimme aus dem Off begleitet sodann ein elegantes Kraftfahrzeug, während es ohne Fahrer und ohne Passagiere sicher die Straßen befährt und den Zuschauer daran erinnert, dass dieser „Salon auf Rädern" der sicherste Ort der Welt wäre, wenn man ihn nur mit einem „Mechanismus für autonomes Lenken" ausstatten könnte: Unfälle geschähen nämlich einzig und allein aus der Unvernunft der menschlichen Fahrzeugführer, welche die Richtung, in die sie zu fahren beabsichtigten, für sich behalten, als wäre es ein Geheimnis, oder das Losfahren an einer Ampel mit dem Start zu einem Autorennen verwechseln.

Das in dem Kurzfilm von 1935 zur Schau gestellte unbedingte Vertrauen in die Unfehlbarkeit und damit in die Sicherheit der Technik zielte nicht nur darauf ab, die Kraftfahrzeugfahrer dazu zu erziehen, sich vorsichtig zu verhalten, sondern vor allem, sie zu beruhigen, da die Straßenverkehrsunfälle in den Vereinigten Staaten die Relevanz eines sozialen Problems angenommen hatten[15]. Ein vergleichbares triumphales Vertrauen (womöglich gestützt auf vergleichbare kommerzielle Gründe) begleitet heute die Einführung der ersten Fahrzeuge mit teilweise autonomer Lenkung[16] mit dem Versprechen einer

14 https://archive.org/details/SafestPl1935 (letzter Zugriff zu dieser und zu den weiteren zitierten websites: 18. April 2019).

15 *F. Kröger*, Das automatisierte Fahren im gesellschaftsgeschichtlichen und kulturwissenschaftlichen Kontext, in: Autonomes Fahren. Technische, rechtliche und gesellschaftliche Aspekte, hrsg. von M. Maurer / J. Gerdes / B. Lenz / H. Winner. Berlin, Heidelberg (Springer) 2015, S. 41–68.

16 Die Autonomie von Fahrzeugen ist von der Society of Automotive Engineers in sechs Stufen klassifiziert worden, von Stufe 0, auf der der menschliche Fahrzeugführer sämtliche Fahroperationen ausführt, bis zu Stufe 5, auf der die Lenkung vollständig automatisiert ist und der menschliche Besitzer des Fahrzeugs lediglich ein Passagier ist (https://www.sae.org/standards/content/j3016_201806/).

Verminderung der Straßenverkehrsunfälle um 90 Prozent[17], ohne dass im Allgemeinen die Möglichkeit berücksichtigt wird, dass es sich um eine bloße Ersetzung der menschlichen Irrtümer durch Irrtümer der Maschinen handelt[18]. Indem mit ideologischem Optimismus[19] weit über den gegenwärtigen Zustand der technischen Entwicklung von selbstfahrenden Automobilen hinausgeblickt wird – mit einem Blick, in dem tödliche Unfälle, bei denen z.B. das Fahrzeug einen weißen LKW mit einem Stück Himmel verwechselt hat und gegen ihn gekracht ist, keinen Platz haben[20] –, meinen viele, dass diese Fahrzeuge in der Lage seien, mit Notstandssituationen fertig zu werden, dank ihrer rechnerischen Fähigkeit und Schnelligkeit aufgrund von verfeinerten und augenblicklichen Analysen der Situation statt, wie die Menschen, mit instinktivem Improvisieren. Um seltenen, unvermeidlichen Unfällen begegnen zu können, sei es daher – so meint man – zweckmäßig, die Fahrzeuge im Voraus zu autonomen Führern zu programmieren, um zu entscheiden, wer in Fällen getötet werden soll, in denen ein tödlicher Schaden unvermeidbar ist und in denen feststeht, dass jedes der vom Fahrzeug vorgenommenen alternativen Manöver ein anderes Opfer treffen wird. Damit hat das Dilemma des Eisenbahnwagens anscheinend eine neue konkrete Dringlichkeit erlangt, und es hat im Rahmen der heftigen Debatte über künstliche Moral sogar einen eigenen Herrschaftsbereich gefunden: die „Ethik der Straßenverkehrsunfälle" bei autonom gelenkten Fahrzeugen[21].

Außer in den – der ursprünglichen Version des Dilemmas entsprechenden – Fällen, in denen die Alternative diejenige zwischen der Aufopferung eines

17 Vgl. die National Highway Traffic Safety Administration, eine Agentur des Transportministeriums der Vereinigten Staaten: https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety; s. auch *M. Bertoncello / D. Wee*, Ten ways autonomous driving could redefine the automotive world. McKinsey & Company: https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/ten-ways-autonomous-driving-could-redefine-the-automotive-world.

18 Die Frage behandelt hingegen *C. Misselhorn*, Grundfragen der Machinenethik. Ditzingen (Reclam) 2018, S. 198 f.

19 *V. F. Operto*, Elementi di roboetica. Analisi di alcune metaetiche nella progettazione e programmazione di veicoli autonomi, in: InCircolo VI, 2018, S. 89–108: 98.

20 Vgl. den Bericht von Tesla, der Herstellerfirma des in den Unfall verwickelten Fahrzeugs: https://www.tesla.com/it_IT/blog/tragic-loss.

21 S. z.B. aus der unübersehbaren Zahl der in weniger als einem Jahrzehnt erschienenen Artikel *P. Lin*, Why Ethics Matters for Autonomous Cars, in: Autonomes Fahren, a.a.O, S. 69–85; *G. Keeling*, Why Trolley Problems Matter for the Ethics of Automated Vehicles, in: Science and Engineering Ethics 2019, S. 1–15. Nützlicher Überblick b. *S. Nyholm*, The Ethics of Crashes with Self-Driving Cars: A Roadmap, I und II: in: Philosophy Compass, XIII, 7, 2018, https://onlinelibrary.wiley.Com/doi/epdf/10.1111/phc3.12506 und /phc3.12507.

Menschen oder mehrerer Menschen ist[22], wird angenommen, dass auch die Fälle betrachtet werden müssen, in denen die alternativen Opfer jeweils einzelne Personen sind:

> Stelle dir vor, dass in einer fernen Zukunft dein autonomes Kraftfahrzeug vor der schrecklichen Wahl steht: Soll ich nach rechts steuern und ein Kind von acht Jahren überfahren, oder nach rechts lenken und eine Großmutter von 80 Jahren überfahren? Wegen der Geschwindigkeit des Fahrzeugs würde mit Sicherheit jedes Opfer beim Aufprall getötet. Ändere ich gar nicht die Richtung, werden beide Opfer überfahren und getötet; es gibt daher einen guten Grund für die Meinung, dass du in die eine oder andere Richtung lenken solltest. Doch welches wäre die ethisch richtige Entscheidung? Wenn du dabei bist, das autonom gesteuerte Auto zu programmieren: was für ein Verhalten würdest du eingeben für den Fall, dass es einmal in eine derartige, freilich seltene, Situation geraten sollte?[23]

Ein Kraftfahrzeug so zu programmieren, dass es von zwei Personen diejenige wählt, mit denen es zusammenstößt, bedeutet zugleich, in die Fahrzeuge zivile Algorithmen einzubauen, die – wenn sie nicht in der Lage sind, weitere Handlungsmöglichkeiten zu berücksichtigen – darauf abgestimmt sind, eine Person statt einer anderen aufgrund von feststellbaren Merkmalen, die den Auswahlkriterien für Ziele im militärischen Bereich ähnlich sind, zu überfahren[24]. Doch wie soll man für jedes der möglichen Szenarien bestimmen, welches der beiden möglichen Opfer überfahren werden soll?

Einige schlagen vor, in Fällen, in denen noch keine allgemein anerkannten ethischen Prinzipien verfügbar sind, die Entscheidungsprozesse der autonom steuernden Fahrzeuge nach einem Berechnungsmodell der sozialen Entscheidung zu automatisieren, das in der Lage ist, die von den Menschen geäußerten Präferenzen für die Lösungen der „moralischen Dilemmata bei unvermeidbaren Unfällen“ zu aggregieren. Um diese Präferenzen zu sammeln, ist beim Media Lab des Massachusetts Institute of Technology eine Plattform für online-voting mit der Bezeichnung „Moral Machine“[25] eingerichtet worden, die fast 40 Millionen Entscheidungen aus 233 Ländern und Territorien gesammelt hat[26].

---

22 *J.-F. Bonnefon / A. Shariff / I. Rahwan*, The social dilemma of autonomous vehicles, in: Science CCCLII, 6293, 2016, S. 1573–1576.

23 *P. Lin*, Why Ethics Matters for Autonomous Cars, a.a.O., S. 69 ff.

24 Ebd., S. 72.

25 http://moralmachine.mit.edu.

26 *E. Awad / S. Dsouza / R. Kim / J. Schulz / J. Henrich / A. Shariff / J.-F. Bonnefon / I. Rahwan*, The Moral Machine experiment, in: Nature DLXIII, 2018, S. 59–64; *R. Noothigattu / S. S. Gaikwad / E. Awad / S. Dsouza / I. Rahwan / P. Ravikumar / A. D. Procaccia*, A Voting-Based System for Ethical Decision Making, 2017,

Die Plattform wird in der Form eines Videospiels präsentiert („ein multilinguales ‘ernstes Spiel’ online“[27]), das Unfall-Szenarien aufgrund von neun Faktoren generiert:

> Menschen schonen (statt Haustiere), Richtung beibehalten (statt ausweichen), Fahrzeuginsassen schonen (statt Fußgänger), mehrere Leben schonen (statt einer geringen Zahl von Leben), Männer schonen (statt Frauen), junge Menschen schonen (statt alte Personen), Fußgänger schonen, die legal die Straße überqueren (statt solche, die sie illegal überqueren), Personen schonen, die körperlich fit sind (statt Personen, die weniger gut in Form sind), diejenigen schonen, die einen hören sozialen Status haben (statt diejenigen mit einem niedrigeren sozialen Status).[28]

Die weltweiten Präferenzen derer, die geantwortet haben, sind eindeutig zugunsten der Kinder gegenüber Alten und – wenn auch nur geringfügig – der Frauen gegenüber Männern, der Geschäftsleute gegenüber Obdachlosen und der Sportler gegenüber übergewichtigen Personen. Dies bedeutet nicht – räumen die Autoren ein – dass die politischen Entscheidungsträger zwangsläufig diese Meinungen übernehmen sollen; sie sollten allerdings gut darauf vorbereitet sein, Entscheidungen zu rechtfertigen, die sich von der öffentlichen Meinung unterscheiden, wenn diese besonders starke Präferenzen betrifft wie diejenige zugunsten von Kindern[29].

Der Rückgriff auf die „öffentliche Moralität“[30] – d.h. auf eine Aggregierung auf der Grundlage eines sozialen Entscheidungsmodells der gesammelten individuellen Ansichten – wird als letztinstanzliches Modell präsentiert, das in Fällen angewendet werden soll, in denen es weder technische Lösungen gibt, um das Dilemma zu umgehen, noch rechtliche Bindungen oder moralische Wahrheiten, die das zu Tuende vorschreiben. Die These lautet, dass die entsprechend zusammengestellten Ergebnisse der empirischen Beobachtung der in einer Art von Videospiel simulierten moralischen Verhaltensweisen das gesetzgeberische Vakuum ausgleichen könnten, sowie das Fehlen – übrigens

---

https://arxiv.org/pdf/1709.06692.pdf. Die Vorstellung, dass die berechenbare soziale Entscheidung, d.h. die Verwendung von Algorithmen für das Aggregieren von individuellen Präferenzen, die Werkzeuge liefern könnte, um zu kollektiven ethischen Entscheidungen zu gelangen, wird entwickelt von *J. Greene / F. Rossi / J. Tasioulas / K. B. Venable / B. C. Williams*, Embedding Ethical Principles in Collective Decision Support Systems, in: D. Schuurmans / M. Wellmam (Hrsg.), Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence and the Twenty-Eighth Innovative Applications of Artificial Intelligence Conference. Phoenix (AAAI) 2016, Bd. VI, S. 4147–4151.

27 *E. Awad et al.*, The Moral Machine experiment, a.a.O., S. 59.

28 Ebd., S. 60.

29 Ibidem.

30 Ebd., S. 59.

trotz jahrhundertelanger Diskussion – einer allgemein anerkannten Auffassung und einer notwendigen Formalisierung der Prinzipien der normativen Ethik[31].

Im Ergebnis enthalten weder das zeitgenössische Recht noch irgend eine universell anerkannte Morallehre eine Werteskala für menschliche Leben, auf deren Grundlage man die entsprechenden Opfer-Algorithmen bestimmen könnte, welche die Hierarchie der Ziele autonom gesteuerter Automobile nach Geschlecht, Alter, körperlicher Verfassung oder Vermögen in Rechengrößen umsetzen. Es handelt sich freilich nicht um eine Rechtslücke: die Antwort auf die Frage, wer getötet werden soll – zwischen Alten und Kindern, zwischen Geschäftsleuten und Obdachlosen – findet sich nicht im Recht, weil schon die Frage verboten ist. Eine Werteskala von Leben würde ohne weiteres ermöglichen, einem System autonomen Lenkens klare und formalisierbare Instruktionen zu erteilen, doch stünde es im Widerspruch zu den Menschenrechten auf Leben und Nichtdiskriminierung, wie sie neben anderen in der Allgemeinen Erklärung der Menschenrechte, in der Grundrechtecharta der Europäischen Union und in vielen Ländern in den Verfassungsurkunden anerkannt sind[32].

Bei der Erörterung des Dilemmas zwischen einem Mädchen von acht Jahren und einer Großmutter von 80 Jahren bemerkt Patrick Lin, dass die Diskriminierung nach Alter vom Ethik-Kodex des Institute of Electrical and Electronics Engineers sowie von der deutschen Verfassung und vom XIV. Amendment der Verfassung der Vereinigten Staaten verboten sei (die er in dieser Reihenfolge untersucht), er meint allerdings – nachdem er das Dilemma in der Weise formuliert hat, dass beide Opfer überfahren werden, wenn man sich entscheidet, weder in die eine noch die andere Richtung auszuweichen – , dass es besser sei, nur eine Person, wenn auch aufgrund einer Diskriminierung, zu töten, als zwei zu töten[33].

Akzeptiert man den Ansatz von Lin – für den alle möglichen Dilemmata im Voraus, ohne jede Bewertung zur Legitimität der Dilemmata selbst, berücksichtigt und gelöst werden müssen – so muss man zu den Dilemmata auch jene

31 *R. Noothigattu et al.*, A Voting-Based System for Ethical Decision Making, a.a.O. Vgl. *L. R. Sütfeld / R. Gast / P. König / G. Pipa*, Using Virtual Reality to Assess Ethical Decisions in Road Traffic Scenarios: Applicability of Value-of-Life-Based Models and Influences of Time Pressure, in: Frontiers in behavioral neuroscience XI, 2017, doi:10.3389/fnbeh.2017.

32 An dieser Stelle genügt der Hinweis auf die Allgemeine Erklärung der Menschenrechte, Art. 2 und 3; die Grundrechtecharta der Europäischen Union nennt in Art. 21 („Nichtdiskriminierung“) ausdrücklich auch das Alter unter den Faktoren, auf die eine Unterscheidung nicht gestützt werden darf.

33 *P. Lin*, Why Ethics Matters for Autonomous Cars, a.a.O., S. 70.

zwischen der Tötung eines weißen Kindes und eines schwarzen Kindes hinzunehmen, ebenso dasjenige zwischen der Tötung eines Katholiken und eines Juden, und sodann den Anspruch erheben, dass es sich um ein moralisch relevantes Dilemma handele und auf die Ergebnisse der nächsten weltweit unternommenen Meinungsbefragung, eventuell in 3D[34], mittels eines Videospiels der Moral, warten[35].

Zu einem Schluss, der die Menschenrechte achtet, ist hingegen die 2016 vom deutschen Bundesverkehrsminister berufenen Ethikkommission gelangt, die damit beauftragt war, einen Ethik-Kodex für den automatisierten und vernetzten Kraftfahrzeugverkehr zu erstellen. Der 2017 veröffentlichte Kodex schreibt vor:

> Bei unausweichlichen Unfallsituationen ist jede Qualifizierung nach persönlichen Merkmalen (Alter, Geschlecht, körperliche oder geistige Konstitution) strikt untersagt. Eine Aufrechnung von Opfern ist untersagt. Es ist somit unstatthaft, in Notstandssituationen eine Person aufzuopfern, um andere zu retten."[36]

Die Kommission beruft sich hierfür auf die bekannte Entscheidung des Bundesverfassungsgerichts, das – als unvereinbar mit dem verfassungsrechtlich garantierten Schutz der Menschenwürde und mit dem Recht auf Leben – die Regelung des Luftsicherheitsgesetzes von 2005 für verfassungswidrig erklärt hatte[37], das den Waffengebrauch durch die Luftwaffe gegen entführte Flugzeuge erlaubt hatte: die Tötung der als Geiseln genommenen Passagiere, um eine zahlenmäßig größere Menschenmenge zu schützen, habe nämlich die Passagiere zu bloßen Sachen erniedrigt, indem sie diese als Mittel zum Zweck machte, und ihnen damit den Wert, der jedem Menschen als solchem zukomme, abgesprochen[38].

---

34 Dies schlagen vor *L. R. Sütfeld et al.*, Using Virtual Reality to Assess Ethical Decisions in Road Traffic Scenarios, a.a.O., S. 5.

35 *Derek Leben,* Ethics for Robots. How to Design a moral Algorithm. London and New York (Routledge) 2019, S. 112, bemerkt, dass bei Anwendung von sozialen Fähigkeiten als Kriterium für moralische Urteile das Experiment des MIT eher diskriminierende Vorurteile der Personen als ihr moralisches Urteil untersucht.

36 Ethik-Kommission, Automatisiertes und Vernetztes Fahren. Eingesetzt durch den Bundesminister für Verkehr und digitale Infrastruktur, Bericht, Juni 2017, S. 11, https://www.bmvi.de/SharedDocs/DE/Publikationen/DG/bericht-der-ethik-kommission.html.

37 Luftsicherheitsgesetz, § 14 Abs. 3.

38 Bundesverfassungsgericht, Urteil des Ersten Senats vom 15. Februar 2006 – 1 BvR 357/05 – Rn. (1–156), http://www.bverfg.de/e/rs20060215_1bvr035705.html.

Die deutliche Zitierung Kants[39] in diesem letzten Teil der Urteilsbegründung kann dazu dienen, daran zu erinnern, dass das Fehlen von allgemein anerkannten Prinzipien im Bereich der Moral weniger dramatisch ist, als es jene Gelehrten ausmalen, die bei der Frage, wie Fahrzeuge mit autonomer Lenkung programmiert werden sollen, summarisch die theoretischen Alternativen der normativen Ethik durchgehen und sich, befremdet durch die jahrhundertelange Nichteinigung, abwenden und sich der Online-Befragung der weltweiten Bevölkerung zuwenden, um daraus im Wege des Algorithmus die auf die „öffentliche Moral" gegründete Lösung des *trolley dilemma* zu gewinnen. Von den Alternativen, die sie präsentieren (Utilitarismus, Deontologismus, Tugendethik und mitunter auch Kontraktualismus), ist nämlich eine grundlegende deontologische Entscheidung bereits nach dem Zweiten Weltkrieg mit der Einführung von strengen, der normalen Gesetzgebung übergeordneten Verfassungen in die Rechtssysteme vollzogen worden, die einige der Menschenrechte in sich aufgenommen haben und sie somit der Willkür des Gesetzgebers entzogen haben – sodass die etwaige Abweichung eines Gesetzes von den Grundprinzipien dem Gesetz seine materielle Gültigkeit nimmt, wie im Falle des deutschen Luftsicherheitsgesetzes – und damit die Rechte auf Leben und Nichtdiskriminierung u.a. auch außerhalb der phantasievollen Überlegungen zum Dilemma des Eisenbahnwagens stellt.

Die deutsche Ethik-Kommission hat auch jenes Prinzip abgelehnt, das den sog. personalisierten „ethischen Geräten" zugrunde liegt, welche – ausgehend von der üblichen Feststellung eines Fehlens allgemein anerkannter moralischer Prinzipien – als Alternative zu den im Voraus definierten Positionen der Automobilhersteller vorgeschlagen worden sind. Derartige Geräte würden es dem Eigentümer bzw. Führer des autonomen Fahrzeuges ermöglichen, im Voraus die „ethische" Einstellung des Fahrzeugs zu bestimmen – und diese eventuell in einem personalisierten elektronischen Gerät über ein Menü mit Auswahlfeldern zu speichern[40], wie es für moralisch indifferente Parameter

39 Kant's gesammelte Schriften, hrsg. von der Preußischen Akademie der Wissenschaften, Berlin (de Gruyter) 1900 ff., 4, S. 429: „Handle so, daß du die Menschheit sowohl in deiner Person, als in der Person eines jeden andern jederzeit zugleich als Zweck, niemals bloß als Mittel brauchst".

40 *W. Loh / J. Loh*, Autonomy and Responsibility in Hybrid Systems. The Example of Autonomous Cars, in: P. Lin / K. Abney / R. Jenkins (Hrsg.), Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence. New York (Oxford University Press) 2017, S. 35–50: 46; s. auch *J. Millar*, Ethics Settings for Autonomous Vehicles, in: Robot Ethics 2.0, a.a.O., S. 20–34. Kritisch *P. Lin*, Here's a Terrible Idea: Robot Cars With Adjustable Ethics Settings, in: Wired 2014, https://www.wired.com/2014/08/heres-a-terrible-idea-robot-cars-with-adjustable-ethics-settings/.

geschieht, oder über einen „Ethik-Knopf“[41], der die (altruistische, egoistische, unparteiische) Entscheidung über die Verteilung der Risiken im Falle von Unfällen zu übernehmen vermag. Die etwaige egoistische Entscheidung, stets Fußgänger oder Passanten statt der Insassen des autonomen Fahrzeugs zu töten, wäre – so meint man – gerechtfertigt durch die angenommene Notstandssituation[42]. Die Kommission trifft die Wertung, dass eine solche Entscheidung sich als eine vorsätzliche und systematische Verletzung der Rechte solcher unglücklichen Passanten darstellen würde, und sie meint daher, dass „die an der Erzeugung von Mobilitätsrisiken Beteiligten […] Unbeteiligte nicht opfern [dürfen]“[43]. Im Übrigen gründet sich, wie schon bemerkt, die Auffassung von der Möglichkeit, Geräte zu programmieren, welche die individuellen moralischen Präferenzen sammeln, um sie sodann auf den Kraftfahrzeugverkehr anzuwenden, auf die Überzeugung, dass der individuellen ethischen Entscheidung Verhaltensweisen überlassen würden, die in Wirklichkeit durch Rechtsnormen geregelt sind oder doch in deren Zuständigkeit fallen[44].

Auch wenn man die von der Ethikkommission aufgeworfenen Fragen einmal außer Betracht lässt, insbesondere von den individuellen Eigenschaften der potentiellen Opfer absieht, macht die begriffliche Struktur des trolley-Dilemma – wegen seines simplifizierenden Charakters mit klaren und sicheren Ergebnissen, typisch für Gedankenexperimente, und losgelöst von den sozialen Folgen seiner etwaigen wiederholten Anwendung – dieses ungeeignet, die ethischen Fragen darzustellen, die sich im realen Straßenverkehr stellen, denn diese

41 *G. Contissa / F. Lagioia / G. Sartor*, The Ethical Knob: Ethically-customisable automated vehicles and the law, in: Artificial Intelligence and Law XXV, 3, 2017, S. 365–378; *Iidem*, La manopola etica: i veicoli autonomi eticamente personalizzabili e il diritto, in: Sistemi intelligenti III, 2017, S. 601–614.

42 S. dazu *F. Santoni De Sio*, Killing by Autonomous Vehicles and the Legal Doctrine of Necessity, in: Ethical Theory and Moral Practice XX, 2, 2017, S. 411–429; zur radikalen Unterscheidung zwischen einer Entscheidung, die ein einzelnes Individuum in einer Notstandssituation trifft, und eine Entscheidung, die von verschiedenen Vertretern von Interessen darüber getroffen wird, wie ein bestimmter Typus von Technologie programmiert werden soll, um auf Situationen zu reagieren, welche eintreten könnten, s. *S. Nyholm / J. Smids*, The ethics of accident-algorithms for self-driving cars: An applied trolley problem?, in: Ethical Theory and Moral Practice XIX, 5, 2016, S. 1275–1289.

43 Ethik-Kommission Automatisiertes und Vernetztes Fahren, a.a.O., S. 11; zur Bedeutung der Unterscheidung zwischen beteiligten und nicht beteiligten Personen vgl. *D. Hübner / L. White*, Crash Algorithms for Autonomous Cars: How the Trolley Problem Can Move Us Beyond Harm Minimisation, in: Ethical Theory and Moral Practice XXI, 3, 2018, S. 685–698.

44 Vgl. *A. Etzioni* / *O. Etzioni*, Incorporating Ethics into Artificial Intelligence, in: The Journal of Ethics XXI, 4, 2017, S. 403–418.

verlangen vielmehr angesichts der konstitutionellen Ungewissheit bezüglich der Reaktionen der involvierten Personen eine Würdigung der Wahrscheinlichkeiten und Risiken[45]. Angezeigter wäre es, statt einer Diskussion über die moralischen Dilemmata eine solche über die Standards der funktionalen Sicherheit zu führen – im Hinblick auf die mit der Sicherheit autonomer Fahrzeuge angestrebten Zielsetzungen wie diejenigen, Zusammenstöße zu vermeiden und im Gleis zu bleiben, sowie auf die Bewertung der Risiken und den Umgang mit ihnen[46]. Es wird auch vereinzelt, auf der Grundlage einer Erörterung der technischen Aspekte, die Auffassung vertreten, dass es angemessen sei, unumstößlich vorzusehen, dass in den verschiedenen Fällen, die üblicher Weise als trolley-Dilemmata angesehen werden, das Fahrzeug einfach bremst[47].

Möglicherweise resultiert die Versuchung, die „Entscheidungen" der autonomen Fahrzeuge in den Grenzen der Lösung von ethischen Dilemmata zu konzeptualisieren, aus der Gleichsetzung dieser Fahrzeuge mit menschlichen Personen, die sie in der Rolle von Fahrzeugführern ersetzen: das trolley-Dilemma ist danach vielleicht im Bereich des automatisierten Verkehrs nicht nur unangemessen, was das Objekt des Dilemmas angeht, sondern ebenfalls, was das Subjekt angeht, das, wie man meint, die Entscheidung im eigentlichen Sinne zusteht.

3. Die Moral des Thermostat

Die Entwicklung von Systemen künstlicher Intelligenz, die imstande sind, ohne direkten menschlichen Eingriff einen unter mehreren alternativen Handlungsverläufen auszuwählen und in Gang zu setzen, und zwar in einem Zu-

45 *S. Nyholm / J. Smids*, The Ethics of Accident-Algorithms for Self-Driving Cars: An applied trolley problem?, a.a.O.; *A. Etzioni / O. Etzioni*, Incorporating Ethics into Artificial Intelligence, in: The Journal of Ethics XXI, 4, 2017, S. 403–418; *N. J. Gogoll & J. F. Müller*, Autonomous cars: In favor of a mandatory ethics setting, in: Science and Engineering Ethics XXIII, 3, 2017, 681–700.

46 *R. Johansson / J. Nilsson,* Disarming the Trolley Problem – Why Self-driving Cars do not Need to Choose Whom to Kill, in: Workshop CARS 2016 – Critical Automotive applications: Robustness & Safety, 2016, https://hal.archives-ouvertes.fr/hal-01375606/document; *N. J. Goodall*, Away from trolleys and toward risk-management, in: Applied Artificial Intelligence XXX, 8, 2016, S. 810–821. Einen getrennten Verkehr für autonome Fahrzeuge schlägt vor *G. Tamburrini*, Autonomia delle macchine e filosofia dell'intelligenza artificiale, in: Rivista di filosofia LVIII, 2, 2017, S. 263–275.

47 *R. Davnall*, Solving the Single-Vehicle Self-Driving Car Trolley Problem Using Risk Theory and Vehicle Dynamics, im Erscheinen in: Science and Engineering Ethics 2019, https://doi.org/10.1007/s11948-019-00102-6.

sammenhang, in dem diese Handlungen – wenn sie von Menschen ausgeführt würden – moralische Bedeutung besäßen, hat seit etwa einem Jahrzehnt eine Debatte über künstliche Moral ausgelöst, die Gegenstand eines spezifischen interdisziplinären Forschungsbereichs ist, der Informatik, Philosophie und Robotik einbezieht – der Maschinenethik, besser gesagt: der Ethik *für* Maschinen[48].

Die Anerkennung der Sinnhaftigkeit eines derartigen Forschungsfeldes gründet sich auf die metaethische Annahme, dass es möglich sei, Maschinen mit der Fähigkeit, moralisch zu entscheiden und zu handeln, auszustatten, d.h. dass Ethik in berechenbare Begriffe übersetzt werden könne und dass Maschinen moralisch Handelnde sein könnten.

Die Angemessenheit, künstliche Wesenheiten als moralisch Handelnde zu qualifizieren – wenn auch als solche ohne Bewusstsein, Emotionen und Gefühle, ohne Geisteszustand und ohne Absichten (und somit nicht verantwortlich, ja nicht einmal der Handlungen bewusst, die sie ausführen) – ist von Luciano Floridi und G. W. Sanders als Überwindung eines anthropozentrischen und anthropomorphen Verständnisses moralischen Handelns bejaht worden; es vermöge neue Perspektiven der Untersuchung von „aufgeteilter Moral" zu eröffnen – auf einem bestimmten Abstraktionsniveau könne eine Wesenheit, welche die drei Anforderungen Interaktivität, Autonomie (Fähigkeit, innere Übergänge zu bewirken, um den eigenen Zustand zu verändern) und Anpassungsfähigkeit (Fähigkeit, die eigenen Übergangsregeln zu modifizieren) erfülle, als eine Handelnde und – falls sie das „moralische Spiel" spielt – als moralisch Handelnde angesehen werden.

Auch ein Thermostat im Krankenhaus, der imstande ist, nicht nur die Raumtemperatur zu überwachen, sondern auch das Wohlbefinden der Patienten, und daraus folgend die Temperatur zu regulieren, könnte daher als moralisch Handelnder qualifiziert werden (moralisch gut, falls seine *Outputs* die Gesundheit der Patienten innerhalb einer vorausbestimmten Schwelle aufrechterhalten)[49]. Der Ansatz von Floridi und Sanders ist erklärtermaßen „phänomenologisch"; das Abstraktionsniveau fällt zusammen mit einer Ansammlung von beobachtbaren Fähigkeiten, von denen jede mit einem wohldefinierten *set* von möglichen Werten verbunden ist; die Einordnung der Fähigkeit zur Anpassung in ein System künstlicher Intelligenz hängt somit ab

48 Vgl. neben den bereits in Fußn. 3 zitierten Bänden den vorzüglichen Beitrag von *C. Misselhorn*, Grundfragen der Maschinenethik, a.a.O.

49 *L. Floridi / J. W. Sanders*, On the morality of artificial agents, in: Minds and Machines XIV, 3, 2004, S. 349–379.

von der Unkenntnis des Beobachters – auf einer zuvor festgelegten Abstraktionsebene – über den ihre Übergänge regelnden Kodex bzw. Algorithmus.

Das von Floridi und Sanders als angemessen angesehene Abstraktionsniveau für die Definition des Begriffs des moralisch Handelnden schließt keinen der „spezifischen inneren Zustände", wie etwa die intentionalen Zustände – d.h. „das Beabsichtigen, Wünschen oder Handeln-wollen in einer bestimmten Weise und das wissensmäßige Bewusstsein des eigenen Handelns" – ein, die „eine hübsche, aber nicht notwendige" Bedingung für die Eigenschaft als moralisch Handelnder sind. Floridi und Sanders führen damit einen Begriff der „nicht-verantwortlichen Moralität" ein, der ihnen für das Cyberspace angemessener erscheint als ein Verständnis von Moralität, das in der menschlichen Person verankert ist (*human-based*).

In einer technischen und rechtlichen Übergangsphase, in der das Fehlen von Kontrolle und Berechenbarkeit der Systeme künstlicher Intelligenz verlangt, dass die Frage der moralischen (vor allem aber der rechtlichen) Verantwortlichkeit der Unternehmen, welche diese Systeme herstellen, mit neuen Begriffen angegangen wird, kann die Position von Floridi und Sanders von den Vertretern dieser Unternehmen nur mit Begeisterung aufgenommen werden, denn sie verteidigt eine „Erweiterung der Klasse der moralisch Handelnden", mit der „wir das Rückwärtsschreiten der Suche nach dem *verantwortlichen* Individuum, wenn etwas Schlimmes geschieht, aufhalten können".

Die Gedankenoperation, Handeln auf einem bestimmten Abstraktionsniveau als moralisch anzusehen, ist zweifellos legitim; es gilt allerdings zu berücksichtigen, welche theoretischen und praktischen Konsequenzen eine solche Operation mit sich bringt. Aus sozialer und rechtlicher Sicht muss man sich des Risikos bewusst sein, dass ein solcher Begriff nicht-verantwortlicher und verteilter Moralität die Menschen von ihren allgemeinen und besonderen Verantwortlichkeiten für die Richtung des technischen Fortschritts, für seine spezifischen Hervorbringungen, befreit, sodass „am Ende niemand mehr die Verantwortung für ihr Handeln übernimmt[50] und das Verhalten der Maschinen abgelöst wird von demjenigen der Menschen, die sie entwerfen, konstruieren und benutzen"[51].

Aus theoretischer Sicht bedeutet die Konzeption einer „Moral ohne Geist" (*mind-less morality*), welche die künstlichen Handelnden in die Klasse der moralisch Handelnden einbezieht, eine extreme Form des Reduktionismus,

50 *C. Misselhorn*, Grundfragen der Maschinenethik, a.a.O., S. 14, 133.

51 *D. G. Johnson*, Computer Systems: Moral Entities, but not Moral Agents, in: Machine Ethics, a.a.O., S. 168–183: 183.

welche die Analogie zwischen dem Verhalten von Maschinen, das den Zwecken angemessen ist, und menschliches Verhalten objektiv anstatt bloß erklärend interpretiert[52]. Um künstliche Systeme als moralisch Handelnde zu qualifizieren, entzieht man den im eigentlichen Sinne menschlichen Merkmalen wie Bewusstsein und Intentionalität – indem man ihren Begriff zum Residuum einer magischen Mentalität degradiert – die begriffliche Legitimität[53], weil sie nicht auf eine Beschreibung im rein funktionalen Sinne reduziert werden können[54].

Anzunehmen, dass ein Abstraktionsniveau, das vom Bewusstsein seiner selbst, von der Intentionalität sowie von Wünschen und Empfindungen absieht, geeignet sei, um einen moralisch Handelnden zu beschreiben[55], ist gleichbedeutend damit, als wenn man paradoxer Weise – während man doch meint, die Theorien von den „Phantomen in der Maschine" zu liquidieren[56] – genau eine metaphysische Position begründet, nämlich eine Art von cartesianischem Dualismus, der Rationalität, Erkenntnis und auch menschliche Moralauffassungen auf körperlose Weise versteht[57] und den Verstand mit einer *software*, und damit im Wege des Algorithmus reproduzierbar gleichsetzt. Ein solches Verständnis steht im Widerspruch zu den neueren wissenschaftlichen Erkennt-

52 S. *F. Fossa*, Fare e funzionare. Sull'analogia di robot e organismo, in: InCircolo VI, 2018, S. 73–88. Vgl. *P. Moro*, Libertà del robot? Sull'etica delle macchine intelligenti, in: R. Brighi / S. Zullo (Hrsg.), Filosofia del diritto e nuove tecnologie. Prospettive di ricerca tra teoria e pratica. Rom (Aracne) 2015, S. 525–544.

53 Zu dieser Frage müssen wir uns an dieser Stelle auf den Hinweis auf zwei klassische Texte beschränken: *T. Nagel*, What Is It Like to Be a Bat? In: The Philosophical Review LVIII, 4, 1974, S. 435–50; *J. R. Searle*, Minds, brains, and programs, in: The Behavioral and Brain Sciences III, 1980, S. 417–424.

54 *W. Wallach / C. Allen*, Moral machines: Teaching Robots Right from Wrong, a.a.O., S. 53: „Das Verständnis ist mitunter gleichbedeutend mit dem Wissen – ein anderes Wort mit magischen Konnotationen".

55 Vielstimmige Diskussion über den Vorschlag von Floridi und Sanders b. *D. J. Gunkel / J. J. Bryson / S. Torrance* (Hrsg.), The Machine Question: Ai, Ethics and Moral Responsibility, Birmingham, 2012, http://events.cs.bham.ac.uk/turing12/proceedings/14. pdf. Von den Beiträgen aus jüngerer Zeit bieten eine zusammenfassenden Überblick über die theoretischen Alternativen *D. Brhdadi / C. Munthe,* Artificial Moral Agency: Philosophical Assumptions, Methodological Challenges, and Normative Solutions, 2019, https://www.researchgate.net/publication/311196481.

56 *L. Floridi / J. W. Sanders*, On the morality of artificial agents, a.a.O.

57 *J. Bryson / P. Kime*, Just an Artifact: Why Machines are Perceived as Moral Agents, in: T. Walsh (Hrsg.), Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence. Menlo Park (AAAI Press) 2011, S. 1641–1646: 1642: „Die Tendenz zur Überidentifizierung mit Maschinen stammt von einer irrigen Auffassung von menschlichem Leben, die als dessen entscheidende Merkmale die Fähigkeiten zur Sprache, zur Mathematik und zur Vernunft ansieht".

nissen über die körperbezogene Natur des Geistes und über die – auch kognitive – Rolle der Emotionen und Gefühle[58]. Auf ein solches Verständnis stützen sich – und bezeugen mit ihren spärlichen Fortschritten seine Unzulänglichkeit – die bisher unternommenen Versuche, Systeme künstlicher Intelligenz mit Gemeinsinn auszustatten[59].

Das von Floridi und Sanders festgelegte Abstraktionsniveau, das geeignet sein soll, einen moralisch Handelnden zu beschreiben, verkennt darüber hinaus die wesentliche Funktion der – von den zeitgenössischen Neurowissenschaften dem „Empathie-Kreislauf"[60], um den sich die philosophischen Reflexionen über den Begriff der Anerkennung drehen[61], zugeschriebenen – Fähigkeit, Personen von Sachen zu unterscheiden und den Ersteren einen unbedingten Wert, eine Würde und Rechte zuzuschreiben. Die Entscheidung, den Status als moralisch Handelnde nur Menschen zuzubilligen – und damit an der Unterscheidung zwischen Personen und Sachen festzuhalten – stützt sich nicht auf eine „bloße psychologische Spekulation"[62], die nicht in der Lage sei, die äußerlich beobachtbaren Fähigkeiten zur Kenntnis zu nehmen: in den zeitgenössischen Verfassungsurkunden ist diese Entscheidung aufgrund tragischer und fürchterlicher Erfahrungen mit den konkreten Folgen einer begrifflichen Reduzierung von Personen auf Sachen rechtlich abgesichert[63].

Es ist natürlich nicht gesagt, dass eine solche Reduzierung von all denen verlangt würde, welche „die ethischen Dilemmata der Maschinen" im wörtlichen und nicht im übertragenen Sinne auffassen. In der kollektiven Vorstellungswelt[64] ist vielmehr die entgegengesetzte Tendenz wirksam, nämlich die

58 *A. Damasio*, Descartes' Error: Emotion, Reason and the Human Brain. New York (Grosset / Putnam) 1994; *Ders.*, The Strange Order of Things: Life, Feeling, and the Making of Cultures. New York (Pantheon) 2018.

59 Vgl. *Y. Shoham / R. Perrault / E. Brynjolfsson / J. Clark / J. Manyika / J. C. Niebles / T. Lyons / J. Etchemendy / B. Grosz / Z. Bauer*, The AI Index 2018 Annual Report. Stanford (Stanford University) 2018, S. 64.

60 *S. Baron-Cohen*, The Science of Evil: On Empathy and the Origins of Cruelty. New York (Basic Books) 2011.

61 Vgl. z.B. *E. Nowak-Juchacz*, Das Anerkennungsprinzip bei Kant, Fichte und Hegel, in: Fichte-Studien XXIII, 2003, S. 75–84.

62 *L. Floridi / J. W. Sanders*, On the morality of artificial agents, a.a.O.

63 *Simon Baron-Cohen,* La scienza del male, a.a.O., S. 1 ff.) verbindet seinen Wunsch, als Wissenschaftler „die Faktoren [zu verstehen], die Personen veranlassen, andere Personen als Objekte zu behandeln", mit einer biographischen Episode: dem Bericht über eine reale Transformierung von Menschen in Objekte (als er sieben Jahre alt war, sagte sein Vater zu ihm, die Nazis hätten die Juden in Seife verwandelt).

64 Vgl. *B. Henry*, Dal Golem ai cyborgs: trasmigrazioni nell'immaginario. Livorno (Belforte) 2013; *F. Battaglia / N. Weidenfeld* (Hrsg.), Roboethics in film. Pisa (Pisa

anthropomorphe Neigung, Maschinen die Fähigkeit, nachzudenken, zu urteilen und eine moralische Entscheidung im engeren Sinne zu vollziehen, zuzusprechen statt nur die bloße Fähigkeit, programmierte oder eingeübte Regeln zu befolgen[65]. Häufig wird das Abgleiten vom übertragenen zum wörtlichen Sprachsinn – indem das aus der Biologie entnommene Wort verwendet wird, um das Funktionieren von Maschinen zu beschreiben, die dann „gehorchen", „handeln" oder „entscheiden" – weder thematisiert noch wahrgenommen[66] und enthält dann Merkmale eines magischen Denkens[67].

Wie Fabio Fossa bemerkt hat, handelt es sich um einen Prozess semantischer Ausdehnung der gewöhnlichen Sprache innerhalb derselben Sprache, dem man sich nicht entziehen kann. Jedoch muss man sich vor Augen halten, dass die Verwendung derselben Terminologie für Menschen und Maschinen „suggeriert, dass es keinen beachtlichen Unterschied gebe" zwischen dem Archetyp und seiner Imitation und dazu verleitet „Maschinen wie Organismen zu behandeln" und damit gegenüber der Technik irreale Erwartungen zu wecken, andererseits „Organismen als Maschinen zu behandeln" und damit die Anwendung von unpassenden begrifflichen Schemata technischer Herkunft auf Menschen zu legitimieren[68].

Die größere Gefahr entsteht nach meinem Dafürhalten aus der Gleichsetzung der Undurchsichtigkeit, d.h. der Nichterklärbarkeit des Verhaltens der Systeme künstlicher Intelligenz mit ihrer moralischen Autonomie, denn die Unvorhersehbarkeit, die für die künstlichen neuralen Netze kennzeichnend ist – die

---

University Press) 2015; *F. Battaglia*, Macchine morali, in: Scienza e Filosofia XIII, 2015, S. 193–202.

65 *K. Weber*, Autonomie und Moralität als Zuschreibung: Über die begriffliche und inhaltliche Sinnlosigkeit einer Maschinenethik, in: M. Rath / F. Krotz / M. Karmasin (Hrsg.), Maschinenethik. Normative Grenzen autonomer Systeme. Wiesbaden (Springer) 2019, S. 193–208.

66 Eine Empfehlung, in die Diskussion über die Ethik der autonomen und intelligenten Systeme eine kritische Bewertung der anthropomorphen Auffassungen einzubeziehen, findet sich in der zweiten Fassung des Dokumentes, das vom Institute of Electrical and Electronics Engineers (IEEE) im Rahmen der Global Initiative on Ethics of Autonomous and Intelligent Systems der öffentlichen Diskussion unterbreitet worden ist (Ethically Aligned Design. A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, 2018, S. 195 ff., https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf.

67 Vgl. *M. Musiał*, Enchanting Robots. Intimacy, Magic, and Technology. Cham (Palgrave Macmillan) 2019, S. 63–113.

68 *F. Fossa*, Creativity and the Machine: How Technology Reshapes Language, in: ODRADEK. Studies in Philosophy of Literature, Aesthetics and New Media Theories III, 1–2, 2017, S. 177–208.

imstande sind, automatisch und unüberwacht zu lernen, indem sie, von enormen Datenmengen ausgehend, Modelle und Strukturen berechnen und Entscheidungen nach Kriterien treffen, die für den menschlichen Beobachter undurchsichtig bleiben – kann nur äußerlich mit derjenigen des menschlichen Verhaltens gleichgesetzt werden. Die Maschine ist nicht in dem Sinne eine „black box", dass uns ihre Wünsche, ihre Absichten und ihr Wille unbekannt wären, sondern nur insoweit, als sie, ausgehend von Daten, Modelle erarbeitet hat, die zu erklären uns nicht generell möglich ist und auf deren Grundlage die Maschine agiert und reagiert: „Kleine Veränderungen in der wahrgenommenen Welt könnte auf unvorhersehbare Weise eine angemessene Reaktion in eine völlig ungeeignete mit potentiell tödlichen Folgen verwandeln"[69].

Die Transparenz bildet daher ein entscheidendes – derzeit nicht erfülltes, obwohl stets zu den *desiderata* und den Empfehlungen gehöriges – Element[70] der Systeme künstlicher Intelligenz: die – menschliche – Entscheidung, zuzulassen, dass Maschinen oder Algorithmen Wirkungen auf andere Menschen ausüben, sollte durch die Einsehbarkeit der Art und Weise, in der sie agieren, und die Möglichkeit, diese Wirkungen der Verantwortlichkeit von Menschen zuzuschreiben, bedingt werden. Die Alternative, einige Entscheidungsprozesse zu automatisieren, indem man sie an Systeme künstlicher Intelligenz delegiert, sodass Modelle konstruiert werden, die uns unbekannt sind und die nicht die elementarsten Regeln des gesunden Menschenverstandes einschließen, zöge die – wie Adriano Fabris gezeigt hat, für die neuen Technologien charakteristische[71] – Folge eines Verlustes der Vorhersehbarkeit und Kontrolle über das Funktionieren dieser Technologien nach sich, deren Kosten und deren Nutzen

69 When Computers Decide: European Recommendations on Machine-Learned Automated Decision Making, Informatics Europe & EUACM, 2018, S. 11, https://www. acm. org/binaries/content/assets/public-policy/ie-euacm-adm-report-2018.pdf. Vgl. *E. Sirgio vanni / G. Corbellini*, IA e neuroetica: evoluzione spontanea e valori morali, in: Giornale italiano di psicologia, Fascicolo 1, marzo 2018, S. 147–152:149.

70 S. z.B. *L. Floridi / J. Cowls / M. Beltrametti / R. Chatila / P. Chazerand / V. Dignum / C. Luetge / R. Madelin / U. Pagallo / F. Rossi / B. Schafer / P. Valcke / E. Vayena*, AI4People–An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations, in: Minds and Machines XXVIII, 4, 2018, S. 689–707; European Commission's High-Level Expert Group on Artificial Intelligence, Ethics Guidelines for Trustworthy AI, 2019, S. 13, https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai; Commission National Informatique & Libertés (CNIL), Comment permettre à l'homme de garder la main? Les enjeux éthiques des algorithmes et de l'intelligence artificielle, 2017, S. 51, https://www.cnil.fr/sites/default/files/atoms/files/ cnil_rapport_garder_la_main_web.pdf.

71 *A. Fabris*, La filosofia e lo specchio delle macchine, in: InCircolo VI, 2018, S. 28–38: 33.

Gegenstand einer sorgfältigen vergleichenden Bewertung im Einzelfall sein müssten.

4. Schlussbemerkung

Ein Dilemma wie das *trolley problem,* das eine Entscheidung darüber verlangt, welche Menschenleben im Kraftfahrzeugverkehr geopfert werden sollen, gehört dem Bereich des Rechts und nicht dem der Moral an. Das Recht umschreibt in den Verfassungsurkunden die Fälle, in denen es erlaubt ist, diese Frage aufzuwerfen, wobei es, wie gezeigt, jene Konstellationen ausschließt, die zu Verletzungen der Grund- und Menschenrechte auf Leben und Gleichbehandlung führen würden, und das Recht, das in der Straßenverkehrsordnung minutiös erheblich weniger wichtige Fälle regelt, muss in etwaigen Fällen, in denen es wirklich notwendig ist (und nicht bloß angemessen), die Frage in Begriffen von Sicherheitsstandards und der Einschätzung und Handhabung der Risiken zu beantworten, selbst die Entscheidung darüber, wer auf systematische Weise im Kraftfahrzeugverkehr getötet werden soll, auf systematische Weise treffen.

Was die moralischen Dilemmata angeht, so können sie weder einem Fahrzeug mit autonomer Lenkung gestellt werden noch irgend einer anderen Maschine, denn in der aktuellen Entwicklungsphase der technischen Entwicklung der Systeme künstlicher Intelligenz ist eine Maschine nicht ein moralisch Handelnder im eigentlichen Sinne[72]: Sie ist sich nicht dessen bewusst, was sie tut, hat keine Neigungen, keine Wünsche und ist ohne Empathie und ohne gesunden Menschenverstand. Dass die Maschine, statt die in einem Programm enthaltenen Instruktionen auszuführen, elaborierte Modelle automatisch und ohne Überwachung anwendet, macht ihr Verhalten unzugänglich für menschliche Erklärungen, damit aber weder intentional noch verantwortlich.

Die entscheidende Frage ist daher nicht, ob es wünschenswert ist, künstliche moralisch Handelnde zu schaffen oder nicht[73], denn dies gehört derzeit noch nicht zu den vorhandenen technischen Möglichkeiten, sondern es gilt, sich

72 Zur Alternative zwischen der Kontinuitäts- und Diskontinuitätsthese zwischen menschlichen und künstlichen moralisch Handelnden s. *F. Fossa*, Artificial Moral Agents: Moral Mentors or Sensible Tools? In: Ethics and Information Technologies XX, 2, 2018, S. 115–126.

73 Vgl. *A. Van Wynsberghe / S. Robbins*, Critiquing the Reasons for Making Artificial Moral Agents, in: Science and Engineering Ethics 2018, https://doi.org/10.1007/s11948-018-0030-8; *A. Poulsen / M. Anderson / S. L. Anderson / B. Byford / F. Fossa / E. L. Neely / A. Rosas / A. Winfield*, Responses to a Critique of Artificial Moral Agents, 2019, https://arxiv.org/ftp/arxiv/papers/1903/1903.07021.pdf.

daran zu erinnern, dass Maschinen, welche in verschiedenen Kontexten menschliches Verhalten nachahmen, neue moralische Fragen für die Menschen aufwerfen, welche diese Maschinen entwerfen, konstruieren, benutzen und ihre Aufgaben und Funktionen rechtlich regeln[74]: den Menschen obliegt die Pflicht, technische Lösungen zu finden, welche sicherstellen, dass bei der Erarbeitung von Daten über menschliche Verhaltensweisen die Maschinen nicht Modelle schaffen, die deren Vorurteile und Diskriminierungen unter gewaltiger Verbreiterung der Wirkungen reproduzieren[75], vielmehr auf eine Weise funktionieren, dass nicht die Rechte irgendeiner Person verletzt werden. Nur in diesem übertragenen, uneigentlichen, metaphorischen und anthropomorphen Sinne kann man heute von einer Ethik der Maschinen sprechen[76].

74 Vgl. *A. Fabris*, Etica delle macchine, in: Teoria XXXVI, 2, 2016, S. 119–136; *M. Verdicchio*, An Analysis of Machine Ethics from the Perspective of Autonomy, in: T. Powers (Hrsg.), Philosophy and Computing: Essays in Epistemology, Philosophy of Mind, Logic, and Ethics. Cham (Springer) 2017, S. 179–191.

75 Vgl. *C. O'Neil*, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. New York (Crown) 2016.

76 Zutreffend ist die Definition von Oliver Bendel: „'Maschinelle Moral' ist ein Terminus technicus wie 'künstliche Intelligenz'. Man spielt auf ein Setting an, das Menschen haben, und man will Komponenten davon imitieren bzw. simulieren. […] Moralische und unmoralische Maschinen sind nicht gut oder böse, sie haben keinen freien Willen und kein Bewusstsein, keine Intuition und keine Empathie", in: Gablers Wirtschaftslexikon. Wiesbaden (Springer Gabler) 2019, https://wirtschaftslexikon.gabler.de/definition/moralische-maschinen-119940). S. auch *M. S. Vaccarezza*, Macchine morali: responsabilità e saggezza pratica degli agenti artificiali, in: F. Miano (Hrsg.), Etica e responsabilità. Neapel (Orthotes) 2018, S. 309–318.