Tobias Matzner

# Autonome Trolleys und andere Probleme. Konfigurationen Künstlicher Intelligenz in ethischen Debatten über selbstfahrende Kraftfahrzeuge

2019

https://doi.org/10.25969/mediarep/12632

Veröffentlichungsversion / published version

Zeitschriftenartikel / journal article

**Empfohlene Zitierung / Suggested Citation:**

Matzner, Tobias: Autonome Trolleys und andere Probleme. Konfigurationen Künstlicher Intelligenz in ethischen Debatten über selbstfahrende Kraftfahrzeuge. In: *Zeitschrift für Medienwissenschaft*. Heft 21: Künstliche Intelligenzen, Jg. 11 (2019), Nr. 2, S. 46–55. DOI: https://doi.org/10.25969/mediarep/12632.

TOBIAS MATZNER

# AUTONOME TROLLEYS UND ANDERE PROBLEME

## Konfigurationen Künstlicher Intelligenz in ethischen Debatten über selbstfahrende Kraftfahrzeuge

Selbstfahrende Kraftfahrzeuge sind eine prominente Anwendung Künstlicher Intelligenz (KI). Sie werden als großer zukünftiger Markt gesehen. Zwei der großen Industriesektoren, die Auto- und die IT-Industrie, stoßen aufeinander. Das könnte auch in der eigentlich sehr stabilen und politisch einflussreichen Kraftfahrzeugbranche zu großen Umstrukturierungen führen. Außerdem sind selbstfahrende Wagen ein sehr anschaulicher Fall: Sehr viele Menschen fahren selbst Auto und genau diese Tätigkeit soll nun durch KI ‹ersetzt› werden. Entsprechend wundert es nicht, dass die Ethik der Künstlichen Intelligenz oft anhand selbstfahrender Fahrzeuge diskutiert wird. Wenn KI Dinge tut, die wir alle auch tun, liegt die Frage nahe, wie es dann um die Moral dieser Tätigkeiten beschaffen ist. Die Ethik selbstfahrender Autos wird sehr oft in Bezug auf das sogenannte Trolley-Problem diskutiert. Genauer handelt es sich hier um eine ganze Gruppe von hypothetischen Fällen, die aus der Moralphilosophie stammen – deshalb ist im Folgenden von Trolley-Problemen die Rede. Immer geht es darum, dass der Tod eines oder mehrerer Menschen unvermeidlich ist. Jedoch gibt es die Möglichkeit, zwischen verschieden Entwicklungen der Situation zu wählen, die sich meist in einer Hinsicht unterscheiden: Anzahl der Opfer, aktive Handlung vs. passives Inkaufnehmen, verschiedenes Alter, Reichtum, Erfahrung der Opfer und vieles mehr. In Bezug auf Kraftfahrzeuge geht es also um einen hypothetischen, nicht mehr zu vermeidenden Unfall, in dem aber beeinflusst werden kann, wer zu Schaden kommen wird. Eine Ethik der Künstlichen Intelligenz bezieht sich in diesem Kontext dann primär auf die Frage, wie und ob selbstfahrende Kraftfahrzeuge programmiert werden sollten, um zwischen den zur Verfügung stehenden Alternativen zu ‹entscheiden›.

Im Folgenden möchte ich zeigen, dass aus den Trolley-Problemen nicht sehr viel über die Ethik Künstlicher Intelligenz abgeleitet werden kann. Die hier angenommenen Dilemmasituationen sind unter den ethischen und politischen Problemen des automatisierten Verkehrs eher unwichtig und unwahrscheinlich, wie

**1** Vgl. Karl Iagnemma: Why We Have the Ethics of Self-Driving Cars all wrong, in: *World Ecocomic Forum*, dort datiert 21.1.2018, *www.weforum.org/agenda/2018/01/why-we-have-the-ethics-of-self-driving-cars-all-wrong/*, gesehen am 12.11.2018.

**2** Vgl. Ian Bogost: Enough With the Trolley Problem, in: *The Atlantic*, dort datiert 30.3.2018, *www.theatlantic.com/technology/archive/2018/03/got-99-problems-but-a-trolley-aint-one/556805/*, gesehen am 12.11.2018.

ich in Abschnitt zwei zeige. Aber selbst wenn man solche Dilemmata für relevant hält, sind die Trolley-Probleme, so wie sie in Philosophie und nun zunehmend auch im Kontext der Debatten um selbstfahrende Kraftfahrzeuge diskutiert werden, keine gute Reflexionsbasis. Davon handelt Abschnitt drei. Abschnitt vier liest die Prominenz dieses Falls als Konsequenz eines implizierten Mensch-Technik-Verhältnisses, das medien- und techniktheoretisch problematisiert werden muss, um eine angemessenere ethische und politische Debatte über KI zu führen.

## I. Ablenkung durch Dilemmasituationen

Es wird erwartet, dass selbstfahrende Fahrzeuge schneller reagieren als Menschen. Diese Einschätzung teilen die Hersteller, die sich kürzlich auf Anfrage des World Economic Forum zu den Gefahren von selbstfahrenden Kraftfahrzeugen äußerten,[1] mit durchaus technikkritischen Stimmen.[2] Darin liegt erst einmal ein großes, auch ethisches Versprechen: Viele Unfälle, welche durch Fahrer_innen verursacht werden (je nach Land bis zu 90 % aller Unfälle)[3] ließen sich verhindern.[4] Das Potenzial dieser Verbesserung ist allerdings umstritten, weil dafür mehr und schwerwiegendere Fälle des Versagens der deutlich komplexeren Technik hinzukommen dürften.[5]

Hier zeigt sich eine erste Schwierigkeit der Debatten um die Trolley-Probleme. Denn die Frage, ob selbstfahrende Kraftfahrzeuge den Verkehr allgemein sicherer machen oder nicht und welche Bedingungen es hierfür braucht, spielt hier gar keine Rolle.[6] Die im Zusammenhang mit Trolley-Problemen diskutierten Situationen schließen die erwartbaren Fälle des technischen Versagens weitgehend aus. Sie setzen voraus, dass alles so weit funktioniert, dass ein unvermeidlicher Unfall sicher detektiert und kontrolliert reagiert werden kann. Insbesondere die schnellere Reaktionszeit von selbstfahrenden Kraftfahrzeugen wird hier oft als Grund dafür angeführt, dass Dilemmasituationen überhaupt rechtzeitig erkannt werden können und somit zum relevanten Problem werden. Aus einem erwarteten Unterschied in der Voraussicht und Reaktionsfähigkeit zwischen Menschen und selbstfahrenden Autos wird hier implizit die Möglichkeit kontrollierter Reaktion abgeleitet. Das ist aber sicher nicht in allen Fällen unvermeidlicher Unfälle möglich. Unter den Prämissen der Trolley-Probleme – z. B. schneller reagierende Fahrzeuge – dürfte sich die Zahl der unvermeidbaren Unfälle verringern. So legen Simulationen basierend auf realen Unfällen nahe, dass bis zur Hälfte davon durch schnelleres Notbremsen vermieden werden könnte.[7] In den verbleibenden Fällen wäre z. B. zu fragen, ob die Detektion der Unvermeidbarkeit des Unfalls auf Bewegungsprognosen für die Umweltobjekte beruhen darf, wie sie seit einiger Zeit für autonome Fahrzeuge erforscht werden,[8] oder ob eine so folgenreiche Entscheidung nur aufgrund der tatsächlichen Situation getroffen werden sollte. Dann dürfte das Fahrzeug in einigen Fällen bereits in einem Notfallmanöver sein (z. B. einer Vollbremsung), wenn klar wird, dass dieses nicht gelingen kann, wodurch eine Reaktion noch einmal

**3** Vgl. Leonard Evans: Death in Traffic: Why Are the Ethical Issues Ignored?, in: *Studies in Ethics, Law, and Technology*, Vol. 2, Nr. 1, 2008, 1–11, hier 5.

**4** Vgl. Alex Roy: Autonomous Cars Don't Have a «Trolley Problem», in: *The Drive*, dort datiert 19.10.2016, http://www.thedrive.com/tech/5620/autonomous-cars-dont-have-a-trolley-problem-problem, gesehen am 12.11.2018.

**5** Vgl. Michael Laris, Ashley Halsey III: Will Driverless Cars Really Save Millions of Lives? Lack of Data Makes it Hard to Know, in: *The Washington Post*, dort datiert 18.10.2016, http://wapo.st/2eCTOyo?tid=ss_mail&utm_term=.b22c5f61bbf7, gesehen am 12.11.2018.

**6** Deshalb wird mitunter argumentiert, dass Tötungsalgorithmen die Akzeptanz und Verbreitung selbstfahrender Autos verringern könnten und damit – unter der Annahme, diese seien im Allgemeinen sicherer – in der Summe zu mehr Unfällen führten. Das geht allerdings von den eben genannten umstrittenen Annahmen aus und verbleibt im allgemeinen Rahmen des Verrechnens von Menschenleben. Siehe dazu Nassim JafariNaimi: Our Bodies in the Trolley's Path, or Why Self-Driving Cars Must *Not* Be Programmed to Kill, in: *Science, Technology, & Human Values*, Vol. 43, Nr. 2, 2018, 302–323.

**7** Vgl. Giovanni Savino, Julie Brown, Matteo Rizzi u. a.: Triggering algorithm based on inevitable collision states for autonomous emergency braking (AEB) in motorcycle-to-car crashes, in: *Proceedings of the 2015 IEEE Intelligent Vehicles Symposium (IV) 2015*, 1195–1200.

**8** Vgl. Rishikesh Parthasarathi, Thierry Fraichard: An Inevitable Collision State-Checker for a Car-like Vehicle, in: *Proceedings 2007 IEEE International Conference on Robotics and Automation*, 3068–3073.

schwieriger zu realisieren wäre. Rein physikalisch wird z. B. in Situationen, in denen die Distanz zum Bremsen bereits zu kurz ist, ein gezieltes Lenkmanöver ohne Schlingern oder Ausbrechen immer unwahrscheinlicher.[9] Schließlich könnte ein kontrollierter Unfall Manöver beinhalten, wie gezielt in eine Wand, auf den Gehweg oder einen Grünstreifen zu fahren. Dafür müsste es speziell für solche Fälle erzeugte Modelle geben, die mit eigens und sehr aufwändig erzeugten Daten trainiert werden und die dann auch immer eine Gefahrenquelle sind, wenn sie fälschlicherweise aktiviert würden. Somit bleibt die Frage, ob Dilemmasituationen in Kombination mit Bedingungen für kontrollierte Reaktionsmöglichkeiten überhaupt häufig genug aufträten, um dafür eine solch komplexe und umstrittene Technik zu implementieren, die auch zusätzliche Risiken birgt. Viel wichtiger als eine genaue Prognose der Häufigkeit solcher Fälle ist mir aber anzudeuten, dass die postulierten Szenarien relativ selbstverständlich ein kontrolliertes Fahrzeug voraussetzen, wobei die dabei implizierte Technologie detaillierter auf ihre (jeweils spezifischen) technologischen und physikalischen Implikationen hin zu durchdenken wäre.

Zugleich stellen sich durch die Anwendung von KI im Verkehr viele politische und ethische Fragen, deren Beantwortung mit großer Wahrscheinlichkeit Auswirkungen auf alle Verkehrsteilnehmer_innen haben werden, die aber wegen des Fokus auf Trolley-Probleme nicht genügend und oder nur unter spezifischen Perspektiven diskutiert werden. Eines der ersten umfassenden Bücher, das soziale, technische und rechtliche Aspekte selbstfahrender Kraftfahrzeuge in Bezug auf aktuelle KI darstellt, behandelt z. B. im Kapitel zur Ethik beinahe nur diese Frage. Weitere Probleme, wie die im Folgenden genannten, bekommen gerade eine Seite Platz.[10] Der Bericht der Ethik-Kommission zum automatisierten Fahren des Bundesministeriums für Verkehr und digitale Infrastruktur kommt gleich im ersten Teil prominent auf Dilemmasituationen zu sprechen – auch wenn er in der Folge weitere wichtige Fragen aufwirft.[11] In der Presse wird das Thema sehr oft als Titel für Texte zu selbstfahrenden Kraftfahrzeugen benutzt. Selbst eine von ver.di veranstaltete Konferenz zum autonomen Fahren bearbeitet diese Frage neben der Veränderung von Arbeitsbedingungen als zweites großes Thema.[12] Ich möchte hier stellvertretend einige wichtige Bereiche kurz zusammenfassen, die alle Verkehrsteilnehmer_innen betreffen dürften, aber im Vergleich weniger Aufmerksamkeit bekommen oder durch die Diskussion um Trolley-Probleme beeinflusst werden.

Letzteres betrifft etwa die in juristischen Kontexten relativ breit geführte Debatte zu Verantwortung und Haftung im Einsatz von selbstfahrenden Kraftfahrzeugen.[13] Hier und auch im Rahmen der entsprechenden Lobbyarbeit[14] treten die Trolley-Probleme immer wieder auf, weil sie dazu dienen können, die Verantwortung von Software und Programmierung zu betonen. Allerdings tauchen sie hier unter falschen Prämissen auf, weil Haftungsfragen sich auf technisches Versagen fokussieren, was ein Versagen des Programms einschließt – statt auf eine falsche ‹ethische› Entscheidung eines korrekt operierenden Programms.

**9** Vgl. Roy: Autonomous Cars Don't Have a «Trolley Problem».

**10** Vgl. Markus Maurer, J. Christian Gerdes, Barbara Lenz u. a. (Hg.): *Autonomous Driving*, Berlin, Heidelberg 2016, 80.

**11** Vgl. Bundesministerium für Verkehr und digitale Infrastruktur (Hg.): Ethik-Kommission. Automatisiertes und Vernetztes Fahren: Bericht Juni 2017, dort datiert 20.6.2017, www.bmvi.de/SharedDocs/DE/Publikationen/DG/bericht-der-ethik-kommission.pdf, gesehen am 21.5.2019.

**12** Vgl. [bendelo (Oliver Bendel)]: Berliner Konferenz zum automatisierten Fahren, dort datiert 17.11.2017, maschinenethik.net/?p=4309, gesehen am 14.4.2019.

**13** Vgl. Georg Borges: Haftung für selbstfahrende Autos, in: *Computer und Recht*, Nr. 32, H. 4, 2016, 272–280.

**14** Vgl. Iagnemma: Why we have the ethics of self-driving cars all wrong.

In Bezug auf die öffentliche Diskussion können Trolley-Probleme aber auch von den hier verhandelten Fragen der Versicherungs- und Haftungskalkulationen ablenken, in der Verkehrstote eine sichere, wenn auch zu minimierende, statistische Größe sind. In den Debatten um die Trolley-Probleme werden sie wieder zu Einzelfällen, die von Algorithmen sorgfältig begutachtet werden.

Ein zentrales Feld ethischer oder politischer Probleme des automatisierten Verkehrs ist die Infrastruktur. Das betrifft Fragen von Besitzverhältnissen, Regulierung, Standardisierung und vieles mehr. Die Frage, ob die Verkehrsinfrastruktur in Zukunft für automatisierte oder von Menschen gesteuerte Verkehrsmittel optimiert wird, betrifft fast alle Menschen in unserer Gesellschaft und ist neben Zugangs- und Gerechtigkeitsfragen auch eine Frage von Leben und Tod.

Ein weiteres wichtiges ethisch-politisches Feld des Einsatzes von KI im Verkehr, von dem durch den Fokus auf die Trolley-Probleme abgelenkt wird, ist der Umgang mit Daten. Beispielsweise ist relativ umstritten, ob es ein ‹Dateneigentum› überhaupt geben kann.[15] Dennoch wird das vernetzte und selbstfahrende Automobil von Unternehmen und Verbänden als zentrales Argument für die Notwendigkeit einer entsprechenden rechtlichen Regelung angeführt.[16] Das verdeutlicht, dass die Debatten und Regelungen bezüglich der KI im Kraftverkehr das Potenzial haben, politische Standards zu setzen, die zukünftig den Umgang mit Daten allgemein strukturieren werden.

## II. Herkunft und Aussagekraft der Trolley-Probleme

Der letzte Abschnitt sollte zeigen, dass der von den Trolley-Problemen vorausgesetzte Fall relativ unwahrscheinlich ist und auf unreflektierten Annahmen beruht. Hier möchte ich wenigstens kurz darauf eingehen, dass, selbst wenn ein solcher Fall eintreten sollte, die Trolley-Probleme keine guten Antworten liefern.[17]

Die aktuelle Debatte bezieht sich auf die Trolley-Probleme in der Form, wie sie die Moralphilosophin Philippa Foot in einem Text von 1967 eingeführt hat.[18] Dort behandelt sie das Verbot der Abtreibung, das auch dann noch gelten soll, wenn das Verbleiben eines Fötus im Mutterleib zum Tod der Mutter führen würde. Für ihr Argument denkt sie sich ein *Paar* von Fällen aus. Der Erste ist das bekannte Trolley-Problem. Eine Straßenbahn fährt auf ein Gleis mit fünf Arbeitern zu, es gäbe die Möglichkeit, die Bahn an einer Weiche auf ein anderes Gleis umzulenken, auf dem sich nur ein Mensch befindet. Dieser Fall wird kontrastiert mit einem Zweiten. Ein Richter weiß, dass ein Unschuldiger angeklagt ist. Vor dem Gerichtsgebäude befindet sich jedoch ein wütender Mob und ohne Verurteilung wird Gewalt ausbrechen, die ebenfalls fünf Menschenleben kosten wird. Die einzige Möglichkeit, das zu verhindern, ist den Unschuldigen hinzurichten. Für das Verständnis von Foots Argument ist es nun wichtig, dass für sie feststeht, dass die Weiche umgestellt werden sollte, aber der Richter keinen Unschuldigen opfern darf.[19] Das Problem, das oft als Trolley-Problem

**15** Vgl. Gerrit Hornung, Thilo Goeble: «Data Ownership» im vernetzten Automobil, in: *Computer und Recht*, Nr. 31, H. 4, 2015, 265–273; Andreas Wiebe: Protection of Industrial Data – a New Property Right for the Digital Economy?, in: *Journal of Intellectual Property Law & Practice*, Vol. 12, Nr. 1, 2017, 62–71.

**16** Vgl. Bundesministerium für Verkehr und digitale Infrastruktur (Hg.): «Eigentumsordnung» für Mobilitätsdaten? – Eine Studie aus technischer, ökonomischer und rechtlicher Perspektive, Berlin 2017, online unter *www.bmvi.de/SharedDocs/DE/Publikationen/DG/eigentumsordnung-mobilitaetsdaten.pdf*, gesehen am 14.4.2019; siehe auch *www.bvdw.org/presse/detail/artikel/bvdw-stellungnahme-daten-sind-nicht-eigentumsfaehig/*, gesehen am 14.4.2019.

**17** Ähnliche Argumente finden sich bei Sven Nyholm, Jilles Smids: The Ethics of Accident-Algorithms for Self-Driving Cars: An Applied Trolley Problem?, in: *Ethical Theory and Moral Practice*, Vol. 19, H. 5, 2016, 1275–1289; Alexander Hevelke, Julian Nida-Rümelin: Selbstfahrende Autos und Trolley-Probleme: Zum Aufrechnen von Menschenleben im Falle unausweichlicher Unfälle, in: *Jahrbuch für Wissenschaft und Ethik*, Nr. 19, H. 1, 2016, 5–24; sowie JafariNaimi: Our Bodies in the Trolley's Path.

**18** Vgl. Philippa Foot: The Problem of Abortion and the Doctrine of the Double Effect, in: *Oxford Review*, Vol. 5, 1967, 5–15.

**19** Vgl. ebd., 8.

diskutiert wird, steht für Foot also gar nicht zur Debatte. Sie interessiert, warum wir beide Fälle unterschiedlich behandeln sollten. Am Ende geht es Foot somit nicht darum, eine Regel für all diese Beispiele zu finden. Vielmehr betont sie die spezifischen Unterschiede zwischen vermeintlich ähnlichen Situationen.[20] Auch in der folgenden philosophischen Debatte[21] sind die Trolley-Probleme als Paar von Fällen gedacht, die Unterschiede zwischen moraltheoretischen Alternativen schärfen. Gleichzeitig erfüllen sie aber auch die ambivalente argumentative Funktion, einen sehr aufgeladenen Sachverhalt – bei Foot: Abtreibung – durch Übertragung auf fiktive Situationen zu entschärfen.[22]

Diese Zusammenfassung zeigt, dass es diverse Gründe gibt, warum die Übertragung der Trolley-Probleme auf selbstfahrende Kraftfahrzeuge problematisch ist. Sie betreffen die Entscheidungen einzelner Personen in konkreten Situationen. Im Falle des selbstfahrenden Kraftfahrzeugs geht es um ein Programm, das basierend auf spezifischen Informationen über die Situation, welche die Sensoren anbieten, alle Fälle entscheiden können soll. Ein solches Programm, so Nyholm und Smids, hat dann aber nicht moralische Fragen in Bezug auf einzelne Handlungen zu beantworten: Es geht um eine im Voraus zu realisierende Programmierung für alle potenziellen, technisch fassbaren Dilemmafälle. Das Zustandekommen eines solchen Programms wäre ein Prozess zwischen verschiedenen Stakeholdern[23] wie Fahrzeugherstellern, Softwarefirmen, Regulierungsbehörden, Verkehrsteilnehmer_innen etc. In dessen Zentrum stünde primär die Verteilung von Verantwortung und Haftung, eine Frage, die in den Trolley-Fällen absichtlich beiseitegelassen wird, um die moralischen Unterschiede zwischen Handlungen herauszustellen.[24]

Auch Unsicherheit schließt Foot explizit aus, um die moralisch relevanten Unterschiede zwischen den jeweiligen Alternativen der beiden von ihr angeführten Konstellationen zuzuspitzen. Dagegen fallen alle Entscheidungen bei der Programmierung von selbstfahrenden Kraftfahrzeugen unter Unsicherheit, weil sie von Sensoren, statistischen Modellen, Prädiktoren etc. abhängen und damit eine Frage von Risikoabwägungen sind.[25] Insofern wird wiederum das Feld der Moral individueller Entscheidungen verlassen und die Frage nach den ethischen und politischen Aspekten des Risikodenkens gestellt. Letzteres ist selbst eine ganz bestimmte epistemische Form mit ihrer eigenen Geschichte, ihren eigenen Problemen und Eigenheiten,[26] welche die bestehenden Praktiken um Fahrzeugbau und Versicherungen prägt.[27] Auch dass eine Debatte zu führen wäre, wie diese Praktiken für das automatisierte Fahren gestaltet werden, wird durch die diskursive Gleichstellung der Programmierung von selbstfahrenden Kraftfahrzeugen mit den moralischen Entscheidungen von hypothetischen Fahrer_innen verschleiert. Wie sich anhand des bisherigen Kraftverkehrs zeigt, setzt sich dabei längst nicht immer die sicherste Variante durch. Zudem ist Sicherheit im Verkehr ungleich verteilt, abhängig von Merkmalen wie Alter, Geschlecht oder Gesundheit,[28] was ebenfalls einer sozialen und politischen Adressierung bedürfte.

**20** Vgl. Foot: The Problem of Abortion, 18.

**21** Vgl. Judith Jarvis Thomson: Killing, Letting Die, and the Trolley Problem, in: *The Monist*, Vol. 59, H. 2, 1976, 204–217.

**22** Diese abstrahierende Funktion des «Weichenstellerfalls» nutzte auch schon Hans Welzels in seinen nicht unkontroversen juristischen Überlegungen zum Notstand im Dritten Reich, was die Relevanz der Frage, wovon wie abstrahiert werden kann, schon sehr deutlich aufscheinen lässt. Vgl. Hans Welzel: Zum Notstandsproblem, in: *Zeitschrift für die gesamte Strafrechtswissenschaft*, Nr. 63, H. 1, 1951, 47–56; sowie reflektierend Heike Stopp: *Hans Welzel und der Nationalsozialismus: zur Rolle Hans Welzels in der nationalsozialistischen Strafrechtswissenschaft und zu den Auswirkungen der Schuldtheorie in den NS-Verfahren der Nachkriegszeit*, Tübingen 2018.

**23** Vgl. Nyholm u. a.: The Ethics of Accident-Algorithms for Self-Driving Cars, 1281.

**24** Vgl. ebd., 1282.

**25** Vgl. ebd., 1286.

**26** Vgl. Bernard E. Harcourt: *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age*, Chicago 2007; François Ewald: *L'état providence*, Paris 1986; Jürgen Link: *Versuch über den Normalismus: wie Normalität produziert wird*, Opladen 1997.

**27** Vgl. Evans: Death in Traffic.

**28** Vgl. ebd.

### III. Menschen, Autos und Künstliche Intelligenzen

In den letzten beiden Abschnitte wollte ich verdeutlichen, dass die Trolley-Probleme nur wenige Erkenntnisse für die Ethik der KI im Allgemeinen und die selbstfahrender Kraftfahrzeuge im Speziellen liefern. Dafür lässt sich aus der Aufmerksamkeit, welche diese Probleme in Wissenschaft, Politik und öffentlichen Debatten bekommen, einiges ableiten über die Art und Weise, wie hier KI konzeptualisiert und verhandelt wird. Das hat zu einem gewissen Teil sicher mit der Faszination der Trolley-Probleme selbst zu tun, ganz unabhängig von KI.[29] Doch es lassen sich hier auch spezifisch mit den Debatten um KI zusammenhängende Aspekte ausmachen.

Anwendungen von KI finden sich bereits in vielen Bereichen, darunter auch solche mit großen Auswirkungen: Sicherheits- und Überwachungstechnologien, Asyl- und Immigrationsverfahren, Entscheidungen in Banken, Versicherungen, Arbeitsämtern und Personalabteilungen, die Auswahl von Werbung und anderen Inhalten im Internet. Während wir also alle zunehmend vom Einsatz der KI betroffen sind, geht es dabei oft um Prozesse oder Entscheidungen, die auch schon vorher innerhalb von Institutionen und Unternehmen verborgen waren. Durch das ‹Internet der Dinge› wird die KI noch mehr zum kaum wahrnehmbaren Teil der uns umgebenden Infrastruktur. Im Vergleich dazu befinden sich selbstfahrende Kraftfahrzeuge in einer interessanten Zwischenposition. Viele Einsatzszenarien selbstfahrender Kraftfahrzeuge sehen diese ebenfalls als Serviceinfrastruktur, die durch Fahrdienstleister betrieben wird. Jedoch ‹übernimmt› die KI dabei eine Tätigkeit, die viele nicht nur selbst ausführen, sondern die zudem noch emotional und kulturell aufgeladen ist. Ein schönes und schnelles Auto zu fahren, ist in der westlichen Kultur des 20. Jahrhunderts eine prominente Art, individuelle Freiheit und ökonomischen Erfolg auszudrücken.

Das legt nahe, Menschen und KI quasi parallel zu setzen und die relevanten Eigenschaften selbstfahrender Kraftfahrzeuge an dieser uns bekannten Stelle ‹hinter dem Steuer› zu suchen. Viele der oben diskutierten Probleme bestehen entsprechend darin, dass die Konsequenzen übersehen werden, die sich daraus ergeben, dass Kraftfahrzeuge Teil der Infrastruktur und zu einem Serviceangebot werden – und genau deshalb eine ganz andere soziotechnische und damit auch ethisch-politische Perspektive benötigen als jene auf die einzelnen Entscheidungen einzelner Fahrer_innen.

Die Parallelsetzung von ‹Mensch hinter dem Steuer› und ‹KI› aktiviert ein Menschenbild, das aus einer bestimmten Kombination kybernetischer und humanistischer Motive besteht. Autofahren ist für Menschen als relativ habitualisierte Tätigkeit möglich. Folglich ist es auch ganz plausibel, dass gerade diese Aufgabe durch künstliche neuronale Netze gelöst werden kann – ein Ansatz der KI, welcher aus der kybernetischen Annahme hervorgegangen ist, dass Lernen durch Feedback zwischen Reizen und Reaktionen erfolgen kann.[30] Eine besondere moralische Begabung ist für das Autofahren somit nicht nötig. Moralisch wird das

**29** Vgl. Christopher W. Bauman, A. Peter McGraw, Daniel M. Bartels u. a.: Revisiting External Validity: Concerns about Trolley Problems and Other Sacrificial Dilemmas in Moral Psychology: External Validity in Moral Psychology, in: *Social and Personality Psychology Compass*, Vol. 8, H. 9, 2014, 536–554.

**30** Vgl. Andreas Sudmann: Szenarien des Postdigitalen, in: Christoph Engemann, Andreas Sudmann (Hg.): *Machine Learning. Medien, Infrastrukturen und Technologien der Künstlichen Intelligenz*, Bielefeld 2018, 55–74.

Fahren dann, wenn das Verhalten der Fahrer_innen maßgeblich zum Entstehen eines Unfalls oder gefährlicher Situationen beiträgt, beispielsweise durch Rasen.

Das bedeutet einerseits eine weitere Runde der kybernetischen Demütigungen, die seit Norbert Wiener die Forschung der Informationstechnologie begleiten:[31] Die von vielen realen wie fiktionalen Helden eingenommene Position am Steuer eines schnellen Wagens lässt sich relativ einfach automatisieren. Andererseits fällt durch die Parallelsetzung von Fahrer_in und KI auf, dass diese Tätigkeit bisher immer von einem Menschen, also einem zur Moral befähigten Wesen, verrichtet wurde. So gesehen erscheint hier ein Mangel von und damit ein Bedürfnis nach Moral für die KI, die Autos fahren soll. Aus der Perspektive einer aus kybernetischen Ansätzen abgeleiteten KI besteht aber der Handlungsspielraum zu moralisch relevanten Verhaltensweisen gar nicht: Eine KI telefoniert nicht beim Autofahren und rast nicht nachts durch die Stadt, um die eigene Identität zu bestätigen.

Daraus folgt nun nicht, dass selbstfahrende Kraftfahrzeuge keine Frage für die Ethik sind. Das wäre nur das humanistische Kippen von einer Betrachtung der KI als mögliche Konkurrenz des Menschlichen hin zur KI als neutrales Werkzeug des Menschen. Vielmehr zeigt sich, dass der Parallelsetzung von einem Menschen hinter dem Steuer mit einer KI eine zweite, kybernetisch-technische Perspektive gegenübergestellt ist, die dazu neigt, ethisch-politische Fragen auszuklammern. Wenn nun mit den Trolley-Problemen ein scheinbar KI-spezifischer, aber moralisch relevanter Handlungsspielraum auftaucht, wird dieser durch ein spezifisches Zusammenspiel beider Perspektiven konfiguriert.

Das lässt sich mit Bezug auf Jutta Webers und Lucy Suchmans Betrachtung der ‹Autonomie› sogenannter autonomer Waffensysteme zeigen. Sie beschreiben, wie das ‹Verhalten› dieser Kampfroboter aus einer komplexen Vielzahl von Entscheidungen, Prozessen und Zufällen entsteht. Dabei durchdringen sich verschiedene institutionelle und technische Logiken, die jeweils unterschiedliche Autonomiebegriffe implizieren. Das betrifft nicht nur die Konstruktion der Waffen selbst. Die Autorinnen lesen ein Strategiepapier des US-Verteidigungsministeriums, in dem sich Konzeptionen der Autonomie als «self-sufficient, adaptive and self-determined performance» einerseits und als «programmed, fully automated execution under perfect human control» andererseits vermischen.[32]

Diese Kombination von Differenzen und Kontinuitäten zwischen menschlich und nicht-menschlich gedachten Konzepten strukturieren viele Anwendungen von KI.[33] In diesem Fall wird die Autonomie der KI da aufgerufen, wo sie sich gegenüber einem menschlichen Soldaten durch kybernetisch gedachtes Verhalten unterscheidet.[34] Ein solcher Unterschied besteht auch bezüglich des nur Menschen zugeschriebenen Handlungsspielraums beim Lenken des Kraftfahrzeugs, der das verlässliche und kontrollierte Fahren durch die KI motiviert. Dieses Verhalten der KI wird nun aber gekoppelt an eine dann wieder menschliche und verantwortliche Stelle, von der im Falle der Kampfroboter die militärischen Entscheidungen über Leben und Tod zu fällen sind. Damit wird die

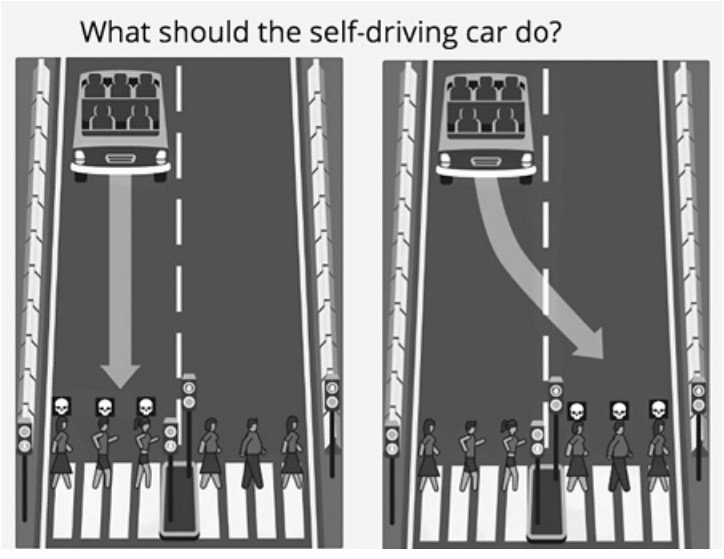
**31** Vgl. N. Katherine Hayles: *How We Became Posthuman*, Chicago 1999, 108.

**32** Jutta Weber, Lucy Suchman: Human–machine autonomies, in: Nehal Bhuta, Susanne Beck, Robin Geiß u. a. (Hg.): *Autonomous Weapons Systems. Law, Ethics, Policy*, Cambridge 2016, 75–102, hier 90.

**33** Vgl. Tobias Matzner: The Human is Dead – Long Live the Algorithm! Human-Algorithmic Ensembles and Liberal Subjectivity, in: *Theory, Culture & Society*, Vol. 36, H. 2, 2019, 123–144.

**34** Vgl. Weber u. a.: Human–machine autonomies, 93.

Illusion eines technisch optimierten Krieges erzeugt, in dem von verantwortungsvollen Offizier_innen gegebene Befehle automatisiert und exakt umgesetzt werden. Auch diese Struktur findet sich beim selbstfahrenden Kraftfahrzeug wieder. Wie im zweiten Abschnitt deutlich wurde, argumentieren Nyholm und Smids für eine Programmierung selbstfahrender Kraftfahrzeuge für Unfälle. Diese soll durch einen entsprechend gestalteten demokratischen Multi-Stakeholder-Prozess gefunden werden, der eine verantwortliche Entscheidung für alle potenziellen Fälle treffen soll.




**Beispielhafter Fall aus dem MIT Moral Machine Experiment. Entschieden werden soll hier, ob das selbstfahrende Auto eine Frau und zwei Athlet_innen (links) oder eine Frau und zwei kräftig gebaute Personen bei Versagen der Bremsen erfasst**

Obwohl Nyholm und Smids die Trolley-Probleme ablehnen, verbleibt ihr Vorschlag innerhalb derselben *boundary practice* zwischen KI und menschlichen Entscheidungen, die einen unvermeidbaren, tragischen Unfall in eine Situation überführt, die richtig gelöst werden kann. Der Schatten des ethisch korrekten Tötens, den besonders Drohnen auf die KI werfen, betrifft also nicht nur Kampfroboter. In gewisser Weise ist der Fall des Kraftfahrzeugs noch prägnanter, weil sich bezüglich einer kriegerischen Handlung ja immer noch fragen lässt, ob militärische Entscheider überhaupt die Macht haben sollten, die sie durch KI bekommen sollen.[35] Für Autos scheint die Entscheidung notwendig, weil der Unfall ja als unvermeidlich postuliert ist, wobei die Unvermeidbarkeit durch die kybernetische Perspektive auf die KI garantiert wird.

Die spezifische Kombination der beiden Vorstellungen von Verhalten als menschlich-moralisch sowie kybernetisch reduziert Ethik somit auf eine Frage deterministisch ausgeführter Vorentscheidungen. Diese *Ethik als Konfiguration* beruht zu wichtigen Teilen auf einer an der KI ausgerichteten Epistemologie. Die schnellere Wahrnehmung und Verarbeitung der Situation wurden schon angesprochen. Aber auch die spezifischen Formen von Mustererkennung und Klassifizierung durch Künstliche Intelligenzen spielen eine Rolle. Das zeigt sich am «MIT Moral Machine Experiment». Hier handelt es sich um eine in *Science* und *Nature* prestigeträchtig publizierte Studie, deren Teilnehmer_innen in einer Reihe von Fällen jeweils entscheiden sollen, ob ein Auto ausweichen oder auf seiner Bahn bleiben soll. In allen Fällen gibt es Todesopfer, die in sehr grobe Kategorien unterteilt werden, darunter «normale» (unmarkierte) und «füllige» (*large*) Menschen, Athlet_innen, Führungskräfte (*executives*), Obdachlose, Kriminelle, Kinder und Babys. Alle bis auf Obdachlose und Kriminelle haben ein binäres Geschlecht, dazu kommen (geschlechtslose) Hunde und Katzen.[36]

**35** Vgl. Brianna Rennix, Nathan J. Robinson: The Trolley Problem Will Tell You Nothing Useful About Morality, in: *Current Affairs*, dort datiert 3.11.2017, *editor.currentaffairs.org/2017/11/the-trolley-problem-will-tell-you-nothing-useful-about-morality/*, gesehen am 12.11.2018.

**36** Vgl. Edmond Awad, Sohan Dsouza, Richard Kim u. a.: The Moral Machine Experiment, in: *Nature*, Vol. 563, H. 7729, 2018, 59–64, hier 61.

Würde eine solche Kategorisierung in anderen Kontexten zur Grundlage von Entscheidungen über Leben und Tod, klänge das verdächtig nach einer Unterteilung in mehr oder weniger wertes Leben. In Bezug auf eine KI scheint die Kategorisierung plausibel, da es ja gerade Klassifizierungsverfahren sind, welche den Erfolg vieler maschineller Lernverfahren ausmachen.[37] Die Kategorisierung verliert also ihre andernorts verdächtigen Aspekte durch die Assoziation mit der Funktionsweise KI-basierter Entscheidungen. Die ganze Studie ist durch autonomes Fahren motiviert. Die Frage hier ist nicht, wer bei einem Unfall sterben sollte, sondern: «What should the self-driving car do?»[38] Es handelt sich also um eine Umfrage über menschliche Intuitionen unter den Bedingungen einer nicht explizit ausgeführten, bei den Proband_innen aber mit aufgerufenen Epistemologie und der technischen Möglichkeiten des KI-basierten Fahrens. Die Kategorien werden sachlich und ohne weitere Begründung eingeführt. Immer geht es hier entsprechend der Logik der Trolley-Probleme um Gegensatzpaare, wie z. B. hoher und niedriger sozialer Status.[39] Dass hoher sozialer Status im Experiment durch die Figur *executive* abgebildet wird, aber nicht durch *athlete*, und ähnliche Entscheidungen bleiben selbst im Artikel implizit.

Bereits in dieser Studie zeigen sich also die viel diskutierten Probleme der KI, die Herkunft der verwendeten Kategorien sowie ihre Bedeutung und Operationalisierung im Einsatzkontext zu verschleiern.[40] Darüber hinaus fehlt die Reflexion der Unterschiede zwischen der in der Studie aufgerufenen Kategorisierung und deren Funktion in KI-basierter Mustererkennung. Selbst unter der Annahme, dass es eine KI gäbe, welche die aufgeführten Kategorien stabil und effizient erkennen könnte, geschähe das aufgrund ganz anderer Merkmale und Prozesse, als Menschen das tun.[41] Zudem sind die von künstlichen neuronalen Netzen gebildeten Kategorien nicht einfach Repräsentationen bekannter Kategorien durch ein konnektionistisches Modell. Stattdessen entstehen durch das Training neue Kategorien, die datengetrieben gebildet und statistisch verifiziert sind. Werden diese als Repräsentation menschlicher Kategorien oder Intuitionen gesehen, können diese Prozesse und die damit verbundene Epistemologie zur Legitimierung von Kategorisierungen beitragen, wie sie eben in der Studie des MIT und ihrem Erfolg aufscheint.[42] Diese durch die Funktionsweise der KI legitimierten Kategorien treffen sich hier nun mit den Zuspitzungen auf eine moralische relevante Eigenschaft, welche die Trolley-Probleme bestimmen. Auch dort werden soziale Positionen, wie Richter_innen, Straßenbahnfahrer_innen, Arbeiter_innen, referenziert, ohne sie in ihrer sozialen oder kulturellen Fülle zu betrachten.

Solche Kategorisierungen werden nochmals folgenreicher, wenn sie mit KI-basierter Prädiktion in Verbindung stehen, die automatisiert zu einem bestimmten Verhalten führt. Unter den Bedingungen KI-basierter Vorhersagen gibt es nur noch ganz bestimmte, durch die Möglichkeiten des Modells beschränkte Pfade in die Zukunft. Diese sind mit unterschiedlichen Wahrscheinlichkeiten belegt und werden anhand dessen evaluiert, was im Modell (angenommen

**37** Vgl. Sudmann: Szenarien des Postdigitalen.

**38** Awad u. a.: The Moral Machine experiment, 59, vgl. auch 60, Abbildung 1b.

**39** Vgl. ebd., 60.

**40** Vgl. Cathy O'Neil: *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, New York 2017.

**41** Vgl. Tobias Matzner: The Model Gap. Cognitive Systems in Security Applications and Their Ethical Implications, in: *AI & Society*, Vol. 31, H. 1, 2016, 95–102.

**42** Vgl. Tobias Matzner: Beyond Data as Representation. The Performativity of Big Data in Surveillance, in: *Surveillance & Society*, Vol. 14, H. 2, 2016, 197–210.

ein künstliches neuronales Netzwerk) Aktivierungen verursacht.[43] Sowohl die vielen kontingenten Möglichkeiten der Umgebung, als auch die des Modells (Fehler) bleiben unberücksichtigt. Die sich selbst erfüllende Zukunft, von der Brian Massumi in Bezug auf präventives Handeln spricht, wird hier, so gut es die Möglichkeiten der Technologie zulassen, herbeigeführt.[44] Der dicke Mann muss sterben, weil berechnet wurde, dass er sterben muss.

## IV. Schluss

Die Herausforderung einer Ethik der KI besteht also darin, das komplexe soziotechnische Zusammenspiel zu durchdringen, das eine vermeintlich recht eindeutige Funktion, wie das Lenken eines Autos, hervorbringt. Das braucht eine andere Form des Nachdenkens als das Suchen nach moralischen (Subjekt-) Positionen, die als Äquivalent zum Menschen hinter dem Steuer fungieren. Einerseits werden damit soziopolitische Fragen ausgeblendet. Andererseits zeigt eine medien- und technikkritische Analyse, dass die Trolley-Probleme ein Mensch-Technik-Verhältnis perpetuieren, welches die Idee einer Ethik als Konfiguration auch dann noch strukturiert, wenn die Trolley-Probleme als Ansatz für Dilemmasituationen kritisiert werden. Eine ethische und politische Debatte über KI müsste auch hier ansetzen, unter anderem weil dieses Mensch-Technik-Verhältnis ein ethisch korrektes Töten impliziert und problematische Formen kategorisierender und wertender Epistemologien legitimiert.

**43** Es läge nahe, hier zu schreiben, dass die Situation eben nur aufgrund der Kategorien evaluiert wird, die sich im Modell finden. Aber das wäre zu eng an der Idee einer, wenngleich vereinfachten, Repräsentation der Welt im Modell, die hier ja gerade vermieden werden soll.

**44** Vgl. Brian Massumi: Potential Politics and the Primacy of Preemption, in: *Theory & Event*, Vol. 10, H. 2, 2007, doi:10.1353/tae.2007.0066, 0 S.