Lehrstuhl für Wissenschaftliches Rechnen
Fakultät für Informatik
Technische Universität München

Maschinen- und Roboterethik: (draft)
Die komplexe Ethik autonomer Kraftfahrzeuge

Florian Schmidt[1, ✉]
[1]Fakultät für Informatik, Technische Universität München (TUM), Boltzmannstr. 3, 85748 Garching, Deutschland
✉ fs.schmidt@tum.de
16. Juli 2020

Das autonome Fahren ist eine sehr gegenwärtige wissenschaftliche, wie technische Errungenschaft, die verspricht, innerhalb der nächsten Dekaden den Straßenverkehr umfassend zu revolutionieren. Am Horizont stehen in diesem Kontext selbstund leer fahrende geteilte Fahrzeuge im Rahmen eines modernisierten Carsharings, die innerhalb von Großstädten maßgeblich zur Entzerrung der fahrzeuggefüllten Innenstädten beitragen können – und auf dem Land selbstredend auch die Verkehrsinfrastruktur ausbauen könnten.

Bis entsprechende Fahrzeuge allerdings vollständig autonom auf den Straßen dieser Welt unterwegs sein können, sind allerdings noch einige Fragen offen. Fragen, auf die es möglicherweise auch keine schwarz-weißen, generalisierbaren Antworten geben könnte – ja womöglich auch nicht geben kann. Fragen, die ethisch und moralisch höchst prekäre Entscheidungen einer algorithmisch denkenden oder künstlich intelligenten Maschine abverlangen, die den Kontext der Situation womöglich gar nicht verstehen kann. Fragen, denen sich die Gesellschaft früher oder später stellen muss, und für die im besten Fall eine globale Lösung gefunden werden könnte.

Die folgenden Überlegungen beschäftigen sich vorrangig mit ebendiesen ethischen Überlegungen des autonomen Fahrens.

1 Grobeinordnung in die Maschinenethik

Zu Beginn des Artikels wollen wir uns grundlegend mit der Maschinenethik beschäftigen, um die vorliegende Spezialisierung in den richtigen Kontext einzuordnen.

1.1 Definition und Abgrenzung

Grundsätzlich beschäftigt sich die Maschinenethik als solches mit den Konzepten der maschinellen Moral beziehungsweise der moralischen Maschine – also mit der Überlegung, wie das doch relativ abstrakte Konzept der Moral mit der konkreten, technischen Maschine in Einklang zu bringen ist.

Kern der Überlegung ist ein naheliegender Gedanke: Mit der Forschung auf dem Gebiet der künstlichen Intelligenz werden technische Systeme geschaffen, die Anzeichen von Intelligenz nachweisen (sollen). Ist eine Maschine nun derart intelligent, liegt es dementsprechend auch nahe, dass sie auch Anzeichen von einer Moralvorstellung aufweisen könnte [1, S. 3f.].

Es zeigt sich, dass die beiden Phänomene der Intelligenz sowie der Moral zwar per se unabhängig sind, aber in ihrer Existenz dennoch korrelieren (siehe Abbildung 1 auf der nächsten Seite). Ähnlich verhält es sich auch mit den Disziplinen der Maschinenethik und der Forschung an der künstlichen Intelligenz. Genauso wie die Maschinenethik an moralischen Maschinen forscht, versucht die Gegenseite künstlich intelligente Systeme zu erschaffen.

Analog kann man die Differenzierung zwischen der schwachen und starken Eigenschaft von der Forschung zur künstlichen Intelligenz übernehmen. Wir sprechen von schwacher maschineller Moral wie schwacher KI, wenn die entsprechende Eigenschaft simuliert oder in Teilen dem Vorbild nachgeahmt wird; erst die starke KI oder starke maschinelle Moral strebt danach, die Eigenschaft vollständig zu erreichen [1, S. 17]. Die starke KI stellt hiermit also eine Art *allgemeine* künstliche Intelligenz dar – spätestens hier sind Überlegungen zur Moral in jedem Fall angebracht[1].

Abgegrenzt wird die Maschinenethik in der Regel von den mit ihr verwandten Bereichsethiken

- der **digitalen Ethik**, die sich im Kern mit informationstechnologischen Systemen und den damit einhergehenden Überlegungen zu informationeller Autonomie auseinandersetzt, sowie

[1]Offen bleibt natürlich die Frage, ob die menschliche Moral in diesem Kontext das erstrebenswerte Ziel ist. Es kann ebenso Ziel sein, eine Art „Moral“ zu erzeugen, die sich gänzlich von der des Menschen unterscheidet [1, S. 23].

Abbildung 1 Begriffliche Abgrenzung der maschinellen Moral und künstlichen Intelligenz [1, S. 17].

- der **Technikethik**, die sich allgemeiner gefasst mit technischen und wissenschaftlichen Entwicklungen beschäftigt und ebendiese mit ethischen Wertungen versieht.

1.2 Kernfragen

Die zentrale Idee der Maschinenethik ist also, die Maschine also als Subjekt und nicht nur als Objekt der Moral zu sehen – sprich die Maschine im Sinne der Moral auf eine dem Menschen gleichgestellte Ebene zu heben.

Die Notwendigkeit dieser Überlegungen ergibt sich direkt aus der wachsenden Autonomie der Maschinen. Trifft eine Maschine Entscheidungen, die ohne Zutun eines Menschen erfolgen, ist die Moral automatisch relevant. Erst recht gilt dies, sofern dieser Maschine eine Entscheidungsgewalt über Leben und Tod zusteht.

Evident ist, dass für diese Art von extremer Entscheidungsfindung ein gewisses Verständnis des Kontext der Handlung nötig ist. Konzepte wie Leben und Tod, Bewusstsein und Menschenwürde sind schwierig in imperative Programmzeilen zu encodieren – genau das verlangen wir allerdings von moralisch entscheidenden Maschinen.

Die Disziplin stellt sich im Prinzip drei Kernfragen, die maßgeblich die Überlegungen charakterisieren (nach [1, S. 13ff.]):

1. **Wie kommt die Moral in die Maschine?**

   Wie können wir es schaffen, das abstrakte Konzept einer „Moral“ in ein für Maschinen verständliches Konzept zu verwandeln, inklusive aller Nuancen und Kontextüberlegungen die daraus folgen?

2. **Wie viel Entscheidungsgewalt überlassen wir Maschinen?**

   Selbst, wenn wir von dem Vorhandensein einer maschinellen Moral ausgehen, bleibt die Frage offen, wie viel Macht wir dieser Maschine überlassen wollen. Ist es erstrebenswert, dass (vor allem in militärischen Anwendungsgebieten) die Maschinen vollständig autonom agieren können?

3. **In welcher Form trägt die Maschine Verantwortung über ihr Handeln?**

   Unklar ist ebenfalls, inwieweit eine Maschine retrospektiv für eine getroffene Entscheidung oder durchgeführte Handlung zur Rechenschaft gezogen werden kann. Zur Illustration sei die folgende Frage gestellt: Bis wir zu vollständig selbstlernenden Maschinen gekommen sind, sind auch die fortschrittlichsten Computer in gewisser Weise an ihre Algorithmik oder ihre Trainingsdatenmenge gebunden; Kann man hier von Verantwortung sprechen?

1.3 Anwendungsgebiet autonomes Fahren

Im Folgenden möchten wir uns auf das Anwendungsgebiet des autonomen Fahrens konzentrieren. Dieses berührt alle drei der im vorigen Abschnitt aufgeführten Kernfragen, und ist nebenbei auch noch sehr Alltagsrelevant. Wir haben als Gesellschaft viel Berührung mit den Überlegungen, die auf den kommenden Seiten folgen werden. Gerade durch die Möglichkeiten einer zukünftig durch hochautomatisierte Fahrzeuge befahrene Innenstadt sprechen wir eben nicht von weit entfernten militärischen Operationen, sondern vom alltäglichen Straßenverkehr.

Die Motivation der technischen Entwicklung ist nach einem kurzen Blick in die Verkehrsunfallstatistik auf deutschen Straßen relativ klar erkenntlich.

Von den in 2018 ca. 2,5 Millionen polizeilich erfassten Verkehrsunfällen fußten ca. $88,4\,\%$ maßgeblich auf dem Fehlverhalten der involvierten Fahrzeugführer (aus [2], [3], [4]). Das zeigt deutlich: der Faktor Mensch verschwindet vorerst nicht aus dem Straßenverkehr. Immer und überall da, wo der Mensch beim Fahren konkrete Aktionen durchführt – sei es Überholen, Einfädeln oder Abbiegen – werden Fehler passieren. Aus diesem Grund helfen seit geraumer Zeit diverse Assistenzsysteme im Fahrzeug mit. Diese tragen maßgeblich zu der bis dato utopischen Vorstellung bei, dass wir mit deren Hilfe effizienter, eleganter und entspannter unterwegs sein werden. Dazu müssen die Systeme ja nicht perfekt fahren – sie müssen lediglich statistisch besser (und sicherer) fahren als wir.

2 Technische Automatisierungsstufen nach der SAE

Zur Konkretisierung wird die technische Entwicklung an dieser Stelle in der Regel in sechs spezifische Automatisierungsstufen (auch *Levels*) aufgeteilt. Diese stellen aufsteigend das Fortschreiten von vollständig manuellem Fahren bis hin zur vollständigen Autonomie des Fahrzeugs dar.

Die Idee hinter einer derartigen Klassifizierung ist, dass hierdurch die Klärung der anfallenden rechtlichen und verantwortungstechnischen Fragen einfacher fällt. Es lässt sich nun sehr einfach sagen, ab welchem Punkt die Maschine hier vorrangig die Verantwortung für ihr eigenes Handeln trägt – eine Information, die in (straf-)rechtlichem Kontext durchaus wertvoll sein kann.

Im Folgenden möchten wir auf diese Automatisierungsstufen genauer eingehen (nach [6], [7]).

2.1 Levels 0–2: menschlich gelenktes Fahren

Level 0. Die überwiegende Mehrheit der Fahrzeuge auf deutschen Straßen ist mit der Automatisierungsstufe 0 unterwegs: als Selbstfahrer, bzw. „Driver only“. Hier übernimmt ganz klassisch der menschliche Fahrer alle Aspekte der Fahraufgabe[2], und ist dementsprechend natürlich auch vollständig für sein eigenes Handeln verantwortlich.

Level 1. Viele moderne Neuwagen bewegen sich nun mindestens auf dem Gebiet der Automatisierungsstufe 1, indem sie ihrem Fahrer bestimmte unterstützende und vor allem auf der Langstrecke der Ermüdung entgegenwirkende Assistenzsysteme anbieten. Dies kann beispielsweise ein adaptiver Tempomat sein, welcher im Kontrast zu einem regulären nichtadaptiven Tempomaten die eingestellte Geschwindigkeit unterschreitet, um den eingestellten Sicherheitsabstand zum vorausfahrenden Fahrzeug zu halten. Konkret geht es hierbei darum, dass die Technik jeweils ausschließlich entweder die Quer-, oder die Längssteuerung des Fahrzeugs übernehmen kann – aber nicht beides gleichzeitig. Auch Spurwechselassistenten im Sinne von Totwinkelwarnern und Spurhalteassistenten mit der Möglichkeit zu einem korrigierenden, aber isolierten Lenkeingriff stellen hier also wertvolle Fahrhilfen dar. Eine Fahrhilfe ist allerdings genau das, was diese Systeme bieten: eine Hilfe beim Fahren für den eigentlichen Fahrer.

Der Mensch fährt immer noch selbst und muss demnach bei der Verwendung der Systeme dauerhaft wachsam bleiben und jederzeit zur vollständigen und sofortigen Übernahme der Steuerung bereit sein. Somit liegt auch hier wie bei Level-0-Fahrzeugen die Verantwortung für Fahrhandlungen eindeutig beim Fahrzeugführer.

Level 2. Als Beispiele für Fahrzeuge der Automatisierungsstufe 2 können sehr anschaulich die Fahrzeuge der Firma *Tesla Motors* dienen. Diese sind mit dem *Autopiloten* ausgestattet, der alle Level 2-Features anschaulich demonstriert [8]:

Diese Fahrzeuge sind gemäß der Spezifikation im SAE-Standard in der Lage, im Gegensatz zum Level 1 nun sowohl die Längs-, als auch die Querlenkung gleichzeitig zu übernehmen. Das System muss zwar vom Fahrer noch überwacht werden, kann allerdings in bestimmten Fahrsituationen (beispielsweise auf der Autobahn) das Fahrzeug eben autonom in der Spur halten und auch Gas und Bremse selbstständig bedienen [7, S. 1]. Dies geschieht durch eine Mischung aus Kameradaten, lokaler Sensorik

[2] Was nicht bedeutet, dass keine technischen Features vorhanden sein können – diese haben dann allerdings lediglich nur informierende oder isoliert eingreifende Wirkungen.

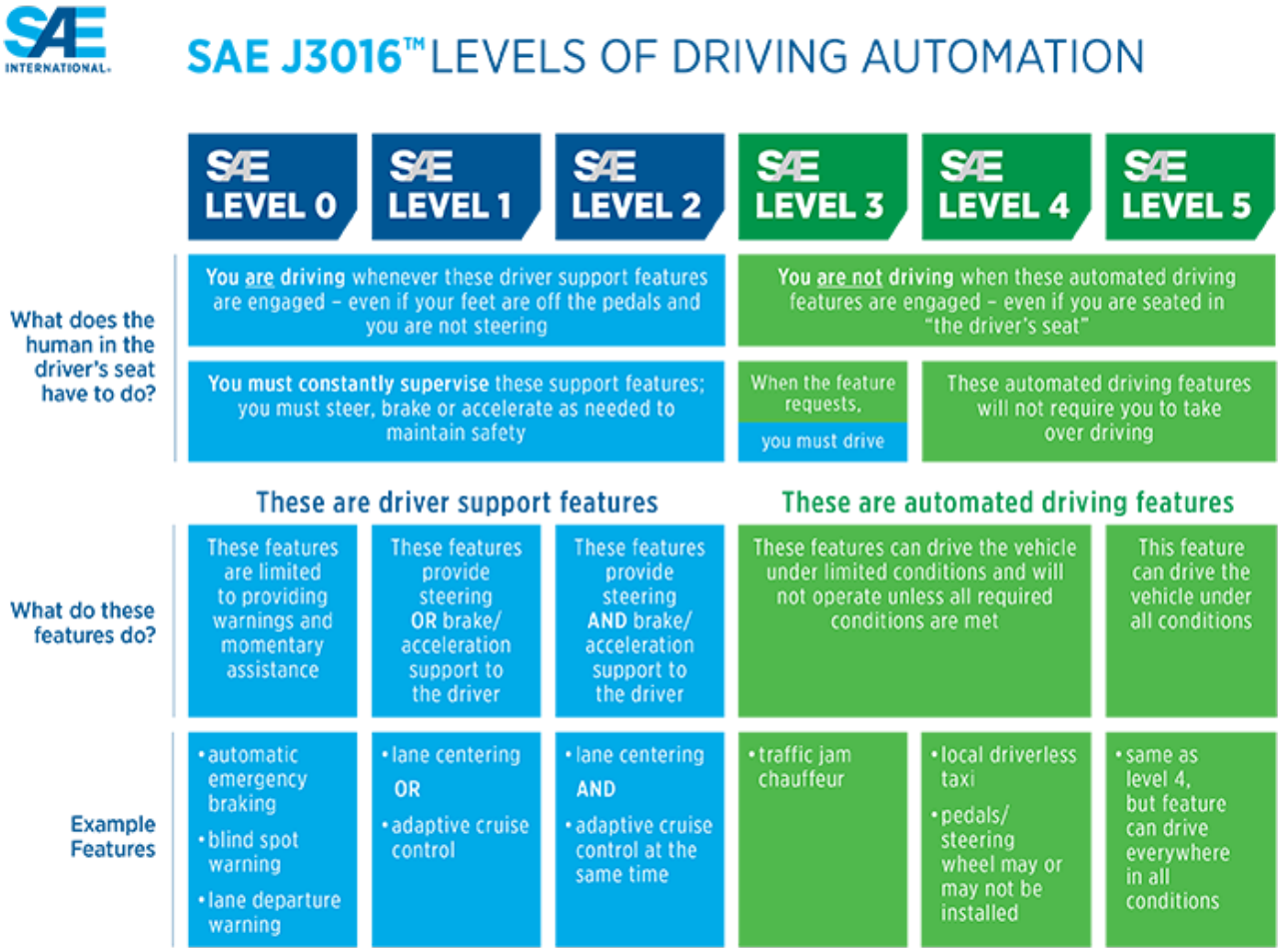

Abbildung 2 Die Automatisierungsstufen nach SAE-Standard J3016 (aus [5]).

Abbildung 3 Ein autonom fahrender Tesla Model 3 auf einem US-amerikanischen Highway [9].

und GPS-basierten Kartendaten aus dem Navigationssystem.

Darüber hinaus kann es bedingt durch die künstlich-intelligente Struktur der Programmierung während der Fahrt aus seinen eigenen Fehlern lernen und mit der Zeit bessere Fahrverhalten entwickeln. Begegnet das System allerdings einer Fahrsituation, mit der es nicht selbstständig umgehen kann, muss es die Steuerung sofort und ohne Verzögerung vollständig an den menschlichen Fahrer abgeben können.

Somit steht auch hier der Fahrer bei allen Fahrsituationen oder im Falle eines Unfalls in der Verantwortung über die Aktionen des teilautonomen Fahrzeugs, und muss im Falle einer rechtswidrigen oder gefährlichen Situation unbedingt eingreifen.

2.2 Levels 3–5: technisch gelenktes Fahren

Aus ethischer Sicht interessant wird die Betrachtung allerdings erst richtig ab der nächsthöheren Automatisierungsstufe 3. Ab hier bewegen wir uns nämlich in Gebieten, in denen das Fahrzeug zumindest streckenweise die Verantwortung für sein eigenes Handeln übernehmen muss.

Level 3. Ab hier übernimmt das autonome System nämlich in konkret abgesteckten Fahrsituationen komplett die Beobachtung der Umgebung sowie die angemessene Reaktion auf äußere Impulse.

Dabei kennt das System seine eigenen Grenzen und muss in der Lage sein, bei jeder auftretenden Situation risikominimierend zu wirken [7, S. 1]. Der menschliche Fahrer muss zwar noch im Fahrzeug anwesend sein und in einem angemessenen Zeitraum auf eine eventuelle Anfrage des Fahrzeugs zur Übernahme der Steuerung eingehen [10, S. 8], entzieht sich allerdings innerhalb der Funktionsgrenzen des Systems vollständig der Verantwortung.

Als anschauliches Beispiel können wir hierzu den sogenannten *Staupiloten* im Audi A8 heranziehen. Dieser kann, so zumindest die Spezifikation, auf Autobahnen im Stau oder bei Kolonnenverkehr mit Geschwindigkeiten unter 60 km/h vollständig das Steuer übernehmen. Bemerkenswert ist hier demnach, dass der Fahrer somit seine Verantwortung an Audi abgibt: er kann die Hände dauerhaft vom Lenkrad und die Füße von den Pedalen nehmen und sich einer anderen Beschäftigung widmen. Soweit leider zumindest nur die Theorie – hier war die technologische Entwicklung schneller als die Gesetzgebung in Europa: nachdem die Zulassung für Fahrzeuge dieser Klasse bis heute nicht in Aussicht steht, hat der Fahrzeughersteller aus Ingolstadt die Pläne für das System nun gestrichen [11] – das Prinzip steht allerdings natürlich trotzdem.

Level 4, Level 5. Die folgenden beiden höchsten Automatisierungsstufen 4 und 5 stellen die restliche Übernahme der Fahrerrolle durch die Technik dar: bei Level 4 muss noch ein Fahrer anwesend sein, der in nicht-definierten Fahrsituationen – also bei Level 3 der Regelfall, hier der Ausnahmefall – übernehmen kann, während Level 5 die vollständige Autonomie darstellt, in der das Auto selbst die kompletten Fähigkeiten eines menschlichen Fahrers ersetzt und, so weit die Theorie, auch leer und vollkommen eigenständig fahren kann [10, S. 8].

3 Unfallprävention und Unfallfolgenminimierung

Spätestens hier eröffnen sich allerdings eine schwierige ethische Dilemmas, die so beim menschlichen Fahren aufgrund der unterschiedlichen Natur des menschlichen beziehungsweise maschinellen Treffens von Entscheidungen nicht auftreten können.

Weiterhin ist klar: auch bei der Prävalenz von autonomen Fahrzeugen auf den Straßen sind Unfallsituationen unausweichlich.

Der Unfallfaktor Mensch wird in naher Zukunft auch bei einem Popularitätszuwachs der autonomen Fahrzeuge nicht verschwinden, da sich auf einen nicht absehbaren Zeitraum menschliche und elektronische Fahrer die Straße teilen werden (auch „gemischter Verkehr“ genannt) [12, S. 1278].

In einer idealen, rein autonomen Fahrwelt ist die Unfallprävention nur bei Systemfehlfunktionen relevant, da Maschinen bei fehlerfreier Programmierung intrinisch fehlerfrei handeln. Von Maschine zu Maschine ist eine Kommunikation der Fahrzeuge untereinander durchaus denkbar und zum kooperativen Informationsaustausch auch in Zukunft angedacht; das noch menschlich gefahrene Auto ist allerdings für das elektronische System immer eine Komponente voller Überraschungen, die selbst fortgeschrittene prediktive Algorithmen nicht umfassend umreißen können.

3.1 Strategien in Systemgrenzbereichen

In den funktionalen Grenzbereichen der aktuellen, teilautonomen Systeme besteht die Vorgehensweise zur Gefahrenminimierung beziehungsweise Unfallvermeidung fast immer in der Kontrollübergabe an den zwangsweise noch vorhandenen menschlichen Fahrer, welcher anschließend durch seine Erfahrung und seinen Instinkt die Situation entschärfen kann.

Problematisch wird dies allerdings, sobald man sich in höhere Automatisierungsniveaus bewegt: Selbst wenn noch ein menschlicher Fahrer im Auto vorhanden ist, ist die Übergabe im Falle einer gefährlichen Situation oder gar eines Unfalls oft nicht rechtzeitig möglich und die Reaktionszeit eines vorher völlig rechtmäßigerweise abgelenkten Fahrers viel zu hoch, um hier präventiv zu wirken. Darüber hinaus stellt sich spätestens in Level-5-Fahrzeugen ein menschliches Eingreifen mangels Lenkrad und Pedalen als eher diffizil heraus.

Daher ist eines klar: das Fahrzeug muss wissen, wie es sich im Falle des Unfalls zu verhalten hat. Unfallvermeidung ist aufgrund der Einzigartigkeit der Situationen speziell im menschlich-mechanisch gemischten Verkehr nicht immer möglich und Unfallfolgenminimierung setzt konkrete Handlungen voraus [13, S. 71].

3.2 Reaktionen vs. aktive Entscheidungen

3.3 Deterministische Perspektiven

4 Typische Abwägungsszenarien

4.1 Die Parallele zum Trolley-Problem

4.2 Ethische Beispielargumentation

4.3 Versuch generalisierbarer Lösungsansätze

4.4 Die Moral Machine des MIT

5 Lösungsansätze der deutschen Ethikkommission

Literatur

[1] Oliver Bendel. *Handbuch Maschinenethik*. Springer, 2019.

[2] Kraftfahrtbundesamt. “Jahresbilanz des Fahrzeugbestandes am 1. Januar 2020”. In: (2020). URL: https://www.kba.de/DE/Statistik/Fahrzeuge/Bestand/pseudo_bestand_node.html.

[3] Statistisches Bundesamt (Destatis). “Verkehrsunfälle in Deutschland (Grafiken)”. In: (2020). URL: https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Verkehrsunfaelle/_inhalt.html.

[4] Statistisches Bundesamt (Destatis). “Verkehrsunfälle (Fachserie 8 Reihe 7)”. In: (2019). URL: https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Verkehrsunfaelle/Publikationen/Downloads-Verkehrsunfaelle/verkehrsunfaelle-jahr-2080700187004.pdf.

[5] URL: https://www.sae.org/binaries/content/gallery/cm/articles/press-releases/2018/12/j3016-levels-of-automation-image.png.

[6] SAE. “Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles”. In: *SAE International (J3016)* (2016).

[7] Bundesanstalt für Straßenwesen (BASt). “Rechtsfolgen zunehmender Fahrzeugautomatisierung”. In: (2012). URL: http://www.bast.de/DE/Publikationen/Foko/Downloads/2012-11.pdf?__blob=publicationFile.

[8] *Tesla - Autopilot*. 2020. URL: https://www.tesla.com/autopilot.

[9] *Tesla Autopilot*. URL: https://electrek.co/wp-content/uploads/sites/3/2019/10/Tesla-Autopilot-hero-4-e1570845324247.jpg.

[10] Conference of European Directors of Roads. *CEDR Transnational Road Research Programme - Call 2014: Mobility & ITS*. (Zuletzt heruntergeladen: 17.12.2018). Dezember 2014. URL: http://www.bast.de/DE/BASt/Forschung/Forschungsfoerderung/Downloads/cedr_call_2014_2.pdf?__blob=publicationFile&v=2.

[11] Herbie Schmidt. “Audi steckt beim autonomen Fahren zurück”. In: *Neue Zürcher Zeitung* (2020). URL: https://www.nzz.ch/mobilitaet/auto-mobil/autonomes-fahren-stufe-3-audi-verzichtet-im-a8-auf-staupilot-ld.1553933.

[12] Sven Nyholm und Jilles Smids. “The Ethics of Accident-Algorithms for Self-Driving Cars: an Applied Trolley Problem?” In: *Ethical Theory and Moral Practice* 19.5 (2016), S. 1275–1289. DOI: 10.1007/s10677-016-9745-2. URL: https://doi.org/10.1007/s10677-016-9745-2.

[13] Markus Maurer et al. *Autonomous Driving - Technical, Legal and Social Aspects*. Springer, 2016. URL: https://link.springer.com/book/10.1007/978-3-662-48847-8.