

Maschinen- und Roboterethik: (draft version 198-1229)

Die komplexe Ethik autonomer Kraftfahrzeuge

Florian Schmidt¹, ✉

¹Fakultät 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 selbst- und 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 angebracht¹.

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

¹Offen 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].

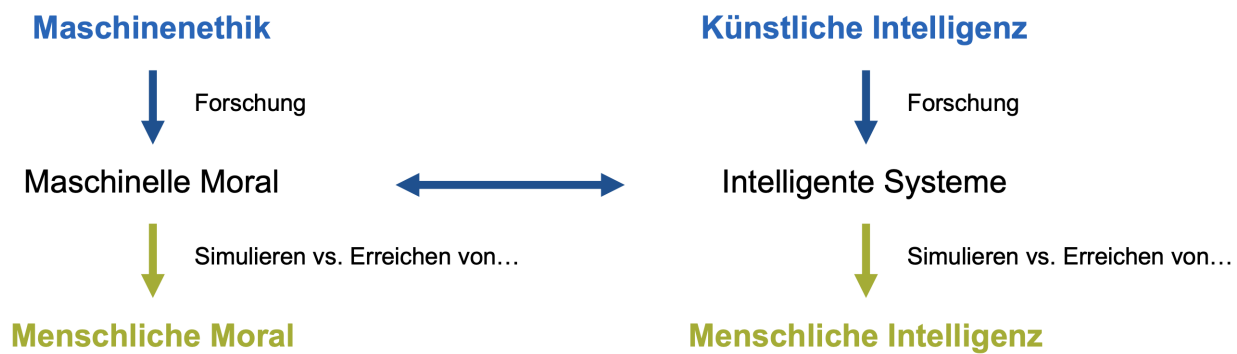


Abbildung 1 Begriffliche Abgrenzung der maschinellen Moral und künstlichen Intelligenz [1, S. 17].

- 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 Maschineneethik 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 [6], [7]. 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.

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 Fahraufgabe², 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. Auch Spurwechselassistenten im Sinne von Totwinkelwarnern und Spurhalteassistenten mit der Möglichkeit zu einem korrigierenden, aber isolierten Lenkeingriff stellen hier 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. Hier geht's weiter...

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



SAE J3016™ LEVELS OF DRIVING AUTOMATION

	SAE LEVEL 0	SAE LEVEL 1	SAE LEVEL 2	SAE LEVEL 3	SAE LEVEL 4	SAE LEVEL 5
What does the human in the driver's seat have to do?	You <u>are</u> driving whenever these driver support features are engaged – even if your feet are off the pedals and you are not steering			You <u>are not</u> driving when these automated driving features are engaged – even if you are seated in “the driver’s seat”		
	You must constantly supervise these support features; you must steer, brake or accelerate as needed to maintain safety			When the feature requests, you must drive	These automated driving features will not require you to take over driving	
	These are driver support features			These are automated driving features		
What do these features do?	These features are limited to providing warnings and momentary assistance	These features provide steering OR brake/acceleration support to the driver	These features provide steering AND brake/acceleration support to the driver	These features can drive the vehicle under limited conditions and will not operate unless all required conditions are met	This feature can drive the vehicle under all conditions	
Example Features	<ul style="list-style-type: none">• automatic emergency braking• blind spot warning• lane departure warning	<ul style="list-style-type: none">• lane centering OR• adaptive cruise control	<ul style="list-style-type: none">• lane centering AND• adaptive cruise control at the same time	<ul style="list-style-type: none">• traffic jam chauffeur	<ul style="list-style-type: none">• local driverless taxi• pedals/steering wheel may or may not be installed	<ul style="list-style-type: none">• same as level 4, but feature can drive everywhere in all conditions

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

2.2 Levels 3–5: technisch gelenktes Fahren

3 Unfallprävention und Unfallfolgenminimierung

3.1 Strategien in Systemgrenzbereichen

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.