3. Unfallalgorithmen als ›moralischer Kompass‹ in ausweglosen Fahrsituationen

3.1 Ethische Problemstellungen und Diskurse im Überblick

3.1.1 Ethische Problemfelder im Kontext des autonomen Fahrens

Die Automatisierung des Verkehrs ist gleichbedeutend mit einer technologischen Revolution, die soziale Auswirkungen von großer Tragweite mit sich bringt. Wie Moor (2005, S. 117–118) betont, sind damit insbesondere ab Level 3 essenzielle ethische Herausforderungen verbunden. So wirft das autonome Fahren zum einen – wie alle Technologien, die unser tägliches Leben tiefgreifend transformieren – technikphilosophische und -anthropologische Fragen auf, die sich um das Verhältnis des Menschen zur Technologie und zu sich selbst drehen. Diese betreffen beispielsweise das (subjektive) Empfinden eines Verlustes an Autonomie durch das Delegieren von Fahraufgaben (vgl. Fossa, 2024), wodurch sich der zum Passagier gewordene Fahrer selbst neu definieren muss.

Zum anderen ergeben sich durch den angedachten Einsatz höherstufig automatisierter Fahrzeuge im praktischen lebensweltlichen Kontext diverse Problemstellungen aus dem Bereich der Angewandten Ethik. Mit diesen beschäftigte sich eine durch das BMVI im Herbst 2016 eingesetzte Ethik-Kommission, die unter der Leitung des ehemaligen Richters Udo Di Fabio aus Experten der Bereiche Verkehr, Rechtswissenschaften, Informatik, Ingenieurwissenschaften, Philosophie und Theologie sowie Vertretern von Verbraucherschutz, Verbänden und Unternehmen bestand. Der 2017 vorgestellte Abschlussbericht enthält die weltweit ersten Leitlinien für das autonome Fahren. Die darin formulierten Thesen und zentralen Prinzipien u. a. in Bezug auf ein Diskriminierungsverbot oder die Priorisierung des Schutzes des Lebens vor anderen Erwägungen und Verantwortungsfragen beziehen sich auf gegenwärtig noch nicht realisierte Systeme der Level 4 und 5. Sie sind daher als praxisori-

entierte Diskussionsgrundlage u. a. an gesetzgebende Institutionen gerichtet, auf deren Basis begleitend zur technischen Weiterentwicklung die Erörterung entsprechender ethischer Aspekte erfolgen und die gesellschaftliche Akzeptanz autonomer Fahrzeuge sichergestellt werden kann.

In der relevanten Forschungsliteratur wird eine ethische Perspektive auf das Phänomen des autonomen Fahrens bereits seit ca. zehn Jahren diskutiert und nimmt seitdem analog zur technischen Entwicklung entlang des Stufenmodells stetig an Komplexität zu. Im Fokus des ethischen Diskurses stehen im Wesentlichen die beiden letzten Level des Stufenmodells: das hoch- und das vollautomatisierte bzw. autonome Fahren. Als anvisiertes Ziel der technologischen Entwicklung ist auf diesen beiden Stufen der Anteil automatisierter Prozesse im Vergleich zu menschlichen Fahranteilen am größten, wobei streng genommen nur auf der letzten Stufe von autonomen Systemen gesprochen werden kann. Im Anschluss an den Forschungsdiskurs werden ethische Fragen im Rahmen dieser Untersuchung ebenfalls primär in Bezug auf Level-5-Systeme betrachtet.

Der beständig wachsende ethische Diskurs autonomer Fahrsysteme lässt sich grob anhand zweier übergeordneter Problemfelder untergliedern.[37] Dies sind zum einen breitere ethische Fragen, die sich aus einer Perspektive auf autonome Fahrzeuge als Teil eines sozio-technischen Zusammenhangs ergeben.[38] Hansson et al. (2021) stellen beispielsweise den durch die vielfältigen ethischen Herausforderungen herbeigeführten sozialen Wandel in den Vordergrund. Als wertgeladene Technologie sind selbstfahrende Fahrzeuge geeignet, wesentliche Aspekte des menschlichen Lebens sowohl im Hinblick

37 Ergänzend zu den im Folgenden erläuterten ethischen Problemstellungen wird das automatisierte Fahren zudem häufig als Anwendungsbeispiel für ethische Fragen im Zusammenhang mit autonomen Systemen im Allgemeinen herangezogen. Vor allem in der Roboter- und Maschinenethik sowie für ingenieurwissenschaftliche Forschungsfragen ist es ein beliebter *Use Case*, anhand dessen spezifische Fragestellungen aus der Perspektive der jeweiligen Disziplinen erörtert werden. Für eine unlängst publizierte Übersicht zu ethischen und rechtlichen Herausforderungen im Kontext des autonomen Fahrens siehe Nyholm (2023a).

38 Ergänzend sei hier auf diejenigen Herausforderungen verwiesen, die sich im Kontext einer ethischen Perspektive auf Algorithmen im Allgemeinen ergeben. Für eine entsprechende Einführung in die Thematik siehe z. B. Mittelstadt et al. (2016), Tsamados et al. (2022) und Zweig (2019).

auf die individuelle Lebensgestaltung als auch auf gesellschaftliche Beziehungen zu konditionieren. Diese Sichtweise fußt auf der technikethischen Prämisse, dass Technologien an sich nicht wertneutral sind;[39] sie müssen als »inhärent moralisch vorprogrammiert verstanden [werden], insofern sie bestimmte moralische Werte und Normen fördern oder behindern.« (Simon, 2016, S. 359) So forcieren selbstfahrende Fahrzeuge den seit einigen Jahren konstatierten Wandel hin zu einer Infrastruktur nachhaltiger, ressourcenschonender *Shared Mobility*, die dem Einzelnen eine effizientere Nutzung von Wegezeiten ermöglicht. »Technik ist immer in gesellschaftliche Zielsetzungen, Problemdiagnosen und Handlungsstrategien eingebettet. In ihr verfestigen sich Wertvorstellungen durch Zielvorgaben und Designentscheidungen«, erläutert Grunwald (2016, S. 28).

In diesem Sinne widmet sich ein Teil der ethischen Literatur denjenigen Schwierigkeiten, die durch bereits im Design der verwendeten Technologie transportierte Werte hervorgerufen werden. So wird thematisiert, auf welche Weise durch vernetzte Infrastruktur ermöglichte Mechanismen der Verkehrssteuerung die Effekte sozialer (Un-)Gerechtigkeit verstärken können. Ein mögliches Beispiel dafür, wie die berechtigten Bedürfnisse und Interessen Einzelner tangiert würden, wäre ein Mechanismus, durch den eine Notfallfahrt zum Krankenhaus durch Verkehrsleittechnik – z. B. Ampelschaltung – beschleunigt werden könnte (vgl. Mladenovic & McPherson, 2016, S. 1132–1137). Als weitere ethisch relevante Themen ergänzen Hansson et al. (2021, S. 1396–1399) noch die zu erwartenden Auswirkungen auf Gesundheit und Umwelt sowie den Arbeitsmarkt, die im

39 Der Diskurs einer vermeintlichen Wertneutralität der Technik verfügt über eine jahrzehntelange Tradition in der Technikforschung sowie Technik- und Wissenschaftsphilosophie. Bis in die 1990er-Jahre war hier die These dominant, dass Technik prinzipiell wertneutral sei, was häufig mit dem instrumentellen Charakter technologischer Artefakte begründet wird. Gemäß dieser Sichtweise werden Technologien lediglich als neutrale Werkzeuge betrachtet, die erst durch menschliche Absichten und Nutzung ethische Relevanz entfalten (vgl. Hubig, 1993). Im Zuge der technologischen Weiterentwicklung wurden moderne Technologien jedoch verstärkt zu komplexen autonomen Systemen, die über eine reine Mittelfunktion hinausgehen. Dabei reifte auch die Einsicht, dass moralisch relevante Konsequenzen nicht erst durch den Gebrauch entstehen können, sondern bereits im Design von Technologien explizit und implizit enthalten sind (vgl. Brey, 2010, S. 43–49).

Forschungsdiskurs bisher allerdings nur eine marginale Rolle einnehmen.

Intensiv erforscht ist dagegen die Thematik von Datenschutz und Privatsphäre, wobei sich ethische und rechtliche Aspekte (als institutionalisierte Moral) nicht immer trennen lassen (vgl. Boeglin, 2015). Grundlegend ist hier die Feststellung, dass Komfortzuwachs einen Preis hat: Roboterfahrzeuge nehmen uns nicht nur Fahraufgaben ab, sondern auch die Fähigkeit zu freien (Fahr-)Entscheidungen; sie kontrollieren in gewisser Weise unser Mobilitätsverhalten und beeinträchtigen dadurch die individuelle Autonomie. Ethische Fragestellungen ergeben sich hier vor allem aus dem Spannungsverhältnis von Freiheit, Autonomie sowie Privatsphäre einerseits und der praktischen Relevanz generierter Daten, z. B. für Haftungsfragen (vgl. Boeglin, 2015; Rannenberg, 2015) oder als Systemtrainingsdaten, andererseits. Dominiert wird die Debatte von Datenschutzproblematiken, die auftreten, wenn persönliche Daten der Passagiere erzeugt und verarbeitet werden, etwa über den aktuellen Standort, Fahrtziel oder Bewegungsmuster. Vordergründige Bedenken drehen sich zum einen um das Risiko einer Überwachung und externen Kontrolle sowie die Weitergabe sensibler Daten an Unbefugte (vgl. Glancy, 2012, S. 1188–1216; Lim & Taeihagh, 2018, S. 6–14).[40] Bei Daten, die den Standort von Personen enthalten, handelt es sich um sensible Informationen, die Beziehungen oder religiöse und politische Zugehörigkeiten offenbaren können; zudem besteht die Gefahr, dass sie für kommerzielle Zwecke missbraucht werden (vgl. Hansson et al., 2021, S. 1395–1396). LaFrance (2016, o. S.) schreibt dazu: »In this near-future filled with self-driving cars, the price of convenience is surveillance.«

Der zweite und größere Teil des ethischen Diskurses beschäftigt sich mit Fragestellungen, die sich um den Sicherheitsaspekt selbstfahrender Fahrzeuge drehen. Besonderes Augenmerk liegt hierbei auf ethischen Fragen rund um Unfallsituationen mit Beteiligung höherstufig automatisierter Fahrzeuge. Die Zahl relevanter, in Fachjournals publizierter ethischer Untersuchungen wächst stetig und hat die Thematik inzwischen als dominanten ethischen Diskurs

40 Die Expertengruppe »Driverless Mobility« der Europäischen Kommission betont in diesem Kontext, dass neue Strategien, Forschung und Industriepraktiken erforderlich sind, um Datenschutz und Privatsphäre weiterhin zu gewährleisten (vgl. Europäische Kommission, 2020, S. 34–51; Santoni de Sio, 2021, S. 721–722).

rund um das autonome Fahren etabliert, der primär über zwei Perspektiven erschlossen wird. Während sich einige Artikel der Problematik über Fragen der Verantwortungszuschreibung nähern, fokussieren sich andere auf ethische Fragestellungen im Zusammenhang mit der Fahrzeugsteuerung in Situationen, in denen sich Schäden nicht vermeiden lassen. In den folgenden beiden Unterkapiteln werden diese Diskurse jeweils grob skizziert.

3.1.2 Problemfeld Unfallsituationen: Der Verantwortungsdiskurs

Wer trägt die Verantwortung für entstehende Schäden im Kontext autonomer und vernetzter Fahrzeuge? Fragen nach Verantwortung und Haftbarkeit weisen hier Überschneidungen auf, wobei letztere Teil des rechtswissenschaftlichen Diskurses sind. Als Synthese juristischer und philosophischer Literatur sind verschiedene anwendungsbezogene Entwürfe möglicher Verantwortungszuschreibung erarbeitet worden. Ein hilfreicher Überblick über die einschlägige Literatur aus dem Bereich der Rechtswissenschaften findet sich bei Nyholm (2018c, S. 2–3). Er konstatiert, dass dem rechtlichen Diskurs vor allem zwei zentrale Erkenntnisse zu verdanken sind, die auch für philosophische Untersuchungen fruchtbar gemacht werden können: Grundlegend für jegliche Betrachtung von Verantwortung im Kontext von Fahrrobotern ist zum einen die Feststellung einer ›existenziellen Krise‹, welche aus der Notwendigkeit resultiert, die Rolle bisheriger menschlicher Fahrer, ihr Verhältnis zum autonomen System und damit auch ihre Verantwortung von Grund auf neu zu deuten.[41] Zum anderen erweitert die rechtswissenschaftliche Literatur den Blickwinkel auf alternative, in der Praxis übliche Modelle der Verantwortung. So assoziieren wir Verantwortung nicht nur mit den Folgen bestimmter Handlungen, sondern schreiben diese auch aufgrund bestimmter (sozialer) Rollen oder zugestandener Rechte zu. Für Verantwortungsfragen rund um autonome Systeme bzw. solche, die eine Kollaboration von Mensch und Maschine erfordern, sind diese Aspekte bis dato weitgehend unberücksichtigt geblieben.

41 Eine einschlägige empirische Untersuchung der psychologischen Aspekte verschiedener Grade geteilter Verantwortung zwischen Nutzern und Herstellern legen Liu et al. (2021) vor.

Aus Sicht der Maschinenethik hängt die Frage, inwiefern autonome Systeme für die Konsequenzen ihres Handelns verantwortlich sind, unmittelbar damit zusammen, ob Maschinen Subjekte moralischen Handelns sein können. Die gegenwärtig dominierende Position maschinenethischer Forschung ist es, dass Maschinen essenzielle Merkmale einer moralischen Handlungsfähigkeit nicht oder nicht in ausreichendem Maße erfüllen, um als Träger moralischer Verantwortung gelten zu können.[42] Daher übt sich der Diskurs weitgehend in Zurückhaltung, wenn es darum geht, künstlichen Systemen eine Verantwortungsfähigkeit zuzuschreiben. So vertritt Sparrow (2007, S. 71–73) die Auffassung, dass eine Maschine nicht in dem Sinne zur Verantwortung gezogen werden kann, dass sie eine daraus folgende Bestrafung tatsächlich als solche empfindet.

Vor diesem Hintergrund dreht sich die einschlägige philosophische Literatur hauptsächlich um die Problematik von Verantwortungslücken (*responsibility gaps*). Diese sind darauf zurückzuführen, dass das Verhalten von Robotern weder für Entwickler noch für Nutzer vollständig vorhersehbar oder kontrollierbar ist (vgl. ebd., S. 70–71). Eine Anwendung traditioneller Modelle der Verantwortungszuschreibung lässt sich hier nur schwer legitimieren:[43]

> Traditionally we hold either the operator/manufacturer of the machine responsible for the consequences of its operation, or ›nobody‹ (in cases, where no personal fault can be identified). Now it can be shown that there is an increasing class of machine actions, where the traditional ways of responsibility ascription are not compatible with our sense of justice and the moral framework of society because nobody has enough control over the machine's actions to be able to assume the responsibility for them. These cases constitute what we will call the responsibility gap. (Matthias, 2004, S. 177)

42 Eine differenzierte Erörterung der Grenzen moralischer Handlungs- und Verantwortungsfähigkeit autonomer Systeme ist an anderer Stelle bereits von der Autorin publiziert worden (vgl. Schäffner, 2022).

43 Nyholm (2018a, S. 1206) weist im Kontext der Begründung einer möglichen Verantwortungslücke darauf hin, dass die bloße Unvorhersehbarkeit und die Unfähigkeit, eine Technologie vollständig zu kontrollieren, menschliche Subjekte noch nicht von jeglicher Verantwortung für deren Handlungen freisprechen. Dies kann nur unter der Voraussetzung erfolgen, dass die Unkontrollierbarkeit darauf zurückzuführen ist, dass das betreffende künstliche System in nicht-trivialer Weise über Autonomie verfügt.

Eine Alternative zur klassischen Verantwortungslücke entwirft Danaher (2016) mit der sogenannten Vergeltungslücke (*retribution gap*), die seit einigen Jahren die Debatte speziell im Hinblick auf künstliche Systeme erweitert. Sie basiert auf der Diskrepanz zwischen dem menschlichen Wunsch, bei schuldhaftem Verhalten Vergeltung zu erwirken, und dem Fehlen eines entsprechend in die Pflicht zu nehmenden Subjekts.

Obwohl die Thematik der Verantwortungslücke im Kontext selbstfahrender Fahrzeuge erst in den letzten Jahren verstärkt in den Vordergrund getreten ist, existiert bereits ein breit gefächertes Diskussionsspektrum. Dies ist der Tatsache zu verdanken, dass analoge Diskurse schon länger in anderen Einsatzbereichen autonomer Systeme geführt werden, so im Rahmen der Debatte über die moralische Zulässigkeit letaler autonomer Waffensysteme[44] (vgl. Misselhorn, 2018b, S. 155–184), zu dem sich einige Anknüpfungspunkte finden lassen (vgl. Jong, 2020; Nyholm, 2018a). Übereinstimmende Erkenntnis des gegenwärtigen Forschungsstands ist es, dass sich individuelle Modelle der Verantwortung nicht ohne Weiteres auf autonome Systeme übertragen lassen. Santoni de Sio und Mecacci

44 Kriegsroboter stellen gewissermaßen einen ethischen Sonderfall dar: Im Kontext von Kriegshandlungen findet die Theorie des gerechten Krieges als bereichsspezifische normative Theorie Anwendung, woraus sich Rechtsordnungen wie das humanitäre Völkerrecht ableiten. Dennoch gibt es einige Gemeinsamkeiten zum autonomen Fahren. So behauptet der prominente Ingenieur Ronald Arkin, dass autonome Systeme Kriege ethischer und humaner machen würden, weil sie Regeln des Völkerrechts besser einhalten könnten als Menschen, welchen sie hinsichtlich sensorischer Situationseinschätzung und Ausführungspräzision überlegen sind (vgl. Arkin, 2010, S. 332–334, 2018, S. 318–319). Dieses Argument erinnert an die durch automatisierte Fahrzeuge angestrebte Kompensation menschlicher Fehleranfälligkeit bei der Fahrzeugführung und das daraus resultierende Sicherheitsversprechen. Eines der zentralen Argumente der Gegenposition betrifft die Problematik der Verantwortungslücke (vgl. Sparrow, 2007, S. 67–68): Nur wenn es jemanden gibt, der die moralische Verantwortung für ausgeführte militärische Aktionen trägt, können solche Systeme grundsätzlich erlaubt werden. Zudem sind autonome Waffensysteme prinzipiell bedenklich in Bezug auf den Verlust menschlicher Kontrolle über den Einsatz von Gewalt, insbesondere Massenvernichtungswaffen, die ethische Grundpfeiler wie das Völkerrecht und die Menschenrechte zu untergraben drohen. Menschliche Urteilskraft und Handlungsvermögen sind bis auf Weiteres unverzichtbar, um den Anforderungen des *ius in bello* gerecht zu werden, welches die Art und Weise ethisch zulässiger Kriegsführung regelt (vgl. Koch & Rinke, 2018, S. 127–130; Sparrow, 2016, S. 99).

(2021, S. 1060–1068) beispielsweise haben eine breite Analyse der Verantwortungslücke vorgelegt, in der sie diese nicht als ein einziges Problem, sondern als eine Reihe von mindestens vier miteinander verbundenen Problemen auffassen. Diese manifestieren sich im Hinblick auf defizitäre Konzepte von Schuldfähigkeit, moralischer und öffentlicher Rechenschaftspflicht sowie aktiver Verantwortung und werden jeweils auf unterschiedliche Weisen verursacht. In einer der jüngsten einschlägigen Publikationen identifiziert Nyholm (2023b) verschiedene vorwärts- und rückwärtsgerichtete Verantwortungslücken im Zusammenhang mit autonomen Fahrzeugen und bewertet mögliche Strategien, wie diese Lücken geschlossen werden können.

Neben der Thematisierung von Verantwortungslücken sehen sich relevante Ansätze auch mit dem sogenannten *Problem of Many Hands*[45] konfrontiert (vgl. Europäische Kommission, 2020, S. 58–63; Santoni de Sio, 2021, S. 723). Dafür ist es notwendig, auf breitere und innovative Konzepte der Verantwortung zurückzugreifen:

> It is important for all stakeholders to move beyond a narrow conception of responsibility for CAVs as involving purely backward-looking responsibility (legal liability or culpability) for accidents and mistakes, towards a broader, forward-looking conception of responsibility as a culture that sustains and shapes the development, introduction, and use of CAVs in a way that promotes societal values and human well-being.[46] (Europäische Kommission, 2020, S. 53)

Eine der ersten philosophischen Untersuchungen zur Verantwortungsfrage im Kontext selbstfahrender Fahrzeuge liefern Hevelke und Nida-Rümelin (2015b).[47] Sie argumentieren zunächst, dass es aus pragmatischen Gründen nicht zielführend sei, Hersteller in die Verantwortung zu nehmen, denn diesen würden durch eine hohe Verantwortungslast Anreize genommen, in die Entwicklung autono-

45 Das *Problem of Many Hands* ist ein Phänomen, welches oft im Kontext von Verantwortungsfragen auftritt und die Schwierigkeit der Zuschreibung individueller Verantwortung in Zusammenhängen kollektiven Handelns aufgreift (vgl. Thompson, 1980; van de Poel et al., 2015). Im Zuge der Entwicklung autonomer Systeme wird es zunehmend auch im Kontext von Technologien thematisiert.

46 Die Abkürzung CAV steht für ›Connected and Automated Vehicle‹.

47 Eine andere Perspektive auf das Thema der Verantwortung rund um autonome Fahrsysteme nimmt Kauppinen (2021) ein. Anstatt Verantwortungsträger zu bestimmen, stellt er die Rolle in den Mittelpunkt, die Verantwortlichkeitsüberlegungen an sich für die Bestimmung der ›richtigen‹ Handlung spielen.

mer Fahrzeuge zu investieren. Stattdessen sprechen sie sich für ein Modell kollektiver Verantwortung aus, welches alle Nutzer autonomer Fahrzeuge als Teil einer risikokreierenden Gemeinschaft einbezieht. Die Operationalisierung dieser gemeinsamen Verantwortung wäre über eine Pflichtversicherung oder Besteuerung denkbar (vgl. ebd., S. 623–628).

Weitere Forschungsbeiträge greifen die Problemstellung auf, dass weder Hersteller noch Nutzer noch Maschine zweifelsfrei als alleinige Verantwortungsträger gelten können. So setzt sich Coeckelbergh (2016) mit den epistemischen und sozial-relationalen Problemen auseinander, welche die Ausübung und Zurechnung von Verantwortung im Kontext von selbstfahrenden Autos erschweren. Borenstein et al. (2017) beleuchten die Verantwortung der Entwicklungsingenieure und betonen die Bedeutung einer Werteorientierung bereits im Designprozess. Liu (2017) analysiert zunächst die Konzepte, welche die Idee der Verantwortung formen, und erörtert sodann, inwiefern Ansätze der Zielorientierung (*targeting*) einerseits und distributiver Fragen andererseits zur Klärung der Verantwortungsfrage beitragen können. Awad et al. (2019) belegen anhand einer empirischen Untersuchung, dass in der öffentlichen Wahrnehmung möglicherweise eine zu geringe Sensibilisierung für die Fehlerhaftigkeit von Technologien besteht. Diese äußert sich dahingehend, dass Maschinen bei folgenreichen Fahrfehlern in der öffentlichen Wahrnehmung tendenziell für weniger schuldig befunden werden als menschliche Fahrer.

Eine der größten Schwierigkeiten bei dem Versuch, autonome Technologien in die Verantwortung zu nehmen, stellt die Frage dar, inwiefern die Aktionen von Maschinen tatsächlich als unabhängig von menschlichem Eingreifen zu sehen sind. Davon ausgehend manifestiert sich in den letzten Jahren eine Tendenz in der Forschungsliteratur, künstliche Systeme nicht als vollständig autonome, sondern als Systeme kollaborativen Handelns zu interpretieren. Diese Perspektive hat direkte Auswirkungen auf die Frage der Verantwortlichkeit: »Auch wenn Maschinen nicht moralisch verantwortlich sein können, haben sie doch Auswirkungen auf die Zuschreibung von Verantwortung.« (Misselhorn, 2018b, S. 126) In diesem Sinne wird oft vorgeschlagen, Maschinen im Rahmen hierarchischer Modelle kollaborativer Verantwortung entsprechend ihres moralischen Status eine partielle Verantwortung zuzusprechen. Misselhorn (2015b) plä-

diert beispielsweise für eine Auffassung von Robotern als kooperative Akteure,[48] während Nyholm (2018c, S. 5) die Beziehung zwischen Fahrer und Fahrzeug als eine Partnerschaft gemeinsamen Handelns charakterisiert:

> […] if we do attribute agency to them, we should think of this as a form of collaborative agency, where the key partners in these human–robot collaborations are certain humans. After all, humans set self-driving cars' goals (e.g., going to the grocery store).

In diesem Sinne propagieren auch Loh und Loh (2017) bei hybriden Fahrzeugen eine geteilte Verantwortlichkeit zwischen Fahrer und System. Loh und Misselhorn (2019) argumentieren, dass Nutzer, Hersteller und autonome Systeme ein Netzwerk der Verantwortung bilden, das die Verantwortung im Hinblick auf ein gemeinsames Ziel teilt, welches in der Verwirklichung maximaler Verkehrssicherheit besteht. Wie Neuhäuser (2015, S. 135) demonstriert, müssen die innerhalb eines Verantwortungsnetzwerks durch unvorhergesehenes Verhalten von Maschinen verursachten Lücken in der Verantwortlichkeit durch menschliche Verantwortungsträger geschlossen werden:

> The ideal of an extensive network of responsibility states that for all matters that are important to people, someone should be responsible, or at least it should be possible to hold someone accountable. The more actions irresponsible robots undertake and the less predictable their actions become, the stronger their potentially negative influence within this extensive network of responsibility will be. […] The unpredictable nature of their actions allows for gaps to emerge within the extensive network of responsibility. People would then have to take responsibility not only for themselves, animals and nature, but also for robots and their doings in order to fill these gaps.

Gunkel (2020) schließlich diskutiert diverse Ansätze auf einem Spektrum, das von voller menschlicher Verantwortlichkeit über einen hybriden Ansatz geteilter Verantwortung zwischen menschlichen und technischen Komponenten bis hin zu einer teilweisen, funktionalen Verantwortungszuschreibung an Maschinen reicht.

48 In einem umfangreichen Sammelband hat Misselhorn (2015a) Ansätze zusammengetragen, die philosophische Konzepte zu Kooperation und kollektivem Handeln mit ingenieurwissenschaftlicher Forschung zu Multi-Agenten-Systemen zusammenbringen.

3.1.3 Praktische Unvermeidbarkeit und dilemmatische Struktur auswegloser Fahrsituationen

Jenseits von Fragen der Verantwortbarkeit fokussiert sich der größte Teil der Forschungsliteratur zur Ethik des autonomen Fahrens auf Fragestellungen, die in Zusammenhang mit unvermeidbaren Unfallsituationen stehen. Auch wenn selbstfahrende Fahrzeuge auf maximal defensives und vorausschauendes Fahrverhalten programmiert werden, lassen sich Unfälle nicht grundsätzlich ausschließen. Wie soll ein autonomes Fahrzeug in derartigen Situationen agieren? Soll es lediglich bremsen oder zusätzlich noch ausweichen? Und wenn ja, wohin?

Die zunehmende Durchdringung unserer Lebenswelt mit autonomen Technologien impliziert, dass diese sich mit Situationen konfrontiert sehen, in denen sie vor moralische Entscheidungen gestellt werden. Der nun mehr als ein Jahrzehnt andauernde, lebhafte Diskurs dreht sich im Kern um die Frage, auf welche ethischen Werte bzw. Normen die Entscheidungsalgorithmen autonomer Fahrzeugsysteme in solchen Notsituationen zurückgreifen sollen. In diesem Zusammenhang ergeben sich zahlreiche Probleme von ethischer Relevanz, z. B.: Wie soll mit Zielkonflikten zwischen Sicherheit und Komfort umgegangen werden?[49] Wie sollen Unfallalgorithmen für autonome Fahrzeuge im Hinblick auf mögliche Schadensfälle programmiert[50] werden? Sollen sie stets die Sicherheit ihrer Insassen priorisieren, dem Prinzip der Schadensminimierung folgen oder einem anderen ethischen Prinzip?

Der einschlägige Forschungsdiskurs setzt sich mit diesen und ähnlichen Fragestellungen unter dem Schlüsselbegriff der ›Unfallalgorithmen‹ auseinander; im englischen Sprachraum wird häufig auch von ›ethics of crashing‹, ›crash optimisation‹ oder ›moral design problem‹ gesprochen. Erste einschlägige Publikationen in philosophischen Fachzeitschriften waren im Jahr 2014 zu verzeichnen; in den Rechtswissenschaften hatte die Debatte um ethisch relevante

49 Ein solcher Zielkonflikt tritt beispielweise im Kontext des ›Paradoxons der Automatisierung‹ auf, welches in Kap. 2.2.3 beschrieben wurde.

50 Wird im Kontext von Unfallalgorithmen von Programmierung gesprochen, so sind damit keine *Hardcoding*-Praktiken gemeint, sondern ein softwaretechnischer, KI-basierter Designansatz für autonome Systeme auf der Basis von *Deep-Learning*-Techniken (siehe auch Kap. 4.1.2).

Rechtsfragen im Kontext selbstfahrender Fahrzeuge bereits einige Jahre früher begonnen (vgl. Nyholm, 2018b, S. 2). Anfangs wurde der ethische Diskurs vor allem durch populärwissenschaftliche und akademische Veröffentlichungen sowohl des Ingenieurwissenschaftlers Noah J. Goodall[51] als auch des Philosophen Patrick Lin[52] forciert und geprägt.[53] Ihre viel zitierten Beiträge (vgl. Goodall, 2014a, 2016a, 2016b, 2017; Lin, 2013a, 2013b, 2014a, 2014b, 2015) zählen bis heute zur Standardliteratur zum Thema der ethischen Problematisierung von Unfallsituationen mit Beteiligung automatisierter Fahrzeuge.

Auf dieser Grundlage wurde die Thematik in der Folge von führenden Experten aus Politik, Wirtschaft und Wissenschaft in interdisziplinären Sammelbänden (vgl. Lin et al., 2017; Maurer et al., 2015) und international renommierten Buchreihen wie den *Lecture Notes in Mobility*, herausgegeben von Gereon Meyer und Sven Beiker, aufgegriffen. Ausgelöst durch diverse Unfälle mit automatisierten (Test-)Fahrzeugen stieg allmählich die mediale Aufmerksamkeit für ethische Fragen, wodurch die Thematik in den letzten Jahren noch stärker in den Fokus akademischer Publikationen rückte. Während die frühen Artikel primär dem Ziel dienten, eine Debatte über ethische Probleme im Zusammenhang mit Unfallsituationen zu entfachen, kam es mit der Zeit zu einer Ausdifferenzierung der Fragestellungen, die sich in der zunehmenden Interdisziplinarität der relevanten Publikationslandschaft widerspiegelt. So wurden vielschichtige Teilaspekte wie Sicherheits- und Gerechtigkeitsfragen, Verantwortung oder politische Aspekte fortan von interdisziplinären Autorenteams aus Philosophie sowie Rechts- und Ingenieurwissenschaften bearbeitet. Zudem integrieren jüngere Forschungsbeiträge

51 Der promovierte Bauingenieur Noah J. Goodall ist Senior Research Scientist des Virginia Transportation Research Council; seine Publikationen zeichnen sich durch eine inter- und transdisziplinäre Denkweise aus, die immer wieder auch ethische Fragen thematisiert.

52 Patrick Lin ist Direktor der »Ethics + Emerging Sciences Group« an der California Polytechnic State University. Aufgrund seiner vielfältigen Affiliationen und seiner Expertise in technikethischen Fragen ist er nicht nur einer der führenden publizierenden Forscher in diesem Bereich, sondern auch ein gefragter Ansprechpartner für internationale Medien.

53 Zum erweiterten Kreis derjenigen Forscher, die sich als erste mit ethischen Fragen im Kontext von Unfallsituationen beschäftigten, gehört auch der Technikethiker Jason Millar (2014a, 2014c, 2015).

verstärkt Verantwortungs- und Designperspektiven in die Problemstellung und setzen diese zueinander in Beziehung.[54]

Die Problematik des Designs von Unfallalgorithmen steht im Kontext des (vermeintlichen) Sicherheitsversprechens, das den zentralen Legitimationsgrund für die Einführung des höherstufig automatisierten Fahrens bildet. Das oft gepriesene Sicherheitspotenzial besteht nun gerade in der Erwartung, dass zuvor durch menschliches Versagen verursachte Unfälle fortan durch die defensive, vorausschauende und stets regelkonforme Fahrweise selbstfahrender Fahrzeuge vermieden werden können. Auch wenn dies für einen Teil der relevanten Unfallsituationen zutreffen mag, so sprechen jedoch plausible Argumente dafür, dass der Anspruch eines Zustands völliger Unfallfreiheit, die sogenannte *Vision Zero*[55], eine Utopie darstellt. Dabei sind laut Fossa (2023, S. 65) verschiedene Aspekte relevant:

> Technical failures, infrastructural problems, and human misconduct will always pose safety threats. As a matter of fact, accidents can occur even when everything runs as it should. Driving is an utterly complicated phenomenon fraught with uncertainty, unpredictability, and risk. Unfortunate situations in which all possible courses of action would lead to an incident simply cannot be theoretically excluded. Some collisions are just unavoidable.

Die innovative Fahrzeugtechnologie kreiert ihrerseits neue Risiken, die in einer Welt ohne autonome Fahrzeuge nicht auftreten würden, denn technische Systeme sind niemals völlig zuverlässig. Durch Sicherheitskonzepte lassen sich zwar Auftrittswahrscheinlichkeit und Schadensausmaß von Systemfehlern, -störungen und -ausfällen v. a. durch redundante Implementierung sicherheitskritischer Komponenten minimieren, jedoch verbleibt stets ein gewisses Restrisiko für ein technisches Versagen. Wachenfeld und Winner (2015, S. 473–474) stufen das vollautomatisierte Fahren als »nicht überwachte Au-

54 Für einen Überblick über bis dato verwendete ethische Konzepte in Bezug auf die Verhaltenssteuerung autonomer Fahrzeuge siehe Németh (2023).

55 Der Begriff der *Vision Zero* wurde erstmals 1995 im Kontext eines schwedischen Programms zur Steigerung der Verkehrssicherheit erwähnt. Inzwischen ist es politisch erklärtes Ziel sowohl der deutschen Bundesregierung als auch auf EU-Ebene, die Zahl der Verkehrstoten mittelfristig auf nahezu null zu senken. Autonome Fahrsysteme stellen dabei ein zentrales strategisches Element dar, um sich der Vision anzunähern, auch wenn sich diese realistischerweise niemals vollständig erreichen lassen wird (vgl. Köllner, 2018; Schäfer, 2018).

tomation« bzw. »Automation ohne Korrekturmöglichkeit« ein, die sich dadurch auszeichnet, dass Systemfehler unmittelbar zu einer Gefährdung von Personen und Umwelt führen. Vallor und Bekey (2017, S. 343) argumentieren, dass dies insbesondere im Fall selbstlernender Systeme gilt, da deren Verhalten sich nur in begrenztem Maße kontrollieren und vorhersehen lässt:[56]

> Statistically they may be competitive with or even superior to humans at a given task, but unforeseen outputs […] are a rare, but virtually ineradicable possibility. Some are emergent behaviors produced by interactions in large complex systems. Others are simple failures of an otherwise reliable system to model the desired output.

Ferner stützen sich die Erwartungen an das Sicherheitspotenzial autonomer Fahrsysteme auf die Annahme, dass die Beachtung der Verkehrsregeln den wichtigsten Faktor eines sicheren Verkehrsgeschehens ausmacht. Im Gegensatz zu Menschen, die spezifische Verkehrslagen situativ abschätzen können, befolgen künstliche Systeme vorgegebene Regeln rigoros. Nun kann jedoch in bestimmten Fällen gerade ein Festhalten an Verkehrsregeln zu Unfällen oder zumindest einer signifikanten Erhöhung des Unfallrisikos führen, wohingegen sich dies durch ein kontrolliertes Abweichen von den Regeln vermeiden ließe.[57] So kann es sinnvoll sein, kurzzeitig die maximal erlaubte Geschwindigkeit zu überschreiten, z. B. zur Prävention von Kollisionen bei Überholmanövern oder von Auffahrunfällen im gebundenen Verkehr (vgl. Reed et al., 2021, S. 781–782).[58]

Weiterhin sind gewisse Gefahrenpotenziale des Straßenverkehrs aufgrund ihrer Komplexität nicht gänzlich eliminierbar (vgl. Gasser,

---

56 In Abhängigkeit von der Phase des Systemlebenszyklus (Forschung, Entwicklung, Betrieb, Service und Nutzerwechsel/Stilllegung), in dem Techniken maschinellen Lernens zum Einsatz kommen, ergeben sich unterschiedliche Herausforderungen und Lösungsstrategien. Intensiver Forschungsbedarf besteht u. a. im Bereich der Laufzeitverifikation und -validierung (vgl. Wachenfeld & Winner, 2015, S. 474–478).

57 Ein gezieltes Übertreten von Verkehrsregeln in Notsituationen wäre freilich nur unter der Voraussetzung zu rechtfertigen, dass autonome Fahrsysteme zweifelsfrei feststellen können, wann eine solche Situation vorliegt (vgl. Reed et al., 2021, S. 783).

58 Eine von Goodall (2021) vorgelegte Studie zeigt, dass die Unfallrate autonomer Fahrsysteme bei Auffahrunfällen 4,8 Mal höher ist als bei von Menschen gesteuerten Fahrzeugen, was vor allem auf plötzliches und unerwartetes Anhalten zurückzuführen ist.

2015, S. 555). Färber (2015, S. 128) beschreibt den Straßenverkehr als »ein selbstorganisiertes, chaotisches System […], das zwar prinzipiell durch Regeln geordnet wird, bei dem aber viele Situationen nicht in einer eindeutigen Regel festgelegt werden können.« Technologien der Fahrzeug-zu-Fahrzeug-Kommunikation bergen zweifellos großes Potenzial, kritische Situationen durch das gemeinsame Finden kooperativer Lösungen bereits in der Entstehung zu verhindern; jedoch lassen sich dadurch nicht alle denkbaren Unfallsituationen vermeiden, insbesondere nicht solche, die Verkehrsteilnehmer außerhalb des Kommunikationsnetzes betreffen (vgl. Reschka, 2015, S. 508–509). Als Folge verbleiben Situationen, die sich durch vorausschauendes Fahren nicht vereiteln lassen oder auf Ursachen zurückzuführen sind, die durch Automatisierung nicht kompensiert werden können.

Der Antizipation von Degradationssituationen[59] kommt eine zentrale Bedeutung zu, wenn es darum geht, vorausschauend zu fahren und dadurch möglichem Schaden vorzubeugen (vgl. ebd., S. 506–507). Allerdings sind dem Antizipationspotenzial autonomer Systeme einerseits durch die Begrenztheit kognitiver Ressourcen (Rechenkraft, Sensorleistung, Prädiktionspräzision) und andererseits durch die Komplexität eines realen Verkehrsumfelds Grenzen gesetzt (vgl. Köllner, 2017). Während Ersteres zu situativ bedingten Fehleinschätzungen durch das System führen kann, sind vor allem das unerwartete Verhalten anderer Verkehrsteilnehmer und plötzlich auftretende Ereignisse oder eine Verkettung davon ursächlich dafür, dass gewisse Situationen schwerlich vorauszusehen sind und daher vom System nicht korrekt eingeschätzt werden können. Das gilt

59 Das Prinzip der funktionalen Degradation impliziert, dass der Funktionsumfang eines Systems bei Auftreten sicherheitskritischer Situationen herabgesetzt wird. Reschka (2015, S. 505–506) erläutert: »Treten Fehler in einem System auf oder sind die Ressourcen eingeschränkt, so werden die ›lebenswichtigen‹ Prozesse erhalten und weniger wichtige Prozesse reduziert oder beendet. Beispielsweise kann bei einem eingeschränkten Sichtfeld die Geschwindigkeit des Fahrzeugs reduziert werden. Unter bestimmten Bedingungen führen jedoch auch diese Aktionen nicht zu einer Reduzierung des Risikos auf einen zumutbaren Wert, sodass ein Anhalten des Fahrzeugs […] oder, falls dies ebenso zu riskant ist, ein Verlassen des Straßenverkehrs notwendig werden.«

insbesondere für die dynamischen Elemente einer Szene;[60] diese sind in zeitlicher oder räumlicher Dimension variabel, d. h. sie ändern ihre Zustände laufend (vgl. Geyer et al., 2014, S. 185–186). Während Lichtsignalanlagen oder Licht- und Wetterbedingungen noch in gewissem Maße vorhersehbar sind, ist das Verhalten anderer Verkehrsteilnehmer prinzipiell unberechenbar und stellt sich als ein wesentlicher Unsicherheitsfaktor bei dem Versuch heraus, eine Situation zu antizipieren.

Unzureichende technische Reife von Sensorik und Perzeptionsmechanismen dürfen jedoch kein Grund sein, ethische Forderungen zu ignorieren; schließlich kann auch eine ausgereifte Technik und vollautomatisierte Fahrzeugsteuerung Unfälle in komplexen Verkehrssituationen nicht vollständig verhindern. Dies ist in erster Linie dann der Fall, wenn die Reaktion des Fahrzeugs zeitkritisch ist oder fahrphysikalische Grenzen erreicht werden, beispielsweise wenn sich Kollisionsobjekte innerhalb der Bremsdistanz des Fahrzeugs befinden. Als klassisches Szenario dient hier das zwischen parkenden Autos plötzlich auf die Straße laufende Kind, dessen Aktionen von den Wahrnehmungssystemen des autonomen Fahrzeugs nicht korrekt oder zu spät gedeutet werden. In derartigen Fällen ist die Situationsprädiktion stark erschwert, sodass das autonome System diejenige Trajektorie nicht zuverlässig ermitteln kann, durch die sich eine Kollision noch vermeiden ließe. Als potenzielle Gefahrenzonen kommen hier vor allem unübersichtliche Verkehrsknotenpunkte sowie Sichtbehinderungen, beispielsweise durch Bäume, Hecken, andere Objekte des Verkehrsgeschehens, Gebäude oder Baustellenaufbauten, in Frage (vgl. Winkle, 2015, S. 372–374). Nach gegenwärtigem technischem Stand wechselt ein selbstfahrendes Fahrzeug in den Notfallmodus, wenn es in eine Situation gerät, in der es nicht weiterweiß. Bei Systemen des Levels 3 fordert es den

60 Der Begriff der Szene wird von Geyer et al. (2014, S. 184–186) im Rahmen ihres Forschungsartikels über eine einheitliche Ontologie zur Erstellung von Test- und *Use-Case*-Katalogen für das automatisierte Fahren als Bezeichnung für die äußeren Merkmale eines Anwendungsfalls verwendet. Demnach besteht eine Szene aus drei Elementen: der Szenerie (statische Umgebung des Fahrzeugs, z. B. Geometrie von vordefinierten Straßentypen, Anzahl an Fahrstreifen, Straßenverlauf, Position von Verkehrszeichen und Lichtsignalanlagen sowie andere statische Objekte), dynamischen Elementen (andere Verkehrsteilnehmer, Lichtsignalanlagen, Licht- und Wetterbedingungen) und optionalen Fahranweisungen.

Fahrer zur Übernahme auf, während es bei höherstufigen Systemen in einen sicheren Zustand übergeht (vgl. Di Fabio et al., 2017, S. 13, Regel 19); das Gesetz zum autonomen Fahren fordert hier, dass das Fahrzeug bei aktivierter Warnblinkanlage an einer möglichst sicheren Stelle anhält (vgl. Artikel 1, §1d Absatz 4). Dieses Vorgehen setzt allerdings voraus, dass ein Abbremsen noch möglich ist, um Schaden im jeweiligen Fall abzuwenden.

Es kann jedoch auch Situationen geben, in denen dies nicht mehr der Fall ist. Einen Spezialfall innerhalb der Kategorie komplexer, beschränkt antizipierbarer Szenarien stellen daher solche Situationen dar, in denen Schaden *unabhängig* von der gewählten Trajektorie und dem Bremsverhalten des betreffenden Fahrzeugs unvermeidbar ist. Arfini et al. (2022) sprechen von »no-win scenarios«. Derartige Notsituationen sind insbesondere dann denkbar, wenn sich Kollisionsobjekte sowohl in der Fahrspur als auch in den Ausweichbereichen des Fahrzeugs befinden. Gasser (2015, S. 555) konstatiert in diesem Zusammenhang eine »Koinzidenz von […] möglichen Schädigungen«. Da selbstfahrende Fahrzeuge darauf programmiert sind Unfälle zu vermeiden, lässt sich aus technischer Sicht bzw. aus Sicht des Fahrzeugs in derartigen Fällen unter der Zielvorgabe der Unfallvermeidung keine korrekte Trajektorie ermitteln; diese Situationen sind im Rahmen des gegebenen Optimierungsproblems rechnerisch nicht lösbar (vgl. Freitas et al., 2021, S. 4; Gerdes & Thornton, 2015, S. 95). In der Literatur werden dazu verschiedene Szenarien diskutiert. Ein simples, häufig auch in medialen Darstellungen zu Veranschaulichungszwecken aufgegriffenes Beispiel beschreibt Lin (2015, S. 70) als die Entscheidung zwischen zwei Trajektorien, bei denen jeweils eine Person zu Schaden kommen würde:

> Imagine in some distant future, your autonomous car encounters this terrible choice: it must either swerve left and strike an eight-year-old girl, or swerve right and strike an 80-year old grandmother. […] Given the car's velocity, either victim would surely be killed on impact. If you do not swerve, both victims will be struck and killed; so there is good reason to think that you ought to swerve one way or another.

Ein komplexeres Szenario, auf das in ethischen Auseinandersetzungen oft Bezug genommen wird, lautet wie folgt:

> Your car is speeding along a bridge at fifty miles per hour when an errant school bus carrying forty innocent children crosses its path.

> Should your car swerve, possibly risking the life of its owner (you), in order to save the children, or keep going, putting all forty kids at risk? (Marcus, 2012, o. S.)

In Fällen wie diesen lässt sich die Wahl einer Trajektorie nicht länger als rein mathematisches Berechnungs- bzw. Optimierungsproblem modellieren; vielmehr ist die Frage, wie unvermeidbarer Schaden verteilt werden soll, eine genuin ethische, die dort zum Tragen kommt, wo Verkehrsregeln und Gesetze keine ausreichende Handlungsorientierung mehr bieten können:

> Some accident scenarios will be such that (a) there are different options open to the self-driving cars; and (b) depending on what option is selected, different people will be put at risk […]. This is the basic reason why the choice between different possible accident-programs is an inherently ethical choice. (Nyholm, 2018b, S. 2)

Kommt es in jedem Fall zur Schädigung von Personen, sind alle zur Verfügung stehenden Handlungsoptionen aus moralischer Sicht problematisch. Im Forschungsdiskurs werden derartige unvermeidbare Notsituationen aufgrund ihrer besonderen Entscheidungsproblematik als Instanzen moralischer Dilemmata aufgefasst. Dabei handelt es sich um einen spezifischen Typ moralischer Entscheidungsprobleme, der für Situationen charakteristisch ist, in denen zwischen Alternativen gewählt werden muss, welche sich gegenseitig ausschließen und die Beteiligten in unterschiedlichem Maße in negativer Weise betreffen: »A moral dilemma is a situation in which an agent has only the choice between two (or more) options which are not without morally problematic consequences.« (Misselhorn, 2018a, S. 162) Für den Kontext autonomer Fahrsysteme lassen sich relevante Situationen folgendermaßen definieren:

> Dilemmas are defined as critical situations in which, at a given point in time, a CAV will inevitably harm at least one road user and/or one group of road users and the CAV's behaviour will eventually determine which group or individual is harmed. (Europäische Kommission, 2020, S. 32)

Weitgehend unproblematisch aus moralischer Sicht sind im Sinne dieser Definitionen solche Fälle, die mittels der Priorisierung von Sach- über Personenschäden gelöst werden können, sofern keine weitreichenden, menschliches Leben gefährdenden Folgeschäden daraus zu erwarten sind (vgl. Di Fabio et al., 2017, S. 17). Sind

hingegen alle möglichen Alternativen mit Personenschäden bzw. entsprechenden Risiken verbunden, liegt ein Dilemma vor:

> Die dem Dilemma zugrunde liegende Annahme ist, dass keine Alternative zur Schädigung von zwei im Wesentlichen gleichrangigen Rechtsgütern im konkreten Einzelfall denkbar ist, obwohl die maschinelle Fahrzeugsteuerung alle alternativ möglichen Steuerungsentscheidungen berücksichtigt hat. (Gasser, 2015, S. 556)

Dilemmatische Entscheidungen zeichnen sich per definitionem dadurch aus, dass sie sich nicht trivial im Sinne eines Abwägens moralischer Argumente auflösen lassen; es liegt im Wesen eines Dilemmas, dass es keine vollkommen zufriedenstellende Entscheidung geben kann: »To make the point in the most obvious possible way: a moral dilemma *is a dilemma;* it has no clear solution by design—or rather, it poses a problem that is inherently difficult, by design.« (LaCroix, 2022, S. 6, Hervorh. i. Orig.)

Über die Struktur moralischer Dilemmata wäre noch Vieles zu erläutern, was an diesem Punkt des Argumentationsganges jedoch noch nicht erforderlich ist; zunächst ist das grobe Verständnis ausreichend, welches sich aus den obigen Ausführungen ergibt. Die hier vorstellte Definition moralischer Dilemmata ist oberflächlich und an dieser Stelle vorläufig. Sie wird im Rahmen der metaethischen Auseinandersetzung in Kap. 5 präzisiert und vertieft. Zunächst wird jedoch im nachfolgenden Unterkapitel begründet, weshalb Dilemma-Szenarien eine zentrale Rolle bei der Gestaltung von Steuerungsalgorithmen autonomer Fahrzeuge zukommt.

3.2 Die Relevanz von Dilemma-Szenarien für das autonome Fahren

3.2.1 Möglichkeit und Existenz von Unfalldilemmata

Moralische Dilemma-Szenarien stehen häufig im Zentrum des ethischen Diskurses autonomer Fahrsysteme und werden von einem großen medialen Interesse begleitet. Sie sind u. a. Gegenstand der ethischen Leitlinien, welche in den vergangenen Jahren von zahlreichen Expertenkomitees und Kommissionen entwickelt wurden. So verfügte die vom BMVI eingesetzte Ethik-Kommission über eine eigens formierte Arbeitsgruppe unter der Leitung des Rechtswissen-

schaftlers Eric Hilgendorf, welche sich mit den ethischen und rechtlichen Besonderheiten unvermeidbarer Schadenssituationen auseinandersetzte. Jedoch sind Experten und Wissenschaftler gespaltener Meinung, in welchem Maße dilemmatische Szenarien für das autonome Fahren tatsächlich relevant sind. Auch wenn diese zumeist im Kontext des hoch- oder vollautomatisierten Fahrens diskutiert werden, sind sie auch auf geringerem Automatisierungslevel grundsätzlich denkbar. Eine spezifische Problematik stellen sie bei Systemen des Levels 3 dar, wenn das System in zeitkritischen Situationen die Kontrolle an die an Bord befindliche Person zurückgibt, die jedoch unter Umständen nicht aufmerksam war und daher ein reduziertes Situationsbewusstsein hat. Denkbar ist z. B., dass das System bei aktiviertem Staufolgefahren ein mit Methoden der KI-gestützten Mustererkennung nicht näher kategorisierbares Objekt erfasst und die Fahrzeugsteuerung dem Fahrer überantwortet. Diesem bieten sich aufgrund der kurzen Reaktionszeit ausschließlich Handlungsoptionen mit resultierendem Schaden, etwa das Auffahren auf das vorausfahrende Fahrzeug oder ein Ausweichen auf eine parallele Fahrspur mit der Gefahr der Kollision mit einem dort fahrenden Fahrzeug. Um einem derart spezifischen Dilemma zu begegnen, ist die Gestaltung der Schnittstelle zwischen Fahrer und System von großer Bedeutung.

Wie zahlreiche Forscher einerseits betonen, sind moralische Dilemmata beispielsweise für die Maschinenethik von hoher theoretischer Bedeutung, indem sie (maschinen-)ethische Kernfragen nach dem moralischen Status von autonomen Systemen tangieren (vgl. Brändle & Grunwald, 2019, S. 284): Darf eine Maschine bzw. ein Algorithmus im Notfall über Menschenleben entscheiden? Es reicht angesichts der Möglichkeit unvermeidbarer Schäden nicht aus, autonome Systeme nur auf Schadensvermeidung, i. e. im Sinne des Klassifikationsschemas nach Moor (2006, S. 19)[61] als implizite ethische Systeme (*implicit ethical agents*) auf die Vermeidung unethischen Verhaltens hin zu konzipieren:

> Unlike explicit ethical agents, implicit ones do not learn or encode ethics explicitly—and thus, they cannot autonomously arbitrate between different kinds of harm. For example, autonomous cars as implicit eth-

61 James H. Moors berühmtes hierarchisches Schema zur Klassifikation moralischer Akteure wird in Kap. 4.1.2 näher erläutert.

> ical agents strive to avoid crashes—but when a crash is unavoidable, when all trajectories are likely to end up in casualties, implicit ethical agents find themselves dumbfounded, and unable to choose among the different ethical choices. (Bonnefon et al., 2019, S. 502)

Vielmehr müssen sie als explizite ethische Systeme (*explicit ethical agents*) in die Lage versetzt werden, anhand implementierter ethischer Kriterien plausible ethische Urteile zu fällen und diese zu begründen (vgl. Moor, 2006, S. 19–20). Die Konstruktion derartiger Maschinen stellt die gegenwärtige Maschinenethik vor eine große Herausforderung, der durch die Einführung des autonomen Fahrens ein praktischer Bezugskontext und eine gewisse Dringlichkeit gegeben werden.[62]

Andererseits werden kritische Stimmen nicht müde hervorzuheben, dass Dilemma-Szenarien jenseits ihres theoretischen Stellenwerts für die maschinenethische Forschung keine nennenswerte Relevanz für die praktische Seite des autonomen Fahrens besitzen. Eines der häufigsten Argumente ist dabei, dass das Auftreten von Dilemma-Szenarien als eher unwahrscheinlich anzusehen ist. Kaum jemand hat eine derartige Situation im Verkehrskontext je praktisch erlebt bzw. wird eine solche zukünftig erleben (vgl. Roy, 2016). Gegründet auf diesen Mangel an unmittelbarer lebensweltlicher Erfahrung werden Dilemma-Szenarien als grundsätzlich unrealistisch erachtet, die Auseinandersetzung mit ihnen als rein hypothetisches Gedankenexperiment ohne unmittelbare praktische Relevanz. Es liege keine Evidenz für das Auftreten derartiger Szenarien vor, die vom autonomen System ohnehin weder zweifelsfrei erkannt noch kontrollierbar gelöst werden könnten (vgl. Freitas et al., 2020, S. 1285–1286). Aus Sicht kritischer Positionen stellt es eine unverhältnismäßige Anstrengung dar, sich mit Dilemma-Szenarien auseinanderzusetzen und dabei andere ethische Probleme in den Hintergrund zu rücken. So schreibt auch der renommierte Ingenieur Rodney Brooks, ehemaliger Professor am Massachusetts Institute of Technology und Mitgründer von Roboterherstellern, auf seinem Blog:

62 Eine differenziertere Problematisierung moralischer Handlungs(un)fähigkeit von Maschinen im Hinblick auf das Anwendungsbeispiel des autonomen Fahrens hat die Autorin bereits an anderer Stelle veröffentlicht (vgl. Schäffner, 2022).

> This is a made up question that will have no practical impact on any automobile or person for the forseeable [sic] future. Just as these questions never come up for human drivers they won't come up for self driving cars. […] The problem is both non existant [sic] and irrelevant. Nevertheless there is endless hand wringing and theorizing […] about how this is an oh so important problem that must be answered before we entrust our cars to drive autonomously. (Brooks, 2017, o. S.)

Gegen dieses Argument lässt sich jedoch einwenden, dass es versäumt, zwischen realen, praktisch unvermeidbaren Situationen einerseits und idealisierten Szenarien andererseits zu differenzieren. Erstere sind, wie bereits in Kap. 3.2.1 erläutert, aufgrund technischer Unvollkommenheiten der Systeme sowie begrenzter Antizipierbarkeit und nicht-eliminierbarer Eigenheiten des Verkehrsgeschehens durchaus realistisch. Die Ethik-Kommission spricht von Situationen, »die sich bei aller technischen Vorsorge als unvermeidbar erweisen« (Di Fabio et al., 2017, S. 11). Anders liegt der Fall bei übersteigerten Extremszenarien. Hier hat der Vorwurf mangelnden Realismus insofern eine gewisse Berechtigung, als in einschlägigen Diskursen häufig besonders tragische, konstruiert wirkende Konstellationen herangezogen werden.[63] Jedoch wird oft übersehen, dass Letztere nicht primär den Anspruch haben, lebensweltliche Zustände exakt abzubilden. Vielmehr handelt es sich um idealisierte Abstraktionen hochkomplexer Situationen aus der realen Lebenswelt, die es erlauben, zugrundeliegende ethische Problemstellungen zu isolieren und ethisch irrelevante Aspekte auszublenden, sodass adäquate Entscheidungsentwürfe für die Praxis entwickelt werden können: »The job of these thought experiments is to force us to think more carefully about ethical priorities, not to simulate reality.« (Lin, 2017, o. S.)[64]

63 Vor allem in der medialen Darstellung wird die Zuspitzung der gewählten Dilemma-Szenarien häufig übertrieben. Dies sollte jedoch eher als Teil einer medialen Strategie zur Erzeugung von Aufmerksamkeit durch eine Skandalisierung des autonomen Fahrens an sich (vgl. Hilgendorf, 2017b, S. 48) verstanden werden denn als wissenschaftliche Auseinandersetzung.

64 Lin (2017) betont, dass philosophische Gedankenexperimente sich im Grunde kaum von der Vorgehensweise empirischer Wissenschaften unterscheiden. Es handelt sich um abstrakte (und daher nicht realistische) Repräsentationen der realen Welt mit dem Zweck, kontrollierte Bedingungen zu schaffen und Variablen zu isolieren, um den Wirkungszusammenhang zwischen abhängigen und unabhängigen Variablen, wie Anzahl beteiligter Personen, persönliche Merkmale, ihre Bewegungsgeschwindigkeit und -richtung, zu untersuchen. Es gilt zu

Als weiteres Contra-Argument gegen die Relevanz von Dilemma-Szenarien führen Kritiker häufig an, dass sie äußerst selten und daher vernachlässigbar seien. Gegen diese Schlussfolgerung lassen sich zwei ethische Einwände ins Feld führen. Zum einen widerspricht es den Grundsätzen ethischer Praxis, aus der durchaus plausiblen Annahme, dass autonome Fahrzeuge selten in Dilemma-Situationen geraten werden, deren Irrelevanz zu folgern. So wird anhand der ethischen Debatte über die Atomenergie deutlich, dass unwahrscheinliche Szenarien – wie das Eintreten nuklearer Katastrophen – eine zentrale Rolle bei der ethischen Bewertung von Technologien einnehmen (vgl. Misselhorn, 2018b, S. 9–10). Die ethische Bedeutung eines problematischen Ereignisses ist prinzipiell unabhängig von seiner Eintrittswahrscheinlichkeit; im Gegenteil: Katastrophenszenarien gehören zu den ethisch brisantesten Fällen. Entscheidend ist dabei vor allem die Höhe des möglichen Schadens (vgl. Bhargava & Kim, 2017, S. 9–10):

> When harm is possible or inevitable, the vehicle will need to make a decision, which means that it needs to have been programmed or trained to be capable of making a decision. And, this is true regardless of how rare the circumstances might be in practice. (LaCroix, 2022, S. 3)

Zum anderen impliziert eine solche Position die Installation einer probabilistischen Ethik, wie sie seit Jahren in vielen Risikodebatten forciert wird. Die Tatsache, dass der Eintrittswahrscheinlichkeit eines möglichen Schadens in diesem Zuge eine »eigene moralische Qualität beigemessen [wird], die ethisch nicht zu rechtfertigen ist«, bezeichnet Ropohl (2017, S. 887) als »Kalamität«. Insbesondere vor dem Hintergrund zeitgenössischer Verantwortungsbegriffe erscheint eine »Moralisierung der Wahrscheinlichkeit« (ebd., S. 904) von Grund auf fragwürdig.

Aus ingenieurtechnischer Sicht ist die Berücksichtigung moralischer Dilemma-Szenarien ein zentrales Kriterium für die Robustheit des Designs automatisierter Fahrzeuge. So ist das proaktive Treffen von Vorkehrungen für das Eintreten eines *worst case* zentraler Bestandteil jedes Sicherheitskonzepts, das den Standards funktionaler Sicherheit genügt. Gemäß dem Prinzip des *Safety by Design* ist es

---

erforschen, wie sich die Veränderung dieser Variablen auf unsere moralische Intuition auswirkt. Wenn fünf anstelle von zwei Personen beteiligt sind, würde ich dann anders entscheiden?

in der Softwareentwicklung seit Langem gängige Praxis, sogenannten *edge cases* besondere Aufmerksamkeit zu widmen, um mögliche kritische Situationen bereits bei der Spezifikation erfassen und qualitativ hochwertige Software entwickeln zu können (vgl. Lin, 2017; Reschka, 2015, S. 500). Dieses Vorgehen hat auch eine verantwortungsethische Komponente:

> Not programming the car for how to respond to situations like this and others like it amounts to knowingly relinquishing the important responsibility we have to try to control what happens in traffic. It amounts to unjustifiably ignoring the moral duty to try to make sure that things happen in good and justifiable ways. We should not do that. Hence the need for ethical accident-algorithms. (Nyholm & Smids, 2016, S. 1278–1279)

Diese Verantwortung schlägt sich direkt in gesetzlichen Bestimmungen nieder: Herstellern obliegt im Rahmen der zivil- und strafrechtlichen Produkthaftung eine Pflicht, sämtliche zumutbaren Maßnahmen zur Risikominderung bei ihren entwickelten Produkten zu ergreifen: »[…] das Hervorrufen von Schäden, wie sie in der Massenproduktion von technischen Produkten praktisch unvermeidlich sind, [ist] nicht als fahrlässig anzusehen, wenn der Hersteller alles in seiner Macht Stehende getan hat, um derartige Schäden zu vermeiden.« (Hilgendorf, 2019, S. 363)

Nun wirkt sich die legitimerweise bemängelte, übertriebene Fokussierung auf Dilemma-Situationen negativ auf deren generelle Glaubwürdigkeit aus (vgl. Bonnefon et al., 2019, S. 503). Eine zentrale Rolle kommt in diesem Zusammenhang der Überbetonung einer vermeintlichen Analogie zwischen unvermeidbaren Unfallsituationen und dem Trolley-Problem zu.[65] Dilemma-Situationen würde per se jegliche praktische Bedeutsamkeit abgesprochen, wenn sie lediglich als anwendungsnahe Instanzen von Trolley-Fällen aufgefasst würden, deren Plausibilität, wie später gezeigt wird, leicht angreifbar ist:

> When the media refers to the trolley problem in the context of vehicle automation, they seem to use it as a stand-in for a range of more

65 Das Trolley-Problem ist ein philosophisches Gedankenexperiment, das moralische Präferenzen in dilemmatischen Entscheidungssituationen untersucht. Mögliche Analogien zwischen Trolley-Problem und Unfallalgorithmen werden in Kap. 4.1.4 diskutiert.

> subtle ethical decisions an automated vehicle may face, many of which will have less obvious moral undertones, uncertain outcomes, and consequences that are not life-threatening. This is a problem because critics of automated vehicle ethics can argue that any research into ethical decision making for automated vehicles is unnecessary or wasteful simply by attacking the trolley problem. (Goodall, 2016a, S. 812)

Die vorherrschende Motivation der Kritiker von Dilemma-Szenarien ist es, den Fokus auf dringendere praktische Probleme zu lenken, welche vor allem im Zusammenhang mit Strategien zur Unfallvermeidung auftreten. Hierbei geht es vordergründig um Abwägungen ethisch relevanter Ziele, wie beispielsweise zwischen Sicherheit und Effizienz im Sinne von Zeitverlust durch eventuell nötige Geschwindigkeitsreduzierung in spezifischen Fahrsituationen (vgl. Hansson et al., 2021, S. 1393; Nyholm, 2018c, S. 8, Endnote 5). Welcher (Mindest-)Abstand soll zu anderen Verkehrsteilnehmern eingehalten werden? Wie soll die Bremsreaktion bei gelber Ampel geregelt werden? Für derartige Situationen lautet die allgemeine Empfehlung, das Verhalten autonomer Fahrzeuge an einer vernunftgesteuerten, Common-Sense-Fahrweise auszurichten, welche im Kern auf einer alltagstauglichen Heuristik, der Minimierung des Gesamtschadens bzw. des absoluten Schadensrisikos, basiert (vgl. Freitas et al., 2021, S. 2–6).

Mögliche Unfalldilemmata jedoch zugunsten wahrscheinlicherer Alltagssituationen gänzlich zu vernachlässigen, zeugt von einer Haltung, die die Bedeutung von dilemmatischen Szenarien für das autonome Fahren unterschätzt. So wird in der Literatur, die Dilemma-Szenarien kritisch gegenübersteht, häufig ein entscheidender Punkt übersehen: Das zugrundeliegende Entscheidungsdilemma tritt implizit weitaus häufiger in alltäglichen Fahrsituationen auf, als uns bewusst ist. Als beispielhafte Situation beschreibt Lin (2017) ein autonomes Auto, das durch eine enge Gasse fährt, in der sich links eine Gruppe von Personen befindet, rechts nur eine Person. Wo soll sich das Auto in der Fahrspur positionieren – mittig, eher rechts, eher links?[66] Oder mit welcher Intensität soll das Fahrzeug bei einem auf die Straße laufenden Tier bremsen, um mögliche Auffahrunfälle im nachfolgenden Verkehr zu vermeiden (vgl. Lin, 2013b)? In bei-

66 Goodall (2016b, S. 31) beschreibt ein ähnliches Szenario auf einer dreispurigen Straße, wo sich das autonome Fahrzeug in der mittleren Fahrspur zwischen einem LKW und einem kleinen PKW positionieren muss.

den Beispielen enthalten die Fahrentscheidungen implizite ethische Werturteile:

> The behavior of the vehicle can have an ethical component, even if the vehicle is not in immediate danger. The decisions of how to position itself within a lane, how much buffer to provide a pedestrian, whether the buffer size should change based on a pedestrian's behavior or physical attributes, what type of headway to allow—these all carry some risk of crashing. […] More subtle but just as difficult choices exist for almost all driving. (Goodall, 2016a, S. 813–814)

In alltäglichen Verkehrssituationen findet eine ethisch problematische Wahl zwischen verfügbaren Trajektorien statt, die gleichbedeutend ist mit der Zuweisung relativer Schadensrisiken an involvierte Personen bzw. Parteien:

> Even if every action of an autonomous car is oriented toward minimizing the absolute risk of a crash, each action will also shift relative risk from one road user to another […]. The cars may not be making decisions between outright sacrificing the lives of some to preserve those of others, but they will be making decisions about who is put at marginally more risk of being sacrificed. (Bonnefon et al., 2019, S. 503)

Auch wenn es in diesen »low stakes scenarios« (Freitas et al., 2021, S. 4) möglicherweise gar nicht um tatsächliche Unfälle, sondern lediglich um entsprechende Risiken geht, ist das zugrundeliegende Entscheidungsproblem des Abwägens von Alternativen mit potenziell ungünstigen Folgen dasselbe wie jenes, das Dilemma-Szenarien stellen. Daraus folgt, dass diese im Grunde lediglich ein Problem pointieren, das auch in vielen alltäglichen Situationen auftritt und notwendigerweise thematisiert werden muss:

> […] mundane driving situations iterated over time can lead to injuries and deaths. In this way, such situations are not all that different from trolley cases or real-world collision scenarios. The only difference lies in the degree of risk and uncertainty: death or harm to at least one party is unavoidable in trolley cases, and almost unavoidable in real-world collision scenarios, while most mundane driving situations have a significantly lower risk of harm. Despite this difference, mundane driving situations are rendered ethically challenging in part precisely because of their structural similarity to trolley cases and real-world collision scenarios: just like these, mundane driving situations will involve decisions about risk distribution between AV users and other road users. (Brändle & Schmidt, 2021, S. 1490)

Vor dem Hintergrund dieser Argumente kann die Auseinandersetzung mit dem ethischen Entscheidungsproblem, das moralische Dilemmata auszeichnet, entgegen der Meinung der Kritiker legitimerweise als essenziell bewertet werden. Dilemma-Szenarien liefern trotz geringer Eintrittswahrscheinlichkeit und teilweise mangelndem Realismus in jedem Fall einen wichtigen Beitrag zur Thematisierung weniger dramatischer, subtil problematischer Alltagsszenarien.[67] Auch aus der Summe vieler kleiner automatisierter Entscheidungen ergibt sich bei Milliarden von zurückgelegten Kilometern letztlich doch eine ethisch sehr brisante Thematik (vgl. Bonnefon et al., 2019, S. 503). Da autonome Fahrzeuge in Dilemma-Situationen einer vorgegebenen Entscheidungslogik folgen, ist es eine praktische Notwendigkeit, diese im Rahmen des (Software-)Designs bereitzustellen (vgl. Millar, 2014c). Ethischen Fragen wird daher eine erfolgskritische ›Klammerfunktion‹ hinsichtlich der Einführung des autonomen Fahrens zugesprochen:

> Erst wenn es gelingt, autonom agierenden Fahrzeugen eine Art von Entscheidungsethik mitzugeben, vermag sich die Fahrrobotik auch in der Praxis zu behaupten. Dies gilt insbesondere für sogenannte Dilemma-Situationen, in denen eine Abwägung getroffen werden muss, welches Verhalten im Falle einer unvermeidbaren Kollision den beteiligten Personen innerhalb und außerhalb des Fahrzeugs den geringsten Schaden zufügt. (Minx & Dietrich, 2015, S. VI)

Unfallszenarien mit dilemmatischen Strukturen sind also nicht nur prinzipiell möglich, sondern sie treten auch tatsächlich auf. Doch wie können sie entschieden werden? Im nächsten Unterkapitel wird ausgeführt, weshalb eine ›einfache Entscheidung‹, welche Unfallalgorithmen mittels Heuristiken normiert, der Komplexität des Problems nicht gerecht wird.

---

67 Eine treffende Formulierung dieser Erkenntnis findet sich auch bei Fried (2012, S. 512), die sich zwar auf das klassische Trolley-Problem bezieht, deren hier zitiertes Argument sich aber verallgemeinern lässt: »By presenting tragic choices only in ›extreme and desperate,' indeed (outside of the context of war) freakish, circumstances, the trolley literature has inadvertently led both authors and consumers of that literature to regard tragic choices *themselves* as rarely occurring and freakish in nature. But they are neither of these things. They are ubiquitous and for the most part quotidian, and typically result [...] from the finite nature of the resources we depend on to realise our projects in the world.«

3.2.2 Sind Unfallalgorithmen normierbar?

Wenn es um mögliche Entscheidungsstrategien für Dilemma-Szenarien im autonomen Fahren geht, taucht im einschlägigen Literaturdiskurs immer wieder die Fragestellung auf, inwiefern Unfallalgorithmen über die Implementierung einfacher Heuristiken wie z. B. ›immer bremsen‹ oder ›immer ausweichen‹ normiert werden können. Voraussetzung für jegliche Überlegungen dieser Art ist, dass eine allgemeingültige, triviale Standardstrategie existiert, die für alle denkbaren Fälle stets die beste aller verfügbaren Optionen darstellt.

Einen praxisorientierten Ansatz, der sich an diesem Ziel orientiert, legt Davnall (2020) vor. Unter Bezugnahme auf fahrphysikalische Mechanismen statischer und kinetischer Reibung argumentiert sie, dass maximales Abbremsen bei gleichzeitigem Spurhalten stets zum geringsten Schadensrisiko führt und somit sämtlichen Ausweichmanövern vorzuziehen ist.[68] Davnalls Vorschlag unterliegt jedoch ernsthaften Limitierungen, die seine Eignung für einen praktischen Einsatz in Frage stellen. So ist er erstens nur für vergleichsweise triviale Szenarien im Stadtverkehr plausibel, an denen nur ein Fahrzeug mit niedriger Geschwindigkeit beteiligt ist. Für Landstraßen und Autobahnen muss realistischerweise davon ausgegangen werden, dass weitere Fahrzeuge involviert sind; hier würden Auffahrunfälle billigend in Kauf genommen, bei denen schwerwie-

68 Dafür führt Davnall zwei Gründe an: Zum einen verringert zeitgleiches Ausweichen die Bremswirkung, wodurch die Wucht des Aufpralls mit dem Kollisionsobjekt weniger stark gedämpft wird. Zum anderen besteht bei Ausweichmanövern die Gefahr, dass das Fahrzeug ins Schleudern gerät und unvorhersehbare Trajektorien einschlägt. Problematisch ist hierbei, dass z. B. für Fußgänger bei seitlichem Aufprall ein deutlich höheres Verletzungsrisiko besteht als bei einem Kontakt mit der frontalen Knautschzone eines Fahrzeugs (vgl. Davnall, 2020, S. 440–441). Für Davnall ist die Kontrollierbarkeit des resultierenden Risikos daher der entscheidende Faktor: »The car does not face a decision between hitting an object in front of it and hitting an object off to one side. Instead, the decision is better described as being between a controlled manoeuvre—one which can be proven with generality to result in the lowest impact speed of any available option— and a wildly uncontrolled one.« (Ebd., S. 442–443) Ließen sich Szenarien auf diese Weise normieren und damit eindeutig entscheiden, handelt es sich streng genommen nicht mehr um Entscheidungsdilemmata. Diese treten nach Davnalls Auffassung nur dann auf, wenn Bremsvorgänge nicht ausgeführt werden können – beispielsweise, wenn die Bremsen versagen, was allerdings sehr selten vorkommt.

gende Folgen nicht auszuschließen sind. Zweitens wird außer Acht gelassen, dass sich dynamische Kollisionsobjekte wie z. B. Fußgänger irrational verhalten und unerwartet ihre Position verändern können, sodass ihr Schadensrisiko nicht allein von den Aktionen des involvierten autonomen Fahrzeugs abhängt, sondern auch von ihren eigenen. Das Verkehrsgeschehen ist ein offenes System mit unendlich vielen Szenarien, die von dynamischen Faktoren abhängig sind. Drittens können fahrphysikalische Eigenschaften je nach Witterung variieren und z. B. Bremswege bei Schnee oder starker Nässe verändern. Viertens – und das wiegt möglicherweise am schwersten – beschränkt sich Davnalls Entwurf auf eine fahrzeugdynamische Perspektive, die jegliche Sensibilität für die ethische Dimension von Entscheidungsdilemmata vermissen lässt. Auf dieser Grundlage lässt sich argumentieren, dass rein physikalische Ansätze grundsätzlich nicht geeignet sind, um ethische Probleme zu entscheiden:

> […] some decisions are more than just a mechanical application of traffic laws and plotting a safe path. They seem to require a sense of ethics, and this is a notoriously difficult capability to reduce into algorithms for a computer to follow. (Lin, 2015, S. 69)

Welche Entscheidungsstrategien resultieren nun aus einer ethischen Würdigung von Dilemma-Szenarien? Angesichts der enormen Vielfalt und des hohen Komplexitätsgrades möglicher Szenarien und der jeweils tangierten ethischen Problemstellungen, die u. a. durch die zuvor beschriebenen Beispielszenarien zum Ausdruck kommen, erscheint es plausibel, Dilemma-Situationen als individuelle Einzelfälle zu beurteilen, die sich nicht mit vergleichsweise einfachen Heuristiken entscheiden lassen (vgl. Birnbacher & Birnbacher, 2016, S. 12; Lin, 2017). Dies steht in Einklang mit der Feststellung der Ethik-Kommission, dass eine abstrakt-generelle Regelung über alle denkbaren Szenarien hinweg prinzipiell fragwürdig ist, was auch die Priorisierung von Sach- über Personenschäden einschließt, sofern Konsequenzen katastrophalen Ausmaßes zu erwarten sind:

> Das Problematische an Dilemma-Situationen ist […], dass es sich um Entscheidungen handelt, die aus dem konkreten Einzelfall heraus bei Betrachtung verschiedener Faktoren heraus getroffen werden müssen. Konkrete Normierungen wie zum Beispiel ›Personenschaden vor Sachschaden‹ erscheinen daher bei Dilemma-Situationen zwar möglich, aber als abstrakt generelle Regelung werfen sie Zweifel in Fällen auf, in denen zum Beispiel die Folge eines Sachschadens das Auslaufen

> eines Tanklasters oder auch der Zusammenbruch des Stromnetzes einer Metropolregion sein könnte. Abstrakt generelle Regelungen wie Sachschaden vor Personenschaden treffen bei der Vielfalt und Komplexität der verschiedenen denkbaren Szenarien auf das Problem, dass eine Normierung aller Situationen nicht möglich ist. Die Prämisse der Minimierung von Personenschäden kann nur dann konsequent eingehalten werden, wenn eine Folgenabschätzung bei Sachschäden versucht wird und mögliche folgende Personenschäden in das Verhalten bei Dilemma-Situationen einkalkuliert werden. (Di Fabio et al., 2017, S. 17)

Ein weiteres Argument, das einer auf trivialen Heuristiken basierenden Entscheidungsstrategie für Dilemma-Szenarien eine Absage erteilt, fußt auf verantwortungsethischen Überlegungen. Die Verantwortung für die Aktionen autonomer Systeme liegt bei denjenigen, die festsetzen, wie diese in einer bestimmten Situation agieren sollen, i. e. beim »Hersteller und Betreiber der technischen Systeme und […] [den] infrastrukturellen, politischen und rechtlichen Entscheidungsinstanzen.« (Ebd., S. 11) Problematisch wird diese Verantwortungsverschiebung dann, wenn die normierte Programmierung eines autonomen Fahrzeugs im konkreten Dilemma-Fall zu einem suboptimalen Ergebnis führt. Lin (2017) erklärt, dass dies besonders im Fall von Standardimplementierungen gravierende Rechtsfolgen für Hersteller bzw. Entwickler nach sich ziehen könnte, denn der Verzicht auf eine sorgfältige Einzelfallbetrachtung käme einer Verletzung der Sorgfaltspflicht bei der Produktentwicklung gleich:

> If a human driver made a fatal snap-decision in an emergency, it'd just be a tragic accident, and we'd be hard-pressed to blame the driver. But if an AI driver made the exact same decision, it's no longer an unfortunate reflex but more like premeditated homicide; a self-driving car must be programmed and its behavior scripted or purposely trained. So, there could be implications for legal liability. […] It might be that there's no ›right‹ decision, but to systematically decide in a certain way—for instance, to always protect the driver *über alles*—could be faulted, especially if that design decision was made unilaterally by the company and in secret. (Ebd., o. S., Hervorh. i. Orig.)

Ferner bildet das Sicherheitspotenzial autonomer Fahrsysteme die zentrale Legitimationsgrundlage für die angestrebte Automatisierung des Verkehrs. Die Ethik-Kommission sieht eine positive Risikobilanz als das entscheidende Kriterium dafür an, dass die Zulassung autonomer Fahrzeuge für den öffentlichen Straßenverkehr

vertretbar ist (vgl. Di Fabio et al., 2017, S. 10). Kurz gesagt: Nur wenn automatisierte Systeme die Sicherheit tatsächlich erhöhen, ist ihre Einführung gerechtfertigt. Eine rein aggregierte Sichtweise auf den potenziellen Sicherheitszuwachs ist hier allerdings nicht ausreichend; vielmehr ist es erforderlich, dass autonome Fahrzeuge in *jeder* möglichen Situation eine nicht nur gleichwertige, sondern *bessere* Fahrentscheidung (im Sinne höherer Sicherheit bzw. geringeren Schadens) treffen als der Mensch. Das Postulat der Optimierung automatisierten Verhaltens schließt demnach auch Unfallsituationen mit dilemmatischen Strukturen ein. Wie zuvor bereits gezeigt, stellt eine Normierung derselben in keiner Weise einen adäquaten Ansatz dar, die den beschriebenen Anforderungen genügt: »Technische Systeme […] sind […] auf eine komplexe oder intuitive Unfallfolgenabschätzung nicht so normierbar, dass sie die Entscheidung eines sittlich urteilsfähigen, verantwortlichen Fahrzeugführers ersetzen oder vorwegnehmen könnten.« (Ebd., S. 11, Regel 8)

Zusammenfassend kann festgehalten werden, dass Dilemma-Szenarien sich nicht pauschal entscheiden lassen, sondern nur im Rahmen einer differenzierten ethischen Einzelfallbetrachtung. So stellt die Ethik-Kommission neben der Absage an eine triviale, heuristische Entscheidungsstrategie für Dilemma-Szenarien fest, dass Letztere ebenfalls »nicht ethisch zweifelsfrei programmierbar« (2017, S. 11) sind. Auch wenn Unfalldilemmata keine eindeutige Lösung haben, müssen sie aus praktischer Sicht kein unüberwindbares Hindernis darstellen, sofern die Programmierung von Unfallalgorithmen nicht länger als die Suche nach der ›einzig richtigen Antwort‹ missverstanden wird. Vielmehr muss der Fokus auf der argumentativen Herleitung ethisch vertretbarer Entscheidungsstrategien liegen, welche sich im Zuge eines Prozesses erarbeiten lassen, bei dem Entscheidungen für oder gegen konkrete Handlungsoptionen sorgfältig auf der Grundlage moralischer Überlegungen getroffen werden:

> […] what's important isn't just about arriving at the ›right‹ answers to difficult ethical dilemmas, as nice as that would be. But it's also about being thoughtful about your decisions and able to defend them – it's about showing your moral math. (Lin, 2014a, o. S.)

Jenseits der ethischen Perspektive auf Unfalldilemmata lässt sich deren Bedeutung auch im Hinblick auf gesellschaftliche und technische Aspekte argumentativ untermauern; dies wird im Folgenden ausgeführt.

3.2.3 Gesellschaftliche und technische Relevanz von Dilemma-Szenarien

In Ergänzung zu den bisher skizzierten Argumenten für die Relevanz unvermeidbarer Unfallsituationen, die sich auf deren Existenz und die Möglichkeit ihres Auftretens beziehen, sollen an dieser Stelle noch zwei weitere relevante, praxisbezogene Perspektiven Beachtung finden. Zum einen bescheinigen zahlreiche empirisch gestützte Forschungsbeiträge Dilemma-Situationen eine zentrale Bedeutung für die Akzeptanz autonomer Fahrzeuge durch potenzielle Nutzer. Dabei wird beispielsweise mangelndes Vertrauen in die Sicherheit der Technologie, welche sich in (ethischen) Notsituationen als besonders kritisch erweist, als eines der gewichtigsten sozialen und psychologischen Hindernisse identifiziert, die einer Masseneinführung des autonomen Fahrens im Wege stehen (vgl. Adnan et al., 2018, S. 824–828; Choi & Ji, 2015, S. 694–700; Edmonds, 2019; Shariff et al., 2017, S. 694–695). Erklärbarkeit (*explainability*) und Transparenz (*transparency*)[69] sind zentrale Anforderungen für die soziale Akzeptanz von algorithmischen Entscheidungssystemen, denn komplexe und intransparente Algorithmen geben den Nutzern das Gefühl, die Kontrolle zu verlieren und der Maschine ausgeliefert zu sein. Verlässlichkeit und Vertrauenswürdigkeit[70] sind essenzielle Qualitäten für die Adoption autonomer Technologien (vgl. Chilson, 2022, S. 232–235; Othman, 2021, S. 358–360). Entsprechende Mängel manifestieren sich in einer grundsätzlichen Skepsis gegenüber statistischen Algorithmen, was unter dem Begriff der *Algorithm Aversion* gefasst wird (vgl. Dietvorst et al., 2015, S. 115; Shariff et al., 2017, S. 695). Menschliche Fehler sind zu einem gewissen Grad abschätzbar[71], algorithmische nicht; die Frage nach dem sozial akzeptablen

69 Chilson (2022, 235–239) beschreibt sechs Merkmale, die als Desiderate für das Design autonomer Fahrsysteme geeignet sind, angemessenes Vertrauen der Nutzer zu generieren: Wiederholbarkeit, Vorhersehbarkeit, Zuverlässigkeit, Transparenz, Rekonstruierbarkeit und Erklärbarkeit.

70 Siehe hierzu auch den Beitrag von Weydner-Volkmann (2021), der den Begriff ›Technikvertrauen‹ als komplementäres Konzept zu ›Akzeptanz‹ und ›Akzeptabilität‹ entwickelt.

71 Erwähnenswert ist in diesem Zusammenhang ein Beitrag von Zerilli et al. (2019), der aufzeigt, dass auch viele menschliche Entscheidungen mit Transparenzproblemen behaftet sind.

Risiko (»Wie sicher ist sicher genug?«) ist zentral, wenn es um die Adoption neuer Technologien geht.[72]

Nun legen im Kontext autonomer Fahrzeuge durchgeführte empirische Untersuchungen nahe, dass moralische Dilemma-Situationen aus Sicht potenzieller Nutzer stark negative Affekte transportieren.[73] Sie werden mit hohen Risiken erheblicher körperlicher Schäden assoziiert und deshalb als bedeutsamer im Vergleich zu anderen technischen, rechtlichen und ethischen Herausforderungen betrachtet (vgl. Gill, 2021, S. 662–669). Negative öffentliche Reaktionen auf reale Unfälle mit Beteiligung autonomer Fahrzeuge sind geeignet, das Vertrauen potenzieller Nutzer in die Fahrautomatisierung zusätzlich zu untergraben. Sind die Verbraucher nicht von der Sicherheit der Fahrzeuge im Notfall überzeugt, verzichten sie unter Umständen auf eine Investition bzw. Nutzung. Bonnefon et al. (2020, S. 109–111) beschreiben dies als »*Opt-Out*«-Problem.[74] Die in der Folge stagnierende Nachfrage würde aufgrund ihrer stimulierenden Wirkung auf

72 Diese Frage lässt sich nicht technologisch, sondern nur psychologisch oder soziologisch beantworten. An dieser Stelle sei auf weiterführende empirische Studien zur Nutzerakzeptanz und Risikowahrnehmung autonomer Fahrsysteme verwiesen. So untersuchen Brell et al. (2019), wie autonome Fahrzeuge hinsichtlich ihres diversen Risikopotenzials wahrgenommen werden. Eine zentrale Rolle für die Risikowahrnehmung spielen dabei die individuellen Vorerfahrungen, die potenzielle Nutzer mit autonomen Fahrfunktionen gemacht haben. Diese These stützen auch Raue et al. (2019), die in ihrer Studie den Einfluss von Gefühlen auf die Akzeptanz erforschen. Relevant ist in diesem Kontext auch, dass von selbstfahrenden Autos begangene Fehler oft anderer Art sind als menschliche. Wie Prototypenfahrten zeigten, kommt es in verhältnismäßig trivialen Situationen zu Problemen, wenn die Fahrsysteme nicht weiterwissen, weil z. B. spontane Baustellen den Fahrtweg versperren.

73 Die dominante verhaltensökonomische Forschung geht von der *Prospect*-Theorie aus, der zufolge ein Schlüsselfaktor für die Nutzerakzeptanz von Innovationen in der relativen Risikowahrnehmung von Individuen liegt, die grundsätzlich verlustavers sind und Risiken höher gewichten als potenzielle Vorteile (vgl. Kahneman & Tversky, 1979, S. 274–288). Dabei wirken sogenannte Affekt-Heuristiken: Ereignisse, die starke affektive bzw. emotionale Reaktionen hervorrufen, erhalten überproportionales Gewicht in der Entscheidungsfindung (vgl. Slovic et al., 2007, S. 1336–1349), wobei die Wahrscheinlichkeit von deren Auftreten vernachlässigt wird (vgl. Rottenstreich & Hsee, 2001, S. 186–190; Sunstein, 2003, S. 123–129).

74 Dieses Problem ist vielschichtig: Auch wenn Dilemma-Szenarien bei der Programmierung von Unfallalgorithmen berücksichtigt werden, ist es dennoch wahrscheinlich, dass Verbraucher von einer Nutzung Abstand nehmen, sofern

Kapitalinvestitionen und Produktionskapazitäten eine großflächige Einführung des autonomen Fahrens in weite Ferne rücken lassen (vgl. Kumfer & Burgess, 2015, S. 130). Damit würden auch erwartete positive Effekte der Verkehrsautomatisierung zunächst ausbleiben; die anvisierte Mobilitätswende würde ausgebremst. Dilemma-Szenarien kommt daher eine hohe emotionale Bedeutung und ein unverhältnismäßig starkes Gewicht bei individuellen und öffentlichen Entscheidungen zu, die sich auf die Entwicklung und Akzeptanz autonomer Fahrzeuge beziehen (vgl. Bonnefon et al., 2015, S. 3; Misselhorn, 2018b, S. 189). Entsprechend konstatiert Lin (2014a, o. S.): »Often, the rare scenarios are the most important ones, making for breathless headlines.«

Zum anderen sprechen auch technische Gründe dafür, dass Dilemma-Szenarien bei der Entwicklung höherstufig automatisierter Fahrzeuge Berücksichtigung finden müssen. Diese hängen mit der prinzipiellen Begrenztheit maschineller Leistungsfähigkeit und Logik zusammen. So besagt das in den 1980er-Jahren entdeckte Moravec'sche Paradoxon, dass – entgegen lange Zeit gehegter Annahmen in der Forschung – intellektuell anspruchsvolle Tätigkeiten des Denkens weniger maschinelle Rechenleistung erfordern und daher von Systemen Künstlicher Intelligenz sehr gut erlernt werden können. Im Gegensatz dazu benötigen jedoch sensomotorische Tätigkeiten, die Menschen tendenziell leichtfallen, hohe maschinelle Rechenkapazitäten:

> […] it has become clear that it is comparatively easy to make computers exhibit adult-level performance in solving problems on intelligence tests or playing checkers, and difficult or impossible to give them the skills of a one-year old when it comes to perception and mobility. (Moravec, 1988, S. 15)

Anders ausgedrückt: Menschen und Maschinen haben unterschiedliche Stärken und Schwächen. Ein Algorithmus hat Vorteile in Bezug auf Prozesse der Datenverarbeitung, er lässt sich nicht ablenken oder von Emotionen leiten (vgl. Kirkpatrick, 2015, S. 19). Jedoch hat er Schwierigkeiten, Objekte korrekt zu erkennen und das Verhalten anderer Verkehrsteilnehmer zu interpretieren; Techniken maschineller

---

das im Notfall aktivierte Unfallverhalten sie selbst einem (subjektiv) inakzeptablen Risiko aussetzt (siehe Diskussion in Kap. 7.3).

Perzeption reichen (noch) nicht an menschliches Urteils- und Reaktionsvermögen heran (vgl. Fagnant & Kockelman, 2015, S. 169–170; Kirkpatrick, 2015, S. 19; Winkle, 2015, S. 369). Auch die menschliche Fähigkeit, in konkreten Situationen intuitiv entscheiden zu können, mit welcher Intensität wir beispielsweise bremsen oder beschleunigen, lässt sich nicht ohne Weiteres in Algorithmen übersetzen (vgl. Himmelreich, 2018, S. 678). Hinzu kommt, dass Maschinenlogik angesichts von hypothetischen, unsicheren oder mehrdeutigen Szenarien an ihre Grenzen stößt. Eine Maschine muss sich stets in einem innerhalb ihres Systems definierten Zustand befinden, d. h. sie muss zu jedem Zeitpunkt die möglichen Ausgänge ihrer Handlungen kennen,[75] und die Übergänge in mögliche Folgezustände müssen entsprechend des vom System erkannten Eingabealphabets definiert sein (vgl. Wallach & Allen, 2008, S. 86–89).[76] Ist dies nicht der Fall, gelangen die Systemalgorithmen unter Umständen nicht zu einem Ende, wodurch es zum Systemabsturz kommen kann. Eben diese Unfähigkeit, mit unvorhergesehenen Ereignissen im Straßenverkehr angemessen umzugehen, macht es erforderlich, Maschinen für eine möglichst große Bandbreite an denkbaren Situationen mit Leitlinien zur Handlungsorientierung auszustatten (vgl. Lin, 2013b). Dies gilt umso mehr, wenn es sich um Situationen handelt, in denen Personenschäden unabwendbar sind.

75 In einer dynamischen Verkehrsumgebung lassen sich die unmittelbaren Folgen einer Handlung in den seltensten Fällen eindeutig bestimmen. Aufgrund des interaktiven Charakters von Verkehrssituationen ist der Raum potenziell möglicher Szenarien prinzipiell unbegrenzt; dennoch ist es im Hinblick auf die Robustheit von Systemen erforderlich, so viele Fälle wie möglich abzudecken.

76 Zustandsübergänge können verschiedener Art sein. Sie können sich beispielsweise an Verkehrsregeln orientieren – das Fahrzeug geht dann aufgrund einer Geschwindigkeitsbeschränkung in einen Fahrzustand mit geringerer Geschwindigkeit über. Oder sie können eine Reaktion auf das Verkehrsverhalten anderer Verkehrsteilnehmer sein, etwa wenn das vorausfahrende Fahrzeug bremst. Im Fall moralischer Dilemma-Situationen erfolgen Zustandsübergänge anhand von moralischen Kriterien, die z. B. darüber entscheiden, ob ausgewichen wird oder nicht.

3.3 Zwischenergebnis: Die zentrale Bedeutung von Dilemma-Szenarien

Ziel des ersten Teils des Buches war es, die Problemstellung der Gestaltung von Unfallalgorithmen in den bestehenden Forschungsdiskurs des autonomen Fahrens einzuordnen und dabei die besondere Bedeutung herauszuarbeiten, die dilemmatischen Unfallszenarien zukommt. An dieser Stelle sollen die zentralen Ergebnisse im Hinblick auf diese Zielsetzung nochmals in prägnanter Form zusammengefasst werden.

Ansätze einer ethischen Untersuchung von Dilemma-Szenarien werden häufig auf der Basis des Arguments kritisiert, dass Letztere keine oder nur eine marginale Bedeutung für drängende Forschungsfragen rund um das autonome Fahren besäßen. Unfalldilemmata seien unwahrscheinlich bzw. sehr selten, oftmals übersteigert dargestellt und unberechtigterweise mit hoher medialer Aufmerksamkeit bedacht. Wie in diesem Teil der Forschungsarbeit deutlich gemacht wurde, sind diese Kritikpunkte jedoch in vielerlei Hinsicht unzutreffend. Aus theoretischer Sicht spielen Dilemma-Szenarien des autonomen Fahrens als *Use Cases* v. a. für den maschinenethischen Diskurs eine große Rolle, indem sie zentrale Fragen maschineller Handlungsfähigkeit in einen praktischen Zusammenhang stellen. Ferner weisen Unfalldilemmata eine hohe praktische Relevanz für die Entwicklung autonomer Fahrzeuge auf. Das zentrale Argument lautet hier, dass die dilemmatische Grundstruktur, die dem Entscheidungsproblem zugrunde liegt, implizit in vielen kleinen, alltäglichen Fahrentscheidungen enthalten ist. Entscheidungen über Abstände, Geschwindigkeiten oder Trajektorien implizieren stets eine Abwägung relevanter Handlungsgründe bzw. Werte, auch wenn es uns nicht immer bewusst ist. Die Auseinandersetzung mit Unfalldilemmata ist daher eine ethische Notwendigkeit und bedeutet keineswegs, dass andere (ethische) Probleme weniger wichtig wären. Ingenieurtechnische Anforderungen an die Robustheit des Designs autonomer Fahrsysteme sowie die psychologische Rolle, die Dilemma-Szenarien für die Akzeptanz der neuen Technologie spielen, machen es unumgänglich, Antworten für möglichst viele denkbare Szenarien zu finden. Diese repräsentieren stets Einzelfälle, die weder normierbar sind noch plausibel anhand von Heuristiken entschieden werden können.