Ethics, Law and AI | 2

Vanessa Schäffner

Unfallalgorithmen

Eine risikoethische Auseinandersetzung mit moralischen Dilemma-Strukturen im Kontext des autonomen Fahrens

VERLAG KARL ALBER

Ethics, Law and AI

Herausgegeben von

Carmine Di Martino (Università degli Studi di Milano)
Federico L.G. Faroldi (Università di Pavia)
Roberto Redaelli (Università degli Studi di Milano)

Band 2

Vanessa Schäffner

Unfallalgorithmen

Eine risikoethische Auseinandersetzung mit moralischen Dilemma-Strukturen im Kontext des autonomen Fahrens

VERLAG KARL ALBER

Die Forschungsarbeit wurde durch ein Promotionsstipendium im Rahmen des interdisziplinären, kooperativen Promotionskollegs „Ethik, Kultur und Bildung für das 21. Jahrhundert" von der Hanns-Seidel-Stiftung gefördert.

Sie entstand weiterhin im Verbundpromotionskolleg „Mobilität & Verkehr" des Bayerischen Wissenschaftsforums (BayWISS) und wurde vom Bayerischen Staatsministerium für Wissenschaft und Kultur gefördert.

Die Publikation als Open-Access-Werk wurde ermöglicht mit Unterstützung der Barbara-Wengeler-Stiftung.

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Zugl.: München, Hochschule für Philosophie, Diss., 2024

u.d.T.: Unfallalgorithmen in risikoethischer Perspektive. Zur Weiterentwicklung des Diskurses moralischer Dilemma-Strukturen im Kontext des autonomen Fahrens

1. Auflage 2024

Publiziert von
Verlag Karl Alber – ein Verlag in der
Nomos Verlagsgesellschaft mbH & Co. KG
Waldseestraße 3–5 | 76530 Baden-Baden
www.verlag-alber.de

Gesamtherstellung:
Nomos Verlagsgesellschaft mbH & Co. KG
Waldseestraße 3–5 | 76530 Baden-Baden

ISBN (Print): 978-3-495-99203-6
ISBN (ePDF): 978-3-495-99204-3

DOI: https://doi.org/10.5771/9783495992043

Onlineversion
Nomos eLibrary

Für meine Familie.

Vorwort

Dieses Buch stellt die gekürzte, redigierte und hinsichtlich ihrer praktischen Kontexteinbettung aktualisierte Fassung meiner Dissertation dar, die im Sommersemester 2024 von der Hochschule für Philosophie München, Philosophische Fakultät S.J. als Promotionsschrift angenommen wurde. Als interdisziplinäre Untersuchung aus dem Bereich der praktischen Philosophie ist die Forschungsarbeit thematisch im Grenzbereich von Technologie, Ökonomie, Gesellschaft und Ethik zu verorten – einer Kombination derjenigen Forschungsbereiche, die meinen wissenschaftlichen Werdegang nachhaltig geprägt haben. Die scheinbare Unaufhaltsamkeit technologischer Innovation steht in einem Spannungsverhältnis zur gesellschaftlichen Transformation des Zusammenlebens und dem, was in zukünftigen Gesellschaften als moralisch wünschenswert gelten kann. Dabei bestimmt das autonome Fahren seit einigen Jahren die inter- und transdisziplinären Diskurse in Forschung, Politik und Wirtschaft wie kaum eine andere derzeit in Entwicklung befindliche disruptive Technologie. Die ethische Auseinandersetzung mit den Implikationen und Herausforderungen einer der dynamischsten technologischen Revolutionen der Gegenwart scheint mir ein lohnendes Forschungsziel, das der nachstehenden philosophischen Abhandlung zugrunde liegt.

Der Weg zur Entstehung dieses Buches war durch persönliche und familiäre Ereignisse, vor allem aber durch viele verschiedene Menschen geprägt, die meine Arbeit begleitet haben. An dieser Stelle möchte ich meinen Dank jenen Personen aussprechen, ohne deren Unterstützung diese Forschungsarbeit nicht möglich gewesen wäre. Mein besonderer Dank gilt zuerst meinem Doktorvater Prof. Dr. Alexander Filipović – nicht nur für die fachliche Begleitung meiner Arbeit, zahllose Impulse und kompetente Ratschläge, sondern vor allem auch für seine Geduld, sein Verständnis und die immer unterstützende Form der Förderung und Betreuung. Während der intensiven Arbeit an diesem Buch waren mir die vielen Gespräche

mit ihm stets Ermutigung und Motivation zugleich. Für das mir entgegengebrachte Vertrauen und den Glauben an mich und mein Projekt über die gesamte Entstehungszeit hinweg bin ich sehr dankbar. Ferner danke ich Prof. Dr. Markus Babo in seiner Funktion als Zweitgutachter meiner Dissertation für seine Hilfsbereitschaft und beratende Unterstützung. Ebenfalls bedanken möchte ich mich bei Prof. Dr. Claus Dierksmeier, der nicht nur das Forschungsthema dieses Buches inspirierte, sondern mich auch seit Beginn meines Philosophiestudiums unterstützt und ermutigt hat, meinen Weg in der Philosophie zu finden und weiterzugehen.

Der wesentliche Teil der Abhandlung entstand zwischen Oktober 2018 und Dezember 2022 im Rahmen des interdisziplinären, kooperativen Promotionskollegs »Ethik, Kultur und Bildung für das 21. Jahrhundert«, das von der Kooperationspartnerschaft Katholischer Hochschulen in Bayern getragen wurde. Der fachliche und persönliche Austausch, der während dieses Zeitraums – und darüber hinaus – auf verschiedenen Ebenen mit den Kollegiatinnen und Kollegiaten, der Kollegleitung und dem professoralen Leitungsgremium stattfand, hat den Fortschritt meiner Arbeit sehr bereichert. Mein Dank gilt der Hanns-Seidel-Stiftung für die finanzielle und ideelle Förderung durch ein Promotionsstipendium, ohne das mir die Durchführung dieses Forschungsprojekts nicht möglich gewesen wäre. Ebenfalls bedanken möchte ich mich beim Bayerischen Wissenschaftsforum (BayWISS) für die mehrjährige Förderung meines Projekts im Rahmen des Verbundpromotionskollegs »Mobilität & Verkehr«, insbesondere für die Teilfinanzierung der Printausgabe dieses Buches. Zudem danke ich der Barbara-Wengeler-Stiftung für die zur Verfügung gestellten großzügigen Fördermittel, dank derer die Forschungsarbeit in ihrer elektronischen Form als frei zugängliches Open-Access-Werk erscheinen kann. Dem Verlag Karl Alber und der Nomos Verlagsgesellschaft danke ich für die Aufnahme in die Schriftenreihe und die professionelle, unkomplizierte und wertschätzende Betreuung während des Publikationsprozesses.

Einen besonderen persönlichen Dank widme ich nicht zuletzt meiner Familie, die mir während der intensiven Zeit der Arbeit an diesem Buch auf vielfältige Weise zur Seite gestanden hat. So danke ich zunächst meinen Eltern, Brigitte und Raimund, für ihre Wertschätzung und Unterstützung meines akademischen Weges, ihre praktische Hilfe in den zahlreichen Stunden, die sie dem Korrektorat

meines Manuskripts gewidmet haben – und dafür, dass sie immer an mich geglaubt haben. Tief verbunden und dankbar bin ich meinem Ehemann Benjamin für seine Rücksichtnahme und fortwährende Unterstützung in all den Jahren sowie für seinen unerschütterlichen Optimismus, mit dem er mich an den Tiefpunkten zum Weitermachen ermutigte und mir Hoffnung schenkte. Von ganzem Herzen danke ich schließlich meinen beiden Töchtern Sophia und Lea, die mir vor allem in schwierigen Phasen der intensiven Forschungsarbeit stets Quelle von Kraft, Zuversicht und Freude waren. Ihnen sei dieses Buch gewidmet.

In Bezug auf die im Rahmen der nachfolgenden Untersuchung verwendeten zentralen Begrifflichkeiten und verfolgten Ziele sind vorab einige klärende Anmerkungen hilfreich, um das Verständnis der dargestellten Sachverhalte, Thesen und Argumente zu erleichtern. Im Verlauf dieser Forschungsarbeit werden Unfallszenarien als unlösbare Dilemmata charakterisiert. Diese sind, wie der Name schon sagt, ›unlösbar‹ in dem Sinne, dass keine triviale, eindeutige Lösung für ihre spezifische Problematik existiert; unlösbare Dilemmata werden nicht gelöst, sie werden *entschieden*. Entsprechend ist das erklärte Ziel des Diskurses moralischer Unfalldilemmata nicht die Entwicklung von Lösungs-, sondern von *Entscheidungs*strategien. Mit ›Entscheidung‹ bzw. ›entscheiden‹ ist dabei das Resultat bzw. der Prozess einer ethischen Reflexion gemeint, die es erlaubt, unter Berücksichtigung aller moralisch relevanten Aspekte die im jeweiligen Einzelfall bestmögliche Antwort zu identifizieren und zu begründen, ohne den für die zurückgewiesene Alternative sprechenden Gründen ihre Geltung abzuerkennen.

Aufgrund der Tatsache, dass autonome Fahrzeuge prinzipiell durch Softwarealgorithmen gesteuert werden, entsteht im Hinblick auf den Entscheidungsbegriff in diesem Kontext ein zusätzlicher Klärungsbedarf, wenn angenommen wird, dass Maschinen nicht in einer dem Menschen ebenbürtigen Weise moralisch handlungsfähig sind. Es sei darauf hingewiesen, dass, sofern in der vorliegenden Arbeit von ›Handlungen‹, ›Entscheidungen‹ oder semantisch ähnlichen Ausdrücken die Rede ist, die im Kontext von autonomen Systemen verwendet werden, diese in einem metaphorischen Sinne zu verstehen sind. Sie beziehen sich nicht auf den philosophischen Handlungsbegriff und schließen explizit die Annahme aus, dass sie das Ergebnis eines kognitiven Prozesses sind, der menschlichem

Handeln und Entscheiden ebenbürtig ist. Die Verwendung des Begriffs ›Entscheidung‹ erfolgt vielmehr analog zu Miller et al. (2017, S. 390): »[…] we will use ›decision-making‹ to describe the following situation: an entity is in a situation, receives information about that situation, and selects and then implements a course of action.« Von ›Handlungen‹ wird im Sinne von maschinell gesteuerten Bewegungsabläufen gesprochen:

> […] the term ›action‹ is elliptical for something more technical: a robot's action is any movement that the robot causes that is not immediately caused by a human programmer or controller. […] When an autonomous car in ›autopilot mode‹ steers the wheel to stay in its lane or avoid a collision, this is the action of a robot. (Talbot et al., 2017, S. 259–260)

Die Arbeit demonstriert, dass Unfalldilemmata sich als unlösbare Konflikte zwischen legitimen individuellen Interessen darstellen, die Grundrechte der Einzelnen berühren. Die Auseinandersetzung erfolgt hier ausdrücklich auf ethischer Ebene; rechtebasierte Perspektiven finden nur stellenweise Erwähnung, um Argumente zu veranschaulichen oder zu vervollständigen. Der entwickelte alternative Problemzugang versteht sich als dezidiert ethischer Entwurf, der bestrebt ist, Impulse für eine politische Regulierung zu liefern, ohne sich selbst politischer Komponenten zu bedienen.

Darüber hinaus sind an dieser Stelle noch einige formale Hinweise vorauszuschicken, welche die Standards wissenschaftlichen Arbeitens erfordern. So sei erstens erwähnt, dass im Verlauf dieses Buches teilweise spätere Ausgaben zitierter, vor allem philosophischer Literatur verwendet werden. Gemäß des gewählten Zitationsstils geben die Kurzbelege im laufenden Text das Jahr der jeweiligen Erstveröffentlichung unabhängig von der verwendeten Ausgabe an, um eine Einordnung der zitierten Quellen in den jeweiligen historischen Kontext zu ermöglichen. Seitenangaben bei direkten Zitaten beziehen sich hingegen auf die jeweils verwendeten Ausgaben. Im Literaturverzeichnis werden sowohl das Erscheinungsjahr der verwendeten Ausgabe als auch dasjenige der Erstveröffentlichung ergänzend ausgewiesen. Eine Ausnahme bilden die Werke von Immanuel Kant; hier wird das Jahr der ursprünglichen Veröffentlichung jeweils separat per Fußnote vermerkt.

Zweitens wird explizit darauf hingewiesen, dass Teilergebnisse der nachfolgenden philosophischen Untersuchung sowohl in sinn-

gemäßer als auch wortgetreuer Form bereits in wissenschaftlichen Sammelbänden und thematisch einschlägigen Fachzeitschriften publiziert worden sind. Die Langbelege der entsprechenden Publikationen sind dem Literaturverzeichnis zu entnehmen.

Drittens ist zu beachten, dass dieses Buch auf die Verwendung von Genderstilen mit Sonderzeichen oder Doppelnennungen zugunsten einer besseren Lesbarkeit verzichtet. Wo immer es möglich ist, werden genderneutrale Formulierungen gewählt. Die verwendeten Personenbezeichnungen sind geschlechtsunspezifisch zu verstehen und beziehen sich – sofern nicht gesondert gekennzeichnet – auf alle Geschlechter (m/w/d/x).

Vanessa Schäffner

Ulm, im Dezember 2024

Inhaltsverzeichnis

1. Einführung in die Untersuchung 21

1.1 Problemaufriss: Autonomes Fahren als sozio-technisches Phänomen mit ethischer Dimension 21
1.2 Erkenntnisinteresse und Relevanz 25
1.3 Methodik und Struktur der Untersuchung 28
1.3.1 Methodischer Ansatz 28
1.3.2 Ziele und Hypothesen 30
1.3.3 Gedankengang 31

I. Autonomes Fahren und Unfalldilemmata: Ethischer Problemhorizont und Relevanz . . . 35

2. Das Phänomen ›Autonomes Fahren‹: Agenda, Ziele und Herausforderungen 37

2.1 Selbstfahrende Fahrzeuge als Treiber der Mobilitätswende 37
2.1.1 Der Autonomiebegriff im Kontext technischer Systeme 37
2.1.2 Motivatoren des autonomen Fahrens 39
2.1.3 Von Informanten über Assistenten zu Automaten: Evolution und Stufenmodell der Fahrautomatisierung 45
2.2 Herausforderungen im Kontext der Entwicklungsagenda 49
2.2.1 Die Wechselbeziehung zwischen technischer Reife und Wirtschaftlichkeit 49
2.2.2 Wo stehen wir heute? Aktueller technischer Stand und regulative Verordnungen 54
2.2.3 Zwischen Utopie und Dystopie: Die Ambivalenz des autonomen Fahrens 63

3. Unfallalgorithmen als ›moralischer Kompass‹ in ausweglosen Fahrsituationen 71

3.1 Ethische Problemstellungen und Diskurse im Überblick 71
3.1.1 Ethische Problemfelder im Kontext des autonomen Fahrens 71
3.1.2 Problemfeld Unfallsituationen: Der Verantwortungsdiskurs 75
3.1.3 Praktische Unvermeidbarkeit und dilemmatische Struktur ausweglosen Fahrsituationen . . . 81
3.2 Die Relevanz von Dilemma-Szenarien für das autonome Fahren 89
3.2.1 Möglichkeit und Existenz von Unfalldilemmata 89
3.2.2 Sind Unfallalgorithmen normierbar? 98
3.2.3 Gesellschaftliche und technische Relevanz von Dilemma-Szenarien 102
3.3 Zwischenergebnis: Die zentrale Bedeutung von Dilemma-Szenarien 106

II. Problemzugänge in zwei Diskursen: Darstellung und Kritik 107

4. Unfallalgorithmen als moralisches Designproblem: Rekonstruktion und Kritik 111

4.1 Entscheidungsalgorithmen, Dilemma-Szenarien und vermeintliche Trolley-Analogien 111
4.1.1 Die ethische Dimension von Entscheidungsalgorithmen 111
4.1.2 Maschinelle Moral, kontextsensitive Systeme und maschinelles Lernen 116
4.1.3 Systematisierung repräsentativer Dilemma-Szenarien und ihre moralphilosophische Problematisierung 126

4.1.4 Dilemma-Szenarien als angewandtes Trolley-Problem? Von Diskrepanzen und Disanalogien 131
4.2 Praktische Kontexteinbettung: Politisch-soziale Dimension und Entscheidungen unter Risiko 136
4.2.1 Die gesellschaftlich-soziale Dimension von Dilemma-Szenarien 136
4.2.2 Politische Regulierung: Unfallalgorithmen im Spannungsfeld zwischen individuellen Präferenzen und pluralistischen Wertvorstellungen 140
4.2.3 Epistemische Diskrepanzen: Sicherheit, Unsicherheit und Risiko im Kontext von Unfallszenarien 151
4.3 Deskriptive Ansätze: Perspektiven aus der Moralpsychologie 156
4.3.1 Moralische Präferenzen der Öffentlichkeit im Fokus einer experimentellen Ethik 156
4.3.2 Zur Relevanz deskriptiver Methoden: Eine Kritik 159
4.4 Normative Ansätze: Begründungsversuche der philosophischen Ethik 164
4.4.1 Klassische philosophische Ansätze zur moralischen Relevanz des Intervenierens 166
4.4.2 Utilitaristische Ansätze 170
4.4.3 Deontologische Ansätze 179
4.4.4 Alternative Ansätze und pluralistische Frameworks 186
4.5 Zwischenergebnis: Ungeklärte Fragen des Diskurses 199

5. Die Komplexität moralischer Dilemma-Strukturen: Rekonstruktion aus metaethischer Sicht 203

5.1 Einführung: Dilemmata als Grenzsituationen moralischen Handelns 203
5.1.1 Beispiele und Narrative aus Philosophie, Literatur und lebenspraktischen Kontexten 203

5.1.2 Kriterien und Definition moralischer Dilemma-Strukturen 207
5.2 Von der (Un-)Möglichkeit und (Nicht-)Existenz moralischer Dilemmata 216
5.2.1 Überblick und Einführung in den Diskurs . . . 216
5.2.2 Phänomenologische und konzeptionelle Perspektiven 218
5.2.3 (Vermeintliche) Inkonsistenzen in Theoriesystemen: Argumente der deontischen Logik und Thesen logischer Widersprüchlichkeit 229
5.3 Lösbarkeit, Inkommensurabilität und (Un-)Vergleichbarkeit in Wertekonflikten 234
5.3.1 Vorrangbeziehungen und *Prima-Facie*-Pflichten 234
5.3.2 Symmetrie versus Inkommensurabilität: Kriterien und Konzeptionen 239
5.3.3 Metaethische Konzepte unvermeidbaren Scheiterns: Von unersetzbaren Verlusten und nicht-verhandelbaren moralischen Werten . . 243
5.4 Anwendungsfall Unfalldilemmata: Interpretation aus metaethischer Sicht 247
5.4.1 Dilemmatische Unfallsituationen als Konflikte inkommensurabler Werte 247
5.4.2 Entscheidungsperspektiven für inkommensurable Wertekonflikte 250
5.4.3 Zwischenergebnis: Argumentative Relevanz der metaethischen Analyse 259

III. Risikoethische Auseinandersetzung: Entwurf eines alternativen Problemzugangs 261

6. Theoretische Grundlagen, begriffliche Reflexion und Ziele einer Risikoethik für Unfalldilemmata . . . 265

6.1 Systematische wissenschaftliche Einordnung der Risikoethik 265
6.1.1 Sozialwissenschaftlicher und sozio-technischer Diskurs 265
6.1.2 Von der Technikanalyse zur Technikbewertung: Technikfolgenabschätzung und technikethischer Diskurs 268
6.2 Risikoethische Grundlagen und Begriffe 274
6.2.1 Risikoethische Grundbegriffe: Unsicherheit, Ungewissheit und Risiko 274
6.2.2 Risiken im Handlungskontext: Risikosituationen und Risikokonstellationen 278
6.2.3 Grundfragen der Risikoethik: Zulässigkeit, Fairness und Verantwortung im Kontext von Risikoübertragungen 284
6.3 Grundzüge der (rationalen) Risikopraxis: Paradigmen und entscheidungstheoretische Ansätze 287
6.3.1 Risikopraktische Paradigmen 287
6.3.2 Entscheidungstheoretische Kriterien rationaler Risikopraxis 291
6.3.3 Zur Kritik traditioneller Risikopraxis 296

7. Unfallalgorithmen als risikoethisches Verteilungsproblem 299

7.1 Die (risiko-)ethische Problematisierung von Mobilitätsrisiken im Kontext autonomer Fahrsysteme 301
7.1.1 Autonome Fahrzeuge im Spannungsfeld zwischen soziologischer Risikoakzeptanz und ethischer Risikoakzeptabilität 301
7.1.2 Unfallalgorithmen und Risikoethik: Ansätze bisheriger (risikoethischer) Forschung 306

7.1.3 Gegenstand und Ziele eines alternativen risikoethischen Entwurfs 312
7.2 Analyse der Risikokonstellationen in Dilemma-Szenarien entlang von Kriterien der Risikoakzeptabilität 315
7.2.1 Akteure, Beziehungsnetzwerke und private Risiken 315
7.2.2 Szenarien der Risikoübertragung 319
7.2.3 Diskussion aus Sicht konsequentialistischer und kontraktualistischer Kriterien 323
7.2.4 Deontologische Risikoethik: Begründung, Ansätze und Konzeptionen 333
7.3 Grundzüge einer deontologischen Risikoethik für Unfallalgorithmen 336
7.3.1 Kohärente Risikopraxis nach Julian Nida-Rümelin: Grundlinien, Ziele und Anwendung 336
7.3.2 Die (absolute) Frage der Zumutbarkeit: Eine moralische Gratwanderung entlang von Risikoschwellen 341
7.3.3 Die (relative) Frage der Gerechtigkeit: Zwischen Reziprozität und Vorteilsausgleich . . . 351

8. Fazit und Ausblick 375

8.1 Ergebnisse der philosophischen Untersuchung: Zusammenfassung 375
8.2 Kritische Reflexion und Ausblick: Wissenschaftliche Relevanz, Forschungsdesiderate und Limitationen . . 384

Literaturverzeichnis 387

Facere docet philosophia, non dicere, et hoc exigit, ut ad legem suam quisque vivat, ne orationi vita dissentiat vel ipsa inter se vita; ut unus sit omnium actio[dissentio] num color [sit].

– **Seneca**, *Epistulae morales ad Lucilium*, Liber II, Epistula 20, 2

Was kann als Kompaß dienen? Die vorausgedachte Gefahr selber! In ihrem Wetterleuchten aus der Zukunft, im Vorschein ihres planetarischen Umfanges und ihres humanen Tiefganges, werden allererst die ethischen Prinzipien entdeckbar, aus denen sich die neuen Pflichten neuer Macht herleiten lassen.

– **Hans Jonas**, Vorwort zu *Das Prinzip Verantwortung* (1979)

1. Einführung in die Untersuchung

1.1 Problemaufriss: Autonomes Fahren als sozio-technisches Phänomen mit ethischer Dimension

Mobilität ist ein zentrales Merkmal des persönlichen und gesellschaftlichen Wohlstands moderner Industriegesellschaften. Sie spielt nicht nur eine gewichtige Rolle für den Warentransport und damit die ökonomische Prosperität einer Gesellschaft, sondern bestimmt auf Individualebene entscheidend die Lebensqualität der Einzelnen mit: Mobilität gilt als Inbegriff von (Bewegungs-)Freiheit und Unabhängigkeit, der Selbstbestimmung über das eigene Leben. Personenkraftwagen sind Teil unserer Geschichte und Kultur, ob zur Freizeitgestaltung, als Transportmittel oder Statussymbol (vgl. Floridi, 2019, S. 571–572). Der stetige gesellschaftliche Wandel, dem sowohl Wirtschaft als auch persönliche Lebensgestaltung unterworfen sind, wird seit Jahrzehnten durch drei transformative Kräfte dominiert: Urbanisierung, Individualisierung, Globalisierung. Um mit der Dynamik dieses Umbruchs Schritt halten zu können, wird eine Mobilitätswende längst als unverzichtbarer Baustein tragfähiger Zukunftsentwürfe angesehen. Sie soll inmitten des Spannungsfelds komplexer Herausforderungen den Erhalt bzw. die Steigerung der Mobilitätsqualität sicherstellen sowie den ständig steigenden Mobilitätsbedarf nachhaltig decken.

Auf gesättigten Märkten wie in Deutschland kann dies notwendigerweise nur durch den Ausgleich negativer Externalitäten der Individualmobilität gelingen, die sich auch als gesellschaftliche Kosten der Massenmobilisierung beschreiben lassen (vgl. Bengler et al., 2014, S. 2). Die Verbreitung städtischer Lebensformen und die damit verbundene kontinuierlich steigende Zahl an Megacitys führen zu einer Zunahme der Verkehrsdichte in Ballungsräumen. In der Folge kommt es zu mehr Staus, weniger Parkplätze stehen zur Verfügung, Unfallgefahr und Umweltbelastung steigen. Daraus resultierende Zeitverluste v. a. für Berufspendler drohen mittelfristig die in-

dividuelle und gesamtwirtschaftliche Effizienz stagnieren zu lassen. Die Orientierung an globalen und regionalen Klimazielen, der Einsatz ressourcenschonender Energieträger, mehr soziale Gerechtigkeit durch sozialverträgliche Berücksichtigung aller Bevölkerungsgruppen, effizientere Wegezeiten und nicht zuletzt ein entschleunigtes, qualitativ hochwertiges Mobilitätserlebnis – die Liste der Anforderungen an Mobilitätskonzepte der Zukunft, die vor allem eines sein sollen: nachhaltig, ist lang. Nicht zuletzt deshalb stellt die anvisierte Verkehrswende eine gesamtgesellschaftliche Aufgabe dar, die eine tiefgreifende Umgestaltung des gesamten Ökosystems der Mobilität bedeutet (vgl. Klima-Allianz Deutschland, 2020).

Die Entwicklung geeigneter Strategien für diese Herausforderungen steht weit oben auf der mittelfristigen Agenda von Politik, Forschung und Industrie. Neben der angestrebten Dekarbonisierung des Verkehrs durch alternative Antriebssysteme wie die Elektromobilität markiert die schrittweise Vernetzung und Automatisierung des Verkehrs einen globalen Megatrend in der Automobilindustrie. Seit Beginn des 21. Jahrhunderts befindet sich die Branche an der Schwelle zu einer neuen technologischen Revolution. Deren zentrales Element sind selbstfahrende Fahrzeuge, die von Fahrrobotern in einer vernetzten Verkehrsinfrastruktur gesteuert und daher auch als ›autonome Fahrzeuge‹ bezeichnet werden.

Während bisherige Innovationen innerhalb des Automobilsektors evolutionär im Sinne einer kontinuierlichen Weiterentwicklung bestehender (Assistenz-)Komponenten verliefen, wird das autonome Fahren von einem revolutionären Wandel der Individualmobilität begleitet (vgl. Beiker, 2015, S. 199–204). In Fahrzeugen verbaute digitale Technologien verändern das Wesen der Mobilität, indem sie das motorisierte Reisen sowohl in quantitativer als auch qualitativer Hinsicht transformieren; Floridi (2019, S. 569–571) spricht von einer »Re-Ontologisierung« der Mobilität. Die Automatisierung des Verkehrs ruft einen Paradigmenwechsel hervor, der das Potenzial besitzt, nicht nur unser Verständnis von Mobilität und unsere (motorisierten) Fortbewegungsroutinen, sondern auch die Gestaltung von Straßen und Städten als unseren Lebensräumen, und somit das gesamte Ökosystem des Fahrzeugverkehrs, tiefgreifend zu verändern (vgl. KPMG LLP, Center for Automotive Research, 2012, S. 24–26).

Die ideengeschichtliche Vision des automatisierten Fahrens als einer Fortbewegungsform, die sowohl maximalen Komfort als auch

Sicherheit verspricht, hat ihren Ursprung bereits in den Anfangstagen des Automobils. Schon im Jahr 1939 präsentierte der Autobauer General Motors auf der New Yorker Weltausstellung (damals utopische) fahrerlose Automobile als Teil der Zukunftsstadt *Futurama*, deren Realisierung nach damaliger Einschätzung in nicht allzu ferner Zukunft, ca. zwanzig Jahre später, erreicht sein würde (vgl. Gurney, 2015, S. 186). Dieses Projekt entpuppte sich jedoch bald als komplexer als erwartet. So blieb das selbstfahrende Fahrzeug lange Zeit lediglich in den Visionen von Futuristen und Science-Fiction-Liebhabern lebendig. Erst in den 1980er-Jahren rückte es schließlich wieder ins Blickfeld der Wissenschaft, als die Grundlagenforschung im Bereich der Fahrautomatisierung allmählich an Fahrt aufnahm (vgl. Anderson et al., 2016, S. 55–56). Heute ist aus dieser Vision längst eine der verheißungsvollsten Technologien in der Automobilbranche geworden. Die amtierende Bundesregierung proklamiert die Automatisierung des Verkehrs in ihrer im Sommer 2022 veröffentlichten Digitalstrategie als Wegweiser für einen digitalen Aufbruch im Bereich Mobilität:

> Digitale Vernetzung und Automatisierung unterstützen das Erreichen eines effizienten, sicheren, inklusiven und leistungsfähigen Mobilitätssystems, das sich flexibel dem Gesamtbedarf für Personen- und Gütertransport anpasst. Die Mobilität der Zukunft ist zunehmend digital. Sie schafft nutzerfreundliche, barrierefreie, intelligente und maßgeschneiderte Mobilitätsangebote, ermöglicht soziale und kulturelle Teilhabe und trägt zum Erreichen unserer Klimaschutz- und Nachhaltigkeitsziele bei. (Bundesministerium für Digitales und Verkehr, 2023, S. 19)

Auch über die Automobilindustrie hinaus erfährt das autonome Fahren als eines der bedeutendsten Anwendungsfelder von Künstlicher Intelligenz große Aufmerksamkeit (vgl. Hilgendorf, 2019, S. 356). Die technische Realisierung selbstfahrender Fahrzeuge stellt ein branchenübergreifendes und transdisziplinäres Entwicklungs- und Forschungsvorhaben dar, das die Expertise der Automobilindustrie zunächst mit Kompetenzen aus Kommunikations- und Informationstechnologie anreichert. Insbesondere der geplante Einsatz lernender Algorithmen ist ein Novum in der Geschichte der Automobilindustrie (vgl. Beiker, 2015, S. 202). Als Forschungsfeld bildet die Verwirklichung autonomer Fahrsysteme eine Schnittstelle verschiedener Forschungsdisziplinen: Neben der anspruchsvollen technischen Agenda mit ihren ökonomischen Abhängigkeiten bringt sie

auch komplexe Problematiken aus dem Gegenstandsbereich sozial- und geisteswissenschaftlicher Disziplinen mit sich. So ergeben sich u. a. praxisnahe rechtliche Fragen, z. B. hinsichtlich Zulassung und Haftung. Als sozio-technische Systeme entfalten selbstfahrende Autos gesellschaftliche Wirkungen, die durch den Umgang mit ihnen entstehen (vgl. Becker & Axhausen, 2017; Boeglin, 2015; Färber, 2015; Fraedrich & Lenz, 2014; Grunwald, 2015; Heinrichs, 2015; Kröger, 2015; Lenz & Fraedrich, 2015; Woisetschläger, 2015). Tiefgreifende Veränderungen unserer gesellschaftlich-sozialen Sphäre sind vorprogrammiert, denn jede Technikgestaltung ist zugleich immer auch Gesellschaftsgestaltung.

Ein thematisch breites, intensiv und kontrovers bearbeitetes Forschungsfeld eröffnen die bislang ungelösten ethischen Herausforderungen, die sich beim Einsatz selbstfahrender Fahrzeuge in real-lebensweltlichen Zusammenhängen ergeben. Sie reichen von Datenschutzproblematiken über Verantwortungsfragen bis hin zur Programmierung sogenannter Unfallalgorithmen, die in die Steuerungssysteme der Fahrzeuge implementiert werden. Bisher erarbeitete Direktiven und Richtlinien betonen, dass Strategien zur Unfallvermeidung das oberste Ziel des Designs autonomer Fahrzeuge ausmachen (vgl. Di Fabio et al., 2017; Europäische Kommission, 2020); auf diese Weise soll die Verkehrssicherheit signifikant erhöht werden. Jedoch muss davon ausgegangen werden, dass Unfälle sich nicht in allen Situationen vermeiden lassen, denn in »manchen Fällen kann eine Verkettung von Ereignissen zu einer Situation führen, die nicht ohne Personenschaden lösbar ist.« (Reschka, 2015, S. 507) Die Frage, wie autonome Fahrzeuge in Situationen agieren sollen, in denen eine Kollision nicht mehr abgewendet werden kann, ist genuin ethischer Natur. Bei der Wahl zwischen verfügbaren Trajektorien geraten unweigerlich die Interessen unterschiedlicher Personen in Konflikt; jedes autonome Fahrzeug kann – zwar selten, aber evident – in eine Lage geraten, »in der […] [es] vor der ›Entscheidung‹ steht, eines von zwei nicht abwägungsfähigen Übeln notwendig verwirklichen zu müssen.« (Di Fabio et al., 2017, S. 10)

Die Problematik rund um die Gestaltung von Unfallalgorithmen hat sich in den vergangenen Jahren als eine der zentralen Forschungsfragen des relevanten ethischen Diskurses etabliert. Welche ethischen Prinzipien sollen als Entscheidungsgrundlage in spezifischen Situationen dienen, in denen es unweigerlich zu Personen-

schäden kommt? Inwiefern lassen sich Abwägungen zwischen individuellen Interessen rechtfertigen? Antworten auf diese und ähnliche Fragestellungen müssen sich einerseits in einen bestehenden Rechtsrahmen einfügen und zugleich technisch umsetzbar sein. Andererseits hängen sie in entscheidender Weise von gesellschaftlich geprägten, normativen Moral- und Wertvorstellungen ab, welche angesichts der fortschreitenden Digitalisierung neu zur Diskussion gestellt werden müssen.

Gibt es im moralischen Sinne keine eindeutige Lösung, ist unklar, wie selbstfahrende Fahrzeuge in entsprechenden Fällen agieren sollen. Aus ethischer Sicht lässt sich das Entscheidungsproblem, das unabwendbaren Unfallsituationen zugrunde liegt, aufgrund seiner komplexen Struktur als kontextspezifische Instanz moralischer Dilemmata interpretieren. Letztere repräsentieren einen spezifischen Typ moralischer Entscheidungsprobleme, bei dem sich miteinander inkompatible Handlungsalternativen gegenüberstehen, die alle aus moralischen Gründen jeweils richtig und falsch zugleich sind: Richtig in dem Sinne, dass ein moralischer Grund die jeweilige Handlung einfordert, und zugleich falsch in dem Sinne, dass durch die Wahl einer Option zwangsläufig diejenigen moralischen Gebote vernachlässigt werden, die mit den Alternativen assoziiert sind. Sowohl in der philosophischen Ethik als auch in der Moralphilosophie haben moralische Dilemmata eine lange Tradition. In Anlehnung an diese wurde bislang häufig versucht, die Problematik von Unfalldilemmata im Kontext des autonomen Fahrens mithilfe traditioneller ethischer Denkmuster zu adressieren. Diese stoßen aufgrund der spezifischen Vielschichtigkeit des Anwendungsproblems allerdings an ihre Grenzen. Trotz des seit nunmehr zehn Jahren andauernden Forschungsdiskurses ist die zentrale Frage, wie sich Entscheidungsstrategien in konkreten dilemmatischen Fällen ethisch begründen lassen, noch immer weitgehend ungeklärt.

1.2 Erkenntnisinteresse und Relevanz

Seit dem Beschluss über das »Gesetz zum autonomen Fahren« im Mai 2021 schreitet die Entwicklung autonomer Fahrzeuge auch in Deutschland rasant voran. Auch wenn öffentlich kommunizierte Zeitpläne meist ambitioniert sind und laufend korrigiert werden,

so ist dennoch klar: Der Tag, an dem von Fahrrobotern gesteuerte Fahrzeuge – in welchem Umfang auch immer – unsere Straßen im Regelbetrieb bevölkern werden, rückt immer näher. Derzeit befinden wir uns in einer Übergangsphase, in der erste Hersteller die Zulassung für Fahrzeuge mit automatisierten Teilfunktionen erhalten haben, so beispielsweise Mercedes-Benz mit seinem Autobahn-Staupiloten »Drive Pilot« (vgl. Rudschies & Kroher, 2024) oder die nächste Generation des BMW-Autobahnpiloten mit integrierter Überholfunktion per Blicksteuerung, verbaut im neuen Modell »i5« (vgl. Geiger et al., 2024). Treiber der Entwicklung des autonomen Fahrens sind einerseits der technische Innovationsdruck, der auf den Schultern der Automobilbauer lastet, und anderseits die Neuverfassung der Rechtslage durch die gesetzgebenden Institutionen.

Vor diesem Hintergrund drängt angesichts der raschen Entwicklung der vernetzten und automatisierten Mobilität auch die Klärung ethischer Fragen. Moralische Dilemma-Szenarien zählen dabei zu den theoretisch und praktisch bedeutungsvollsten ethischen Herausforderungen. Eine vollumfängliche Automatisierung der Individualmobilität muss den Steuerungsalgorithmen autonomer Fahrsysteme idealerweise für jegliche denkbare Situation eine Handlungsempfehlung an die Hand geben. Besonders brisant aus ethischer Sicht ist diese Thematik nicht zuletzt deshalb, weil in einem automatisierten Verkehrsgeschehen zukünftig Algorithmen moralische Entscheidungen ›treffen‹ müssen, die bisher Menschen getroffen haben; aus instinktiven Handlungen werden systematische (vgl. Baker et al., 2018). In Extremfällen erfordern Unfalldilemmata Entscheidungen über Leben und Tod; ethisch relevant sind sie jedoch auch ohne die tragische Komponente, die Szenarien aus Tages- und Wochenzeitungen oft anhaftet. Dilemma-Szenarien pointieren ein Problem, das sich auch in alltäglichen Fahrsituationen stellt: Entscheidungen über Abstände, Geschwindigkeiten oder Trajektorien implizieren allesamt ein Abwägen der Interessen derjenigen, die von möglichen Handlungen in einer spezifischen Situation betroffen sind.

Die ethische Gestaltung von Unfallalgorithmen wurde im bisherigen Forschungsdiskurs mehrheitlich unter der zentralen Fragestellung diskutiert, an welchen ethischen Prinzipien sich die betreffenden Systemalgorithmen bei der Entscheidungsfindung orientieren sollen. Von Expertengremien entwickelte Richtlinien erwiesen sich bisher als zu unkonkret; sie sind lediglich als Empfehlungen zu

verstehen, denen es an Durchsetzungskraft mangelt. So wurden im Bericht der vom Bundesministerium für Verkehr und Digitale Infrastruktur (BMVI) eigens zur Reflexion ethischer Fragen in Bezug auf das autonome Fahren eingesetzten Kommission aus dem Jahr 2017 die Bandbreite ethischer Herausforderungen aufgezeigt sowie zentrale ethische Prinzipien festgelegt, z. B. ein Diskriminierungsverbot, der Schutz des Lebens vor anderen Erwägungen wie Mobilitätschancen oder eine Transparenz in Verantwortungsfragen (vgl. Lütge et al., 2020).[1] Was verbleibende offene Fragen anbelangt, wurde hingegen lediglich auf weiteren Forschungsbedarf verwiesen. Auch der nunmehr zehn Jahre währende internationale ethische Forschungsdiskurs wurde zwar – analog zur fortschreitenden technischen Entwicklung – mit der Zeit komplexer und vielschichtiger, konnte aber noch keine durchschlagskräftigen Strategien präsentieren. Dies ist vor allem darauf zurückzuführen, dass bisherige Forschungszugänge zu viele Fragen offenlassen, um als praktische Entscheidungshilfe in Erwägung gezogen zu werden.

In den letzten zwei bis drei Jahren kristallisierte sich eine neue, bis dato noch unterrepräsentierte Forschungsströmung heraus, die eine alternative Sichtweise auf die Problemstellung einnimmt: Dilemma-Szenarien werden dabei nicht länger als ein moralisches Designproblem betrachtet, das mittels traditioneller Ansätze der Moralphilosophie zu problematisieren ist, sondern vielmehr als risikoethisches Entscheidungsproblem im Hinblick auf eine rechtfertigbare Risikopraxis. Eine Auffassung als risikoethische Fragestellung erfreut sich zunehmender Aufmerksamkeit innerhalb des Forschungsfelds. Diese spiegelt sich nicht zuletzt in einer Vielzahl laufender oder kürzlich abgeschlossener Forschungsprojekte wider, die auf Bundesebene gefördert und zumeist in enger Verzahnung von technologischer Entwicklung durch Unternehmen oder unternehmensnahe Forschungsinstitutionen einerseits sowie Theoriebildung in universitären oder HAW-Instituten andererseits bearbeitet werden. Das kürzlich in Kraft getretene Europäische Gesetz zur Künstlichen Intelligenz (*EU Artificial Intelligence Act/AI Act*), demzufolge autonome Fahrsysteme

1 Auf den Bericht der Ethik-Kommission wird im Verlauf dieser Forschungsarbeit an verschiedenen Stellen zurückgegriffen. Auch wenn es sich dabei nicht um starke Argumente, sondern vielmehr um Richtlinien mit pragmatischem Fokus handelt, werden sie jedoch nicht ganz unbegründet in den Raum gestellt.

als Hochrisikoanwendungen eingestuft werden, rückt ebenfalls eine risikobasierte Betrachtungsweise in den Vordergrund.

Die Forschungsarbeit lässt sich im weiteren Kontext dieses neuen, risikofokussierten Forschungszugangs verorten. In Abgrenzung zu bisher publizierten Beiträgen wird in diesem Buch eine systematisch entwickelte, dezidiert risikoethische Untersuchung vorgelegt, auf deren Basis sich sowohl Dilemma-Szenarien als auch alltägliche Fahrsituationen ganzheitlich ethisch wie auch gesellschaftlich-sozial interpretieren lassen. Das Fehlen entsprechender integrativer Beiträge, die dies im gegebenen Anwendungskontext leisten, kann als ›blinder Fleck‹ eines Diskurses verstanden werden, der sich gerade neuformiert:

> The discussion of risk and AVs is just beginning. We're at the stage where (a) a good case has been made for the importance of the discussion, and where (b) a smattering of different scenarios and questions about risk has been posed. (Evans, 2022, S. 8)

Der erarbeitete risikoethische Entwurf versteht sich dabei als Versuch, anhand einer risikoethischen Neuinterpretation des zugrundeliegenden Entscheidungsproblems zukünftige Auseinandersetzungen mit Unfallalgorithmen zu inspirieren und auf diese Weise den Diskurs voranzubringen. Die Ergebnisse dieser Forschungsarbeit bewegen sich auf einer Mesoebene, d. h. sie stellen ausdrücklich keine unmittelbar implementierbaren, konkreten Entscheidungsstrategien bzw. Normen dar, sondern schlagen vielmehr Rahmenbedingungen in Form risikoethischer Grenzkriterien vor, die einer rechtfertigbaren Risikopraxis als Grundlage dienen können.

1.3 Methodik und Struktur der Untersuchung

1.3.1 Methodischer Ansatz

Als praktisches Entwicklungsprojekt und zugleich gesellschaftliches Phänomen befindet sich das autonome Fahren an der Schnittstelle verschiedener Forschungsdisziplinen. Obwohl die Problematik moralischer Dilemma-Situationen primär eine ethische Fragestellung darstellt, muss sie aufgrund der spezifischen Merkmale ihres Kontextes als inter- und transdisziplinäres Problem aufgefasst werden. So bezieht sie an verschiedenen Stellen auch Methoden und Konzepte

anderer wissenschaftlicher Disziplinen, z. B. der Ingenieurwissenschaften, Soziologie und Rechtswissenschaften, mit ein. Als philosophische Untersuchung einer real-lebensweltlichen Fragestellung ist das Forschungsvorhaben im Bereich der Angewandten Ethik anzusiedeln. Ihr methodisches Vorgehen entspricht dem Selbstverständnis einer Angewandten Ethik, die sich selbst stets in interdisziplinärer Perspektive begreift, wobei sie anwendungsnah, aber zugleich auch theoretisch-wissenschaftlich reflektierend und begründend vorgeht. Damit beschreitet sie einen methodischen Pfad, den Filipović (2016, S. 46) als »Mittelweg zwischen ethischer Theoriebildung und erfahrungsbezogener Normfindung« beschreibt. In dieser Arbeit wird beabsichtigt, das Anwendungsproblem in seinen ethisch relevanten Facetten aufzugreifen und eine Antwort zu entwickeln, die auf der Basis einer kritischen Reflexion bestehende Entscheidungsstrategien zurückweist bzw. in begründeter Weise ergänzt.

Einer der originären Beiträge, welche dieses Buch zum Forschungsdiskurs leistet, besteht in der metaethischen Rekonstruktion der spezifischen Problemstruktur, die Unfalldilemmata kennzeichnet. Dabei bleiben metaethische Methoden jedoch auf den Rahmen des fünften Kapitels beschränkt; ihre Funktion ist es, die Problemstellung aus der Sichtweise des metaethischen Diskurses zu beleuchten und das zentrale Argument zu erweitern, welches sich aus der kritischen Analyse bisheriger Forschungszugänge ergibt.

Die interdisziplinäre Dimension des Anwendungsproblems bringt direkte Implikationen für die wissenschaftliche Standortbestimmung dieser philosophischen Abhandlung mit sich. Sie steht im weiteren Sinne im Kontext jenes technikethischen Diskurses, der sich mit solchen algorithmischen Entscheidungen befasst, die eine moralische Dimension aufweisen. Im engeren Sinne ist sie an der Schnittstelle von Maschinenethik, Digitaler Ethik und Ethik der Künstlichen Intelligenz zu verorten. Durch die Ausarbeitung einer risikoethischen Perspektive auf das Anwendungsproblem ist sie vor allem auch in der Risikoethik anschlussfähig. Diese widmet sich als Teilgebiet der Ethik denjenigen Problemstellungen, die im Zusammenhang mit der moralischen Bewertung von unsicheren und risikobehafteten Handlungen im Kontext gesellschaftlicher Risiken anzusiedeln sind. Dies schließt technikinduzierte Risiken ausdrücklich ein. Während beispielsweise die Maschinenethik danach fragt, inwiefern Entscheidungen über Leben und Tod von Maschinen getroffen werden kön-

nen bzw. sollen, verfolgt die Risikoethik das Ziel, rechtfertigbare risikopraktische Kriterien und Rahmenbedingungen für den Einsatz künstlicher Systeme zu definieren. Die forcierte Integration verschiedener disziplinspezifischer ethischer Zugänge ist ein Charakteristikum dieses Buches, das Elemente der Digitalen Ethik, der Metaethik und der Risikoethik zusammenführt.

1.3.2 Ziele und Hypothesen

Ein *erstes Teilziel* der vorliegenden Untersuchung ist es, die Problemstellung innerhalb des Forschungsdiskurses zu verorten und dabei die Bedeutung und Relevanz von Unfallszenarien sowohl in praktischer als auch theoretischer Hinsicht zu evaluieren (*Zwischenergebnis*). Das *zweite Teilziel* besteht in einer kritischen Reflexion moralphilosophischer Problemzugänge, die Unfallalgorithmen als moralisches Designproblem interpretieren und den Forschungsdiskurs seit seinen Anfängen dominieren. Eine differenzierte Auseinandersetzung legt offen, dass unter dieser Problemperspektive zentrale Fragen ungeklärt bleiben; diese stehen der finalen Begründbarkeit eines moralphilosophisch fundierten Designs von Unfallalgorithmen entgegen und beinhalten zugleich die Forderung nach alternativen Zugängen (*zweites Zwischenergebnis*). Anhand des *dritten Teilziels* wird die Absicht verfolgt, mithilfe einer metaethischen Rekonstruktion der spezifischen Wertekonflikte, die den relevanten moralischen Dilemma-Strukturen zugrunde liegen, eine ganzheitliche Perspektive auf die Problemstellung zu eröffnen. Auf diese Weise wird die Notwendigkeit einer pragmatischen, der praktischen Sache dienlichen Ausrichtung möglicher Strategien vor dem Anwendungshintergrund von Unfalldilemmata begründet (*drittes Zwischenergebnis*). Schließlich wird im Rahmen des *vierten Teilziels* ein alternativer Problemzugang entworfen, der eine risikoethische Perspektive auf die spezifische Problematik von Unfallalgorithmen einnimmt und durch eine pragmatische Herangehensweise an zum jetzigen Zeitpunkt ungeklärte ethische Fragen den Forschungsdiskurs bereichert.

Die Ziele der Abhandlung werden in Form zweier forschungsleitender Arbeitshypothesen konkretisiert. Die Bearbeitung des zweiten Teilziels erfolgt im Hinblick auf die *(erste) These*, dass der bisher dominante moralphilosophische Zugang zur Problemstellung aufgrund methodischer sowie inhaltlicher und problemstruktureller Schwä-

chen zentrale Fragen offenlässt. Er ist daher inadäquat, um glaubwürdige und begründbare Entscheidungsstrategien hinsichtlich des Anwendungsproblems zu etablieren. Im Zuge des dritten Teilziels wird die Plausibilität dieser These gestützt. Anhand einer Analyse der metaethischen Grundlagen dilemmatischer Wertekonflikte wird das Fehlen systematischer Strategien zur Entscheidung von Unfalldilemmata aufgezeigt und gleichzeitig das vielversprechende Potenzial pragmatischer Entscheidungsstrategien akzentuiert. Im Anschluss an die erste These wird im Rahmen des vierten Teilziels eine risikoethische Auseinandersetzung vorgelegt. Dieser liegt die *(zweite) These* zugrunde, dass sich unter einem risikoethischen Zugang zentrale Fragen des Anwendungsproblems klären lassen und normative Implikationen freigelegt werden können, die neuen, vielversprechenden Entscheidungsperspektiven den Weg bereiten.

1.3.3 Gedankengang

Der Argumentations- und Gedankengang dieses Buches folgt einem Narrativ, das sich in drei Teile gliedert und dabei an den forschungsleitenden Thesen orientiert.

Teil I: Autonomes Fahren und Unfalldilemmata: Ethischer Problemhorizont und Relevanz

Die Untersuchung beginnt mit einer einführenden Darstellung des autonomen Fahrens sowohl als technologisches Entwicklungsprojekt als auch gesellschaftliches Phänomen. Hinsichtlich der Funktion der Verkehrsautomatisierung als eines zentralen Elementes der anvisierten Mobilitätswende westlicher Gesellschaften werden in Kap. 2 Evolution, Agenda, Ziele und Herausforderungen der Entwicklung autonomer Fahrzeuge knapp skizziert und einer kritischen Betrachtung unterzogen. Ziel ist es hierbei, den praktischen Forschungsgegenstand in der Bandbreite seiner vielschichtigen Facetten darzustellen und die Problematiken anzudeuten, die sich für verschiedene Forschungsdisziplinen ergeben. In Kap. 3 werden sodann ethische Herausforderungen fokussiert, die sich im Kontext von Entwicklung und Einsatz autonomer Fahrsysteme stellen. Neben einer überblicksartigen Darstellung verschiedener ethischer Diskurse wird das Hauptaugenmerk auf jene Unfallalgorithmen gelenkt, die im Fall unvermeidbarer Kollisionen aktiviert werden und in dabei entste-

henden moralischen Entscheidungsdilemmata die Aktionen autonomer Fahrzeuge steuern. Diese Problemstellung, die im Zentrum der Forschungsarbeit steht, wird sowohl in ihrer ethischen als auch informationstechnischen Dimension thematisiert. Schließlich werden Bedeutung und Relevanz moralischer Dilemma-Szenarien für das autonome Fahren erörtert.

Teil II: Problemzugänge in zwei Diskursen: Darstellung und Kritik

Im zweiten Teil wird das zentrale Argument entwickelt, welches die erste These des Vorhabens begründet: Ein Zugang zur Gestaltung von Unfallalgorithmen, wie er bisher im einschlägigen Forschungsdiskurs diskutiert wurde, lässt (zu) viele Fragen offen. Hier erfolgt eine kritische Auseinandersetzung mit dem Fokus bisheriger Forschung, unter dem moralische Dilemma-Situationen im Kontext autonomer Fahrsysteme beinahe ausschließlich als Problematik moralischer Designentscheidungen betrachtet werden. Anhand einer umfassenden Rekonstruktion der relevanten Forschungsliteratur wird in Kap. 4 der Nachweis geführt, dass der bisher dominante Forschungszugang sowohl in methodischer als auch inhaltlicher und struktureller Hinsicht erhebliche Schwächen an den Tag legt. Im Rahmen von Kap. 5 wird dieses Argument um eine metaethische Perspektive ergänzt, mittels derer die zugrundeliegende Problemstellung moralischer Dilemmata aus metaethischer Sicht erörtert und hinsichtlich des Anwendungskontextes von Unfalldilemmata spezifiziert wird. Es wird substanziiert, weshalb pragmatische Strategien zur Bewältigung echter moralischer Dilemmata besonders vielversprechend sind.

Teil III: Risikoethische Auseinandersetzung: Entwurf eines alternativen Problemzugangs

Im dritten Teil wird auf Basis der bisherigen Ergebnisse schließlich ein alternativer Zugang zum Anwendungsproblem entworfen. Dabei wird das Feld der Risikoethik beschritten, die – wie die zweite These der Forschungsarbeit postuliert – adäquate Mittel bereitstellt, um sowohl die Spezifika des praktischen Problemkontextes als auch deren theoretische Dimension im Sinne einer pragmatischen Herangehensweise zu integrieren. In Kap. 6 werden im Rahmen einer theoretischen Grundlegung und Einführung in wissenschaftliche Verortung, Historie, Gegenstandsbereich und Paradigmen der Risikoethik

zunächst begriffliche und konzeptionelle Grundlagen gelegt. Es wird gezeigt, dass eine rationale Risikopraxis angesichts der spezifischen Problematik von Unfallalgorithmen nicht glaubwürdig ist. In Kap. 7 werden sodann die Grundzüge einer deontologischen Risikoethik systematisch entwickelt. Im Zuge einer eingehenden risikoethischen Analyse wird die im Diskurs bisher dominante Frage nach der moralphilosophischen Begründung von Entscheidungsprinzipien für Unfallalgorithmen zur Frage der ethischen Rechtfertigbarkeit von Dilemma-Risiken sowie der fairen Verteilung daraus resultierender Vor- und Nachteile transformiert. Es wird für eine kohärente Risikopraxis plädiert, welche die Gestaltung von Unfallalgorithmen als Optimierungsproblem begreift, dem durch deontologische Grenzkriterien unverhandelbare Beschränkungen auferlegt sind.

Zum Abschluss werden die zentralen Ergebnisse der Untersuchung in Kap. 8 nochmals zusammengefasst und im Hinblick auf ihre Bedeutung für weiterführende Forschung kritisch evaluiert.

I.
Autonomes Fahren und Unfalldilemmata: Ethischer Problemhorizont und Relevanz

Der erste Teil des Buches widmet sich einer Einführung in das Phänomen des autonomen Fahrens, dessen überblicksartiger Darstellung und Problematisierung sowie der Verortung der zentralen Fragestellung innerhalb des Forschungsdiskurses, zu dessen Weiterentwicklung die Forschungsarbeit beitragen will. In zwei Kapiteln wird auf diese Weise das erste Teilziel der vorliegenden Untersuchung erarbeitet. Besonderes Augenmerk wird dabei auf die Evaluation der Bedeutung und Relevanz von Unfallszenarien sowohl in praktischer als auch theoretischer Hinsicht gelegt.

Als Annäherung an das Anwendungsproblem werden in Kap. 2 Agenda, Ziele und Herausforderungen im weiteren Kontext autonomer Fahrsysteme beschrieben. In Kap. 2.1 werden selbstfahrende Fahrzeuge als Treiber der Mobilitätswende in ihrer technischen, wirtschaftlichen und gesellschaftlichen Dimension konturiert. Ausgehend von einer begrifflichen Reflexion über den zugrundeliegenden Autonomiebegriff werden Motivatoren, Vision und evolutionäre Aspekte der Fahrautomatisierung erläutert. Darauf aufbauend werden in Kap. 2.2 die spezifischen Herausforderungen dargestellt, die

durch Zusammenhänge und Wechselwirkungen zwischen den vielfältigen Anforderungen an die neue Technologie einerseits und ambivalente Wirkungen derselben andererseits entstehen. Im Anschluss an eine Bestandsaufnahme des gegenwärtigen technischen und regulatorischen Entwicklungsstands wird schließlich kritisch reflektiert, inwiefern das autonome Fahren die an seine Vision geknüpften Erwartungen erfüllen kann.

In Kap. 3 wird der ethische Diskurs des autonomen Fahrens in systematischer Weise rekonstruiert sowie die Bedeutung aufgezeigt, die Unfalldilemmata für theoretische und praktische Fragen des Anwendungskontextes besitzen. Zunächst erfolgt in Kap 3.1 eine überblicksartige Skizze ethischer Problemfelder und Diskurse, in deren Rahmen die relevanten Literaturströmungen identifiziert, gegeneinander abgegrenzt und in ihrer thematischen Tiefe erläutert werden. Besondere Aufmerksamkeit erfährt dabei die Problematik unvermeidbarer Unfallsituationen. In Kap. 3.2 wird vertiefend auf die Relevanz von Dilemma-Szenarien für die Gestaltung von Unfallalgorithmen eingegangen, welche das Fahrverhalten autonomer Fahrzeuge im Fall einer unabwendbaren Kollision steuern. Anhand der relevanten Forschungsliteratur wird ausführlich argumentiert, dass Entscheidungsprobleme mit dilemmatischen Strukturen im Kontext autonomer Fahrsysteme sowohl theoretisch möglich als auch praktisch existent sind. Zudem wird begründet, weshalb Unfallalgorithmen nicht normierbar sind und welche gesellschaftliche und technische Relevanz sie aufgrund dessen entfalten. In Kap. 3.3 werden schließlich die Ergebnisse von Teil I in einem ersten Zwischenergebnis zusammengefasst.

2. Das Phänomen ›Autonomes Fahren‹: Agenda, Ziele und Herausforderungen

2.1 Selbstfahrende Fahrzeuge als Treiber der Mobilitätswende

2.1.1 Der Autonomiebegriff im Kontext technischer Systeme

Selbstfahrende Fahrzeuge fungieren als die zentralen Treiber der anvisierten Mobilitätswende. Gemäß ihrer Vision entbinden sie Menschen von deren Fahraufgaben und schaffen Potenziale für mehr Sicherheit, Effizienz und Teilhabe. Auf diese Weise transformieren sie das Wesen der Mobilität, wie wir sie heute kennen. Die revolutionäre Dynamik der neuen Technologie liegt in einer ihrer zentralen Eigenschaften begründet, die ihr auch den Namen gibt: ihrer Autonomie. Der Begriff der Autonomie stammt aus dem Altgriechischen (*αὐτονομία* bzw. *autonomía*), wo er so viel wie ›Eigengesetzlichkeit‹ bedeutet. Im Zuge einer Auseinandersetzung mit komplexen praktischen Fragen wird allerdings schnell deutlich, dass der Autonomiebegriff im Kontext verschiedener Forschungsdisziplinen jeweils unterschiedliche, fachspezifische Bedeutungen aufweist. Es erscheint daher zweckmäßig, eine wissenschaftliche Abhandlung über autonome Fahrzeuge mit einer Klärung des Autonomieverständnisses zu beginnen, das hier zugrunde liegt.

Die Bezeichnung ›autonomes Fahrzeug‹ als Anwendungsbeispiel eines autonomen Systems bezieht sich auf den in der Informationstechnik verwendeten Autonomiebegriff. Aus anwendungsbezogener und operationeller Sicht zeichnen sich autonome Systeme zunächst dadurch aus, dass sie unabhängig von direktem menschlichen Eingreifen operieren können. Eine solche triviale Auffassung von Autonomie ist jedoch nur für vollständig kontrollierbare Umgebungen, beispielsweise eine intelligente digitale Fabrik (*smart factory*), plausibel. Sobald das Umfeld komplexer wird, muss der Begriff erweitert werden. Autonomie ist dann als die Fähigkeit eines Systems zu interpretieren, das ohne menschliches Eingreifen eine Handlungsoption

wählt: »When such machines are called ›autonomous‹, it is meant that they are able to choose by themselves, without human intervention, the appropriate course of action in the manifold situations they encounter.« (Totschnig, 2020, S. 2474) Diese Verwendungsweise des Autonomiebegriffs setzt voraus, dass der Agent sich bei seiner Wahl an einem Ziel oder einer Nutzenfunktion orientiert, die extern festgelegt ist. In eine ähnliche Richtung weist auch die Definition von Hilgendorf (2017b, S. 47): »Unter einem autonomen technischen System soll ein System verstanden werden, das auf Probleme unabhängig von menschlichem Input situationsangemessen und somit ›intelligent‹ reagieren kann.« Bradshaw et al. (2013) betonen, dass das Design technischer Systeme in der Regel auf einen bestimmten Kompetenzbereich limitiert ist. Die Europäische Gruppe für Ethik der Naturwissenschaften und der neuen Technologien erläutert in ihrer »Erklärung zu künstlicher Intelligenz, Robotik und ›autonomen‹ Systemen« (2018, S. 10–11):

> Allerdings hat sich der Begriff des ›autonomen‹ Systems zur Bezeichnung eines Höchstmaßes an Automatisierung und maximaler Unabhängigkeit vom Menschen im Sinne einer operativen und entscheidungsbezogenen ›Autonomie‹ in der wissenschaftlichen Literatur und der öffentlichen Debatte sehr stark durchgesetzt.

Eine stärker technisch orientierte Konzeption definiert Autonomie als »die Fähigkeit eines rechnergestützten Systems, selbstständig aus Daten, die sowohl durch Programmierung vorgegeben als auch aus der Umwelt mittels Sensoren gewonnen sein können, zielgerichtete Pläne und Aktionen zu generieren.« (Deutscher Bundestag, 2020, S. 31) Eine solche Perspektive auf das Konzept der Autonomie setzt vor allem zwei zentrale Designelemente voraus. Dies ist einerseits die Implementierung einer Belohnungs- oder Nutzenfunktion, mittels derer das System seine Aktionen hinsichtlich gewünschter Zustände abgleichen kann. Andererseits sind Mechanismen des maschinellen Lernens unverzichtbar, die es dem System ermöglichen, sich an veränderte Umgebungsbedingungen anzupassen und sein Verhalten aufgrund von bisherigen, als unerwünscht bewerteten Aktionen zu optimieren.

Autonomie tritt grundsätzlich in verschiedenen Ausprägungen entlang eines Kontinuums auf (vgl. Etzioni & Etzioni, 2017, S. 408–409). Eine einflussreiche Auffassung, die sich an der Schnittstelle von informationstechnischer und (maschinen-)ethischer Perspekti-

ve bewegt, liefert der deskriptive Ansatz von Floridi und Sanders (2004, S. 357–364). Sie schreiben künstlichen Systemen genau dann eine Selbstursprünglichkeit ihrer Handlungen zu, wenn diese drei relevante Kriterien erfüllen. Neben der Fähigkeit zur Interaktivität (*interactivity*) und Adaption (*adaptability*) besteht eines dieser Kriterien in der Autonomie (*autonomy*):

> Autonomy means that the agent is able to change state without direct response to interaction: it can perform internal transitions to change its state. So an agent must have at least two states. This property imbues an agent with a certain degree of complexity and independence from its environment. (Ebd., S. 357)

Die im Kontext künstlicher Systeme verwendeten Autonomiekonzeptionen unterscheiden sich deutlich von anspruchsvolleren philosophischen Auffassungen. Autonomie im Sinne Immanuel Kants, der sie als »Beschaffenheit des Willens, dadurch derselbe ihm selbst (unabhängig von aller Beschaffenheit der Gegenstände des Wollens) ein Gesetz ist« (1900ff., GMS, AA 04: 440.16-18)[2], interpretiert, wäre ausschließlich im Rahmen der Realisierung starker Künstlicher Intelligenz denkbar. Die Erschaffung einer menschenebenbürtigen Denkfähigkeit, die die Ziele ihres Handelns selbst definiert, ist zum gegenwärtigen Zeitpunkt allerdings lediglich eine Zukunftsvision; bei derzeitigen Systemen sind die relevanten Handlungsziele extern festgelegt.

Analog zur Vielschichtigkeit des ihnen zugrundeliegenden Autonomiebegriffs stellen autonome Systeme ein großes Forschungs- und Entwicklungsvorhaben an der Schnittstelle verschiedener Ingenieur-, Sozial- und Geisteswissenschaften dar. Die Konzeption autonomer Fahrzeuge ist dabei kein Selbstzweck, sondern durch spezifische praktische Erwartungen motiviert. Diese werden im nachfolgenden Unterkapitel skizziert.

2.1.2 Motivatoren des autonomen Fahrens

Die Vision einer automatisierten Mobilität ist verbunden mit der Erwartung, auf diese Weise die sozialen Kosten motorisierter Individualmobilität entscheidend zu senken. Die Idee des autonomen

2 Kants *Grundlegung zur Metaphysik der Sitten* wurde erstmals 1785 veröffentlicht.

Fahrens begegnet den essenziellen Herausforderungen moderner Industriegesellschaften anhand dreier zentraler Kriterien, die die Mobilitätsqualität maßgeblich beeinflussen (vgl. Beiker, 2012, S. 1149–1152).

1. Sicherheitspotenzial

Primärer Motivator der Verkehrsautomatisierung ist ihr Potenzial, durch die Reduzierung bzw. Eliminierung menschlichen Fahrversagens mittel- bis langfristig die Zahl der Verkehrsunfälle signifikant zu reduzieren (vgl. Beiker, 2012, S. 1149–1150; Crew, 2015, o. S.; Fleetwood, 2017, S. 532). Wie ältere Studien belegen, gilt der ›Faktor Mensch‹ in der Forschung zur Verkehrssicherheit schon seit Langem als primäre Risikoquelle. Veränderungen im Fahrverhalten bieten dementsprechend eine vielversprechende Möglichkeit, um Schäden im Kontext von Verkehrsunfällen zu reduzieren: »Human factors are far more important than engineering factors. Among human factors, driver behavior (what the driver *chooses to do*) has much greater influence on safety than driver performance (what the driver *can do*).« (Evans, 1996, S. 784) Gemäß Datenerfassung des Statistischen Bundesamtes waren 88 % der Unfälle mit Personenschäden[3] im Jahr 2021 auf menschliche Fahrfehler zurückzuführen. Die häufigsten Fehler geschahen dabei beim Abbiegen, Wenden, Rückwärtsfahren, Ein- bzw. Anfahren, bei Nichtbeachtung der Vorfahrtsregeln, aufgrund von ungenügendem Abstand oder nicht angepasster Geschwindigkeit (vgl. Statistisches Bundesamt, 2022, S. 50). Analog nennt eine in den USA durchgeführte Studie inkorrekt erkannte Fahrsituationen aufgrund von Unaufmerksamkeit oder Ablenkung als häufigste Unfallursache, gefolgt von schlechten Fahrentscheidungen wie zu hoher Geschwindigkeit, Fehleinschätzung anderer Verkehrsteilnehmer, aggressive Fahrweise und schließlich Fehlern im Fahrverhalten wie Überkompensation oder mangelnde Richtungskontrolle (vgl. National Highway Traffic Safety Administration, 2008, S. 24–25). Auch der Konsum berauschender Mittel spielt eine

3 Bei Unfällen mit Personenschäden handelt es sich um solche Verkehrsereignisse, bei denen Personen verletzt oder getötet werden, wobei die Höhe des Sachschadens irrelevant ist (vgl. Statistisches Bundesamt, 2023).

nicht unerhebliche Rolle.[4] Die Ergebnisse einer neueren amerikanischen Studie decken sich mit diesen Erkenntnissen: Selbstfahrende Fahrzeuge sind insbesondere auf Schnellstraßen und bei Nebel sicherer als von Menschen gesteuerte Autos, nicht hingegen auf Landstraßen, bei Dämmerung und bei Abbiegemanövern (vgl. Abdel-Aty & Ding, 2024; Beck, 2024).

Die Idee des autonomen Fahrens setzt an ebendieser Fehleranfälligkeit menschlichen Fahrverhaltens an, um diejenigen zu schützen, die selbst Quelle hoher Risiken sind:

> It is the behavior of those whose lives are at stake in traffic that most influences risk in traffic. The least safe vehicle driven on the least safe road by some drivers poses far less risk than the safest vehicle driven on the safest road by other drivers. (Evans, 2008, S. 1)

Würden automatisierte bzw. autonome Systeme im Sinne eines defensiven Fahrstils programmiert, sodass sie die Fahrzeugumgebung ständig überwachen, im Bedarfsfall schnell reagieren können und dabei stets die Verkehrsregeln beachten, ließen sich fahrerbezogene Unfallursachen potenziell minimieren bzw. im Fall vollständig autonomer Fahrzeuge sogar gänzlich eliminieren (vgl. Goodall, 2020, S. 1). Auch wenn sich eine genaue Bezifferung des positiven Effekts[5] derzeit noch nicht ausreichend mit harten Fakten untermauern lässt, so geben statistische Auswertungen jedoch Anlass zu der Annahme, dass ein solcher – zumindest in gewissem Maße – eintreten könnte. Die Zahlen polizeilich erfasster Unfälle aller Schweregrade sind in den letzten Jahren tendenziell rückläufig (vgl. Statistisches Bundesamt, 2022, S. 44), wobei ein nicht unerheblicher Anteil dieses Trends dem Einsatz automatisierter Teilfunktionen im Fahrbetrieb zugeschrieben werden kann (vgl. Anderson et al., 2016, S. 14–16).[6]

---

4 Der Anteil unter Alkoholeinfluss begangener Fahrfehler an der Gesamtzahl der Unfälle mit Personenschäden lag im Jahr 2021 in Deutschland zwar nur bei 3,3 % (vgl. Statistisches Bundesamt, 2022, S. 50). Schlüsselt man allerdings die Ursachen von Unfällen mit tödlichem Ausgang weiter auf, so zeichnete Alkohol- oder Drogenkonsum beispielsweise im Jahr 2011 für mehr als 39 % der Fälle in den USA verantwortlich (vgl. Anderson et al., 2016, S. 16).

5 Neben einer Reduzierung der Personenschäden hat eine erhöhte Verkehrssicherheit auch wirtschaftliche Auswirkungen, beispielsweise indem die Kosten für Versicherung, Reparaturen und Administration wegfallen bzw. sinken.

6 In einer kürzlich veröffentlichten Studie dokumentieren Abdel-Aty und Ding (2024), dass mit Fahrerassistenzsystemen operierende Autos in vielen Szenarien

Die Ergebnisse einer vergleichenden Studie der Unfallraten von konventionellen Fahrzeugen und solchen, die im autonomen Modus operieren, legen auf Basis der aktuellen Datenlage nahe, dass autonome Fahrzeuge in weniger Unfälle aller Schweregrade involviert sein könnten (vgl. Blanco et al., 2016, S. i–iv). Während frühe Testfahrten diese These noch zu stützen schienen,[7] sorgten in den letzten Jahren vermehrt tragische Unfälle, welche die bis dato technische Unreife der eingesetzten Fahrsysteme schonungslos aufzeigen, für öffentliches Aufsehen.[8] Inwiefern autonome Fahrzeuge tatsächlich zu mehr Verkehrssicherheit beitragen, wird an anderer Stelle in diesem Buch kritisch diskutiert (siehe Kap. 2.2.3). Hier soll das (vermeintliche) Sicherheitspotenzial lediglich in seiner Funktion als primärer Motivator des autonomen Fahrens verstanden werden.

2. Effizienzsteigerung

Positive Wirkungen des autonomen Fahrens erhofft man sich auch in Sachen Zeit- und Verkehrseffizienz. Neben einer allgemeinen Steigerung der Lebensqualität für Personengruppen, die häufig im Verkehr unterwegs sind, hätte dies vor allem vorteilhafte ökonomische Auswirkungen. Zum einen würden die Beförderten durch die (vollständige) Automatisierung der Fahraufgabe von (aktiven) Fahrern zu (passiven) Passagieren, die ihre Zeit an Bord freier gestalten können. So würde es etwa Berufspendlern möglich, Transferzeiten produktiv als Arbeitszeit zu nutzen (vgl. Brändle & Grunwald, 2019, S. 282; Gurney, 2015, S. 192–193; KPMG LLP, Center for Automotive Research, 2012, S. 29). Zum anderen verspricht die Vernetzung und zentrale Koordination der Fahrzeuge eine effizientere Steuerung des Verkehrsflusses (vgl. Friedrich, 2015, S. 339–349). Von den verkürzten Transferzeiten durch weniger Staus würden nicht nur Menschen auf dem Weg zum Arbeitsplatz profitieren, sondern auch Lieferfahrzeuge des Gütertransports. Freigesetzte Optimierungspotenziale für

in weniger Unfälle verwickelt sind als menschengesteuerte Fahrzeuge, nicht jedoch in allen.

7 Googles selbstfahrender Prototyp legte zwischen Februar und Oktober 2015 knapp 3 Millionen Testkilometer mit lediglich geringfügigen Zwischenfällen zurück, welche er jedoch in keinem Fall selbst verursachte (vgl. Gurney, 2015, S. 188; Hulverscheidt, 2015).

8 Beispielhafte prominente Unfälle werden in Kap. 2.2.2 beschrieben.

logistische Prozesse könnten dazu beitragen, die gesamtwirtschaftliche Produktivität zu steigern.

Weiterhin birgt die neue Art der Beförderung das Potenzial, freie Kapazitäten in Fahrzeugen für innovative Geschäftsmodelle einzusetzen, ob als Shuttles, Taxis oder gemeinsam genutzte Fahrzeuge. Insbesondere im Bereich Carsharing ist ein aufsteigender Trend zu beobachten (vgl. Bagloee et al., 2016, S. 289; Fagnant & Kockelman, 2018, S. 147–156; Gogoll & Müller, 2017, S. 685; Gurney, 2015, S. 194; Lenz & Fraedrich, 2015, S. 184–189). Sollte sich dieser hinreichend etablieren, könnte er sogar eine Auflösung der Grenzen zwischen Individualmobilität und öffentlichem Verkehr bewirken und so das Tempo der Transformation bestehender Mobilitätsstrukturen weiter beschleunigen (vgl. Beiker, 2015, S. 204–206; Lenz & Fraedrich, 2015, S. 189–192).

Auch in städtebaulicher Hinsicht tun sich durch das automatisierte Fahren im Zuge einer Neugestaltung von Parkflächen (vgl. Bennett, 2022, S. 197) und einer zentralen Koordination von deren Auslastung neue Perspektiven auf. So könnten städtische Parkraumflächen künftig für alternative Nutzungszwecke zur Verfügung stehen, beispielsweise zur Schaffung neuen Wohnraums oder für Grünflächen (vgl. Anderson et al., 2016, S. 25–27; Sparrow & Howard, 2017, S. 212). Indem Emissionen, Energie- und Kraftstoffverbrauch durch eine zentrale, intelligente Routenplanung und vorausschauende, sparsame Fahrweise optimiert werden, sind nicht zuletzt auch positive Effekte für die Umweltbilanz zu erwarten (vgl. Anderson et al., 2016, S. 28–38; Bennett, 2022, S. 197–198; Brändle & Grunwald, 2019, S. 292; Gurney, 2015, S. 193–194; Lim & Taeihagh, 2018, S. 5; Lin, 2013a). Der Verkehrsautomatisierung wird in diesem Sinne gar eine führende Rolle zugeschrieben, um den Herausforderungen des Klimawandels zu begegnen (vgl. Hula et al., 2018, S. 91–93).

3. Veränderte Mobilitätsbedürfnisse

Nicht zuletzt thematisiert das automatisierte bzw. autonome Fahren die sich stetig im Wandel befindenden Mobilitätsgewohnheiten und -bedürfnisse heutiger und zukünftiger Generationen. Es treibt Veränderungen in intergenerationellen Einstellungen gegenüber Mobilität und Fahrzeughaltung weiter voran: Die Generation der Babyboomer empfand den Erwerb der Fahrerlaubnis und des ersten eigenen Autos noch als Inbegriff persönlicher Freiheit und eines gewissen

sozialen Status. Nun haben sich die Prioritäten der ständig vernetzten, jüngeren Generationen in Richtung völliger Flexibilität – sogenannter *mobility on demand* – verschoben (vgl. KPMG LLP, Center for Automotive Research, 2012, S. 7; Pavone, 2015, S. 400–402). Diese Bedürfnistransformation steht im Zeichen eines seit einigen Jahren zu beobachtenden Trends, der Ausdruck sich verändernder Konsumgewohnheiten ist: ›Nutzen statt besitzen‹ oder ›Zugang statt Besitz‹ lauten die Devisen, die sich als ökonomisches und soziales Phänomen unter dem Begriff ›Sharing-Economy‹ längst einen Namen gemacht haben (vgl. Sundararajan, 2017).

Das Konzept selbstfahrender Fahrzeuge stimuliert diese neuen Prioritäten, indem es die traditionelle Abhängigkeit zwischen ›Mobil-Sein‹ und Fahreignung bzw. Fahrfähigkeit auflöst. Höherstufig automatisierte Fahrzeuge operieren unabhängig von menschlichem Eingreifen und ermöglichen dadurch mobilitätseingeschränkten Personengruppen, z. B. Menschen mit körperlichen und geistigen Beeinträchtigungen,[9] aber auch älteren Personen und Kindern, Zugang zu individueller Mobilität. Damit geht ein erheblicher Zuwachs an Lebensqualität durch mehr persönliche Unabhängigkeit und soziale Teilhabe einher (vgl. Beiker, 2012, S. 1151–1152; Brändle & Grunwald, 2019, S. 282–283; Gurney, 2015, S. 193; Hansson et al., 2021, S. 1398; Howard, 2013; Mladenovic & McPherson, 2016, S. 1137; Owens et al., 2019).

Hiermit sind die Ziele des autonomen Fahrens grob umrissen. Doch wie gestaltet sich deren Autonomie eigentlich aus technischer Sicht und vor allem im Zusammenspiel mit menschlichen Fahrern bzw. Insassen? Das folgende Unterkapitel illustriert, wie die Idee der Fahrautomatisierung im Verlauf der letzten sechzig bis achtzig Jahre schrittweise durch verschiedene Systemkomponenten realisiert wurde. Zudem wird das sogenannte Stufenmodell erläutert, anhand dessen der rote Faden einer immer weniger an menschliche Fahrzeugsteuerung gebundenen Mobilität in den kommenden Jahren weitergesponnen werden wird.

9 Für eine Diskussion spezifischer ethischer und rechtlicher Fragen in diesem Kontext siehe Bradshaw-Martin und Easton (2014).

2.1.3 Von Informanten über Assistenten zu Automaten: Evolution und Stufenmodell der Fahrautomatisierung

Die Idee, die menschliche fahrzeugführende Person bei ihren Fahraufgaben zu unterstützen und von diesen zu entlasten, ist schon seit beinahe einem Jahrhundert Gegenstand technischen Innovationsbestrebens. Bereits in den 1940er-Jahren wurden erste grundlegende Bausteine der Fahrunterstützung konzipiert und im Laufe der folgenden Jahrzehnte mit dem Automatikgetriebe (1940), der Servolenkung (1952) und dem Bremskraftverstärker (1955) erprobt, bevor das Konzept der Fahrerassistenzsysteme in den 1960er-Jahren endgültig Einzug in die Forschungsagenden hielt (vgl. Beiker, 2012, S. 1146–1147). In der Folge wurden mechanische und elektronische Komponenten intensiv weiterentwickelt und zur Marktreife gebracht. Diese frühen Assistenzsysteme übernahmen primär Aufgaben in der Regelung der Fahrdynamik, insbesondere der Lenk- und Fahrstabilität, z. B. in Form des Antiblockiersystems ABS (1978) und des Elektronischen Stabilitätsprogramms (ESP) (*Electronic Stability Control*) (1995), mit dem sich bahnbrechende Erfolge in Bezug auf die Fahrsicherheit einstellten (vgl. Verband der Automobilindustrie (VDA) e.V., 2023).

Durch technologischen Fortschritt vor allem in der Perzeption des verkehrlichen Umfelds entwickelten sich aus reinen Assistenzfunktionen bald komplexere Systeme, die spezifische Fahrfunktionen teilautomatisiert ausführen. Als erste das Fahrzeugumfeld erfassende Technologie kamen Ultraschallsensoren in der ersten Hälfte der 1990er-Jahre in der aktiven Parkassistenz zum Einsatz. Im Zusammenwirken mit erstmals eingesetzten Radarsensoren ermöglichten sie die Realisierung des Abstandsregeltempomats (*Adaptive Cruise Control* (*ACC*)), durch den sich das Fahren im gebundenen Verkehr teilautomatisieren ließ und eine Vielzahl von potenziell gefährlichen Situationen bereits in der Entstehung vermieden werden konnte. Dies markierte einen Meilenstein in der Geschichte der Fahrerassistenz. Die zunehmende technologische Reife von Ultraschall- und Radarsensorik sowie ein Durchbruch in der Kameratechnologie machten seit der Jahrtausendwende die Entwicklung komplexerer Systeme möglich, die verschiedene Technologien integrieren. Mithilfe der über Sensorfusion verknüpften Daten unterschiedlicher Sensortypen gelang die Erzeugung von Fahrzeug-Umfeldmodellen mit

stark verbesserten Präzisionsgraden, womit eine wichtige Voraussetzung für das automatisierte Fahren geschaffen war (vgl. Bengler et al., 2014, S. 6–7).[10]

Während traditionelle Fahrerassistenzsysteme als automatisierte Teilfunktionen der fahrzeugführenden Person lediglich vorübergehend in spezifischen Anwendungsfällen assistieren, übernimmt beim höherstufig automatisierten Fahren ein Fahrroboter (nahezu) alle Fahraufgaben. Die Basisarchitektur gegenwärtig entwickelter Systeme der Fahrzeugautomatisierung umfasst im Wesentlichen drei Komponenten: Perzeption, Prädiktion und Aktion. Die Hardware besteht aus einer Kombination fortschrittlicher Sensoren (Stereokameras, Radar unterschiedlicher Reichweite, Laser, GPS, Ultraschall), die es ermöglichen, Strukturen und Objekte in der Umgebung des Fahrzeugs sowie dessen Position zu erkennen. Diese Sensoren wirken in Verbindung mit Aktuatoren, Steuergeräten und integrierten Software-Algorithmen, welche die produzierten Daten verarbeiten und durch Ansteuerung der Aktuatoren die Aktionen des Fahrzeugs determinieren (vgl. Beiker, 2012, S. 1147; KPMG LLP, Center for Automotive Research, 2012, S. 10). Hierbei spielt die Integration von Anwendungen, welche auf Künstlicher Intelligenz basieren, eine große Rolle; zentrale Komponenten bei der Entwicklung autonomer Fahrsysteme sind vornehmlich neuronale Netze und datengestützte *Deep-Learning*-Techniken des maschinellen Lernens.[11]

---

10 Diese Fusionstechnik kommt beispielsweise beim Stauassistenten zum Einsatz, wo Radar- und Kamerasensoren zusammenwirken, um die Funktionalitäten des *ACC* und des Spurhalteassistenten für den Einsatz im gebundenen Verkehr zu integrieren.

11 Das in diesem Zusammenhang angewandte Konzept Künstlicher Intelligenz begreift diese im Sinne der schwachen KI-Hypothese als Systeme, die in der Lage sind, menschliche Intelligenz in Bezug auf spezifische, genau definierte Aufgaben zu simulieren, zu erweitern oder mit ihr zu konkurrieren (vgl. Vallor & Bekey, 2017, S. 339–340). Ein zentraler Aspekt bei der Entwicklung derartiger Systeme sind Techniken maschinellen Lernens, deren Realisierung über neuronale Netze (*neural networks*) erfolgt, welche dem Aufbau des menschlichen Gehirns nachempfunden sind und Eingabedaten in mehreren komplexen Schichten (*layers*) via *Deep-Learning*-Techniken (vgl. Reed et al., 2021, S. 784) zu kontrollierten Aktionen verarbeiten: »[…] the network gradually ›learns‹ from repeated ›experience‹ (multiple training runs with input datasets) how to optimize the machine's ›behavior‹ (outputs) for a given kind of task.« (Vallor & Bekey, 2017, S. 340).

Das autonome Fahren stellt das Endziel einer technologischen Entwicklungsagenda dar, die sicherheitskritische Fahrfunktionen schrittweise automatisiert. Die Evolution der Fahrautomatisierung kann verstanden werden als eine Agenda zunehmender Systemintegration; sie führt von passiven Warn- und Informationssystemen über assistiertes hin zu automatisiertem Fahren und schließlich zum autonomen, selbstfahrenden Fahrzeug (vgl. Beiker, 2012, S. 1147–1148). Experten in Forschung und Entwicklung unterscheiden daher verschiedene Level der Automatisierung, anhand derer sich bestehende und zukünftige Fahrzeugsysteme einordnen lassen. Grundlage der Kategorisierung ist eine deskriptive Taxonomie, die Teil der erstmals 2014 von der SAE International (ehemals »Society of Automotive Engineers«) veröffentlichten Norm J3016 ist. Anhand funktionaler Mindestanforderungen an das jeweilige System werden dabei sechs Stufen beschrieben, die sich an den spezifischen Rollen orientieren, die menschlichen Nutzern einerseits und dem automatisierten Fahrsystem andererseits im Hinblick auf die dynamische Fahraufgabe[12] zukommen. Bei der Einstufung des Automatisierungsgrades werden dabei nur solche Systeme berücksichtigt, die sich dauerhaft auf die dynamische Fahraufgabe oder Teile davon auswirken und nicht nur kurzzeitig in potenziell gefährlichen Situationen aktiviert werden, wie beispielsweise ein Notbremsassistent (vgl. SAE On-Road Automated Vehicle Standards Committee, 2014, S. 1–2).[13]

12 Der Begriff ›dynamische Fahraufgabe‹ (*dynamic driving task*) beschreibt die Kontrolle des Fahrzeugs. Er umfasst die operativen (Lenken, Bremsen, Beschleunigen etc.) und taktischen (Reagieren auf Ereignisse) Aspekte der Fahraufgabe, wohingegen er den strategischen Aspekt (Bestimmen von Zielen und Wegpunkten) außer Acht lässt (vgl. SAE On-Road Automated Vehicle Standards Committee, 2014, S. 2).

13 Hinsichtlich der Fokussierung auf spezifische Aspekte existieren alternative Klassifizierungen der Automatisierungsgrade des automatisierten Fahrens. So liegt der Schwerpunkt bei der vom VDA konzipierten Version, die inhaltlich deckungsgleich ist mit der Taxonomie der SAE, auf der Beschreibung der Aufgabenteilung zwischen menschlichen Fahrern und dem System in laienverständlicher Sprache. Dabei werden bereits Systeme der Stufe 3 als hochautomatisiert, 4 entsprechend als vollautomatisiert und 5 als fahrerlos bezeichnet. Auch die amerikanische National Highway Traffic Safety Administration bietet ein alternatives Schema an, das die Level 4 und 5 zusammenfasst. Der Fokus liegt dabei auf dem Anteil, den automatisierte Komponenten im Hinblick auf sicherheitskritische Funktionen des Fahrzeugs innehaben (vgl. National Highway Traffic Safety Administration, 2013, S. 4–5).

Die einzelnen Level des Stufenmodells implizieren keine zwangsläufige Reihenfolge der Markteinführung; jedoch lässt der inkrementelle Entwicklungsprozess automatisierter Fahrfunktionen die verwendeten Technologien reifen und legt somit den Grundstein für das Erreichen der jeweils nächsten Stufe (vgl. Bengler et al., 2014, S. 12). Die Automatisierungslevel sind wie folgt charakterisiert:

Level 0 (*No automation*): Der Fahrer hat jederzeit die Kontrolle über das Fahrzeug, i. e. er führt dauerhaft alle dynamischen Fahraufgaben aus, wobei ihm Warn- und Informationssysteme wie Spurhalte- und Notbremsassistenten Hilfestellung geben.

Level 1 (Assistiertes Fahren/*driver assistance*): Eine oder mehrere spezifische Kontrollfunktionen des Fahrzeugs, die unabhängig voneinander operieren, sind automatisiert. Sie unterstützen den Fahrer in Form von Fahrerassistenzsystemen (*driver assistance systems*), die entweder das Steuern (z. B. *lane centering*) oder Beschleunigen bzw. Bremsen in einem bestimmten Fahrszenario (*driving mode*) übernehmen. Ein Beispiel für ein solches System ist der Einparkassistent.

Level 2 (Teilautomatisiertes Fahren/*partial automation*): Im Wesentlichen wie Level 1, nur wirken hier mindestens zwei automatisierte Kontrollfunktionen zusammen; ein oder mehrere Fahrerassistenzsysteme übernehmen zeitgleich sowohl das Steuern als auch Beschleunigen bzw. Bremsen in bestimmten Situationen. Ein Beispiel für ein entsprechendes System ist hier der Stauassistent.

Im Allgemeinen gilt für die Level 0 bis 2, dass der menschliche Fahrer ›fährt‹, d. h. er muss das System und die Fahrumgebung jederzeit überwachen, um gegebenenfalls eingreifen zu können. Er ist vollständig verantwortlich für den sicheren Betrieb des Fahrzeugs. Im Übergang von Level 2 zu 3 findet gewissermaßen ein Bruch statt, es kommt zu einer grundsätzlichen Veränderung in der Aufgabenteilung und Verantwortlichkeit. Der Fahrer gibt schrittweise weite Teile der Fahraufgabe an das System ab, das fortan die Fahrzeugkontrolle und Überwachung der Fahrumgebung übernimmt und ›fährt‹.

Level 3 (Bedingt automatisiertes Fahren/*conditional automation*): Sicherheitskritische Funktionen sind unter bestimmten Fahrbedingungen komplett automatisiert; das System übernimmt in spezifischen Szenarien die Fahrzeugkontrolle vollständig. Der

Fahrer muss das System nicht mehr jederzeit überwachen, aber verfügbar sein, wenn es im konkreten Anwendungsfall an seine Grenzen stößt und den Menschen zur Übernahme des Steuers auffordert. Der Fahrer fungiert als Absicherung bzw. Rückfallebene für das System. Tritt ein solcher Fall ein, ›fährt‹ fortan der Fahrer. Ein beispielhaftes System für diese Stufe ist das Staufolgefahren.

Level 4 (Hochautomatisiertes Fahren/*high automation*): Ein Eingreifen des Fahrers ist nicht mehr erforderlich. Das System übernimmt die Kontrolle in bestimmten Situationen vollständig. Wenn der Fahrer nicht oder nicht angemessen auf eine Aufforderung zur Übernahme reagiert, geht das Fahrzeug in einen sicheren Zustand über, indem es z. B. auf dem Seitenstreifen anhält.

Level 5 (Vollautomatisiertes Fahren/*full automation*): Der Fahrmodus wird als ›autonom‹ bzw. das Fahrzeug als ›selbstfahrend‹ bezeichnet. Alle Fahrfunktionen sind vollständig automatisiert; das System übernimmt zu jeder Zeit und in allen Situationen sowie unter allen Umwelt- und Fahrbahnbedingungen die Kontrolle über das Fahrzeug und überwacht die Fahrumgebung ständig. Der Fahrer greift ins Fahrgeschehen nicht mehr ein.

An dieser Stelle kann das Phänomen des autonomen Fahrens als ausreichend eingeführt gelten. In den folgenden Unterkapiteln wird sich nun der ethischen Dimension zugewandt, die selbstfahrende Fahrzeuge entfalten. Der Einstieg in diese komplexe Thematik erfolgt zunächst über eine Betrachtung der Zusammenhänge zweier Perspektiven: autonome Fahrzeuge als technische Innovation einerseits und als ökonomisches (Konsum-)Gut andererseits.

2.2 Herausforderungen im Kontext der Entwicklungsagenda

2.2.1 Die Wechselbeziehung zwischen technischer Reife und Wirtschaftlichkeit

Die Einführung des automatisierten Fahrens in den öffentlichen Straßenverkehr wird maßgeblich durch drei Aspekte stimuliert: die technische Reife der Fahrzeugsysteme und deren Komponenten, die Wirtschaftlichkeit des Gutes ›automatisiertes Fahrzeug‹ und den

jeweils geltenden Rechtsrahmen. Zwischen diesen bestehen Abhängigkeiten und Wechselwirkungen unterschiedlichen Ausmaßes.

Fortschritte in der technologischen Entwicklung entlang des Stufenmodells sind gleichbedeutend mit einem höheren Automatisierungsgrad; sie lassen sich nur erzielen, wenn automatisierte Fahrzeuge zunehmend unabhängiger von menschlichem Eingreifen operieren, ohne dabei die anvisierten Sicherheitsziele aus den Augen zu verlieren. Die zentralen Aufgaben für Design und Entwicklung autonomer Fahrsysteme (Level 5 und teilweise 4) stellen sich hinsichtlich Perzeption, Kognition, Verhaltensentscheidung und -ausführung (vgl. Wachenfeld & Winner, 2015, S. 466). Durch die intensive Forschungsarbeit und die hohen Entwicklungsinvestitionen des letzten Jahrzehnts sind die wesentlichen technischen Voraussetzungen für das automatisierte Fahren zum heutigen Tag bereits in Ansätzen geschaffen, jedoch noch nicht zur vollständigen technischen Reife gelangt.

Die mitunter größten Herausforderungen bei der technischen Realisierung höherstufiger Automatisierung bestehen in einer kontinuierlichen Verbesserung der maschinellen Wahrnehmungsleistung (vgl. Beiker, 2012, S. 1148–1149; Bengler et al., 2014, S. 6; KPMG LLP, Center for Automotive Research, 2012, S. 12). Maschinenlesbare Informationen sind die Basis für algorithmische Verhaltensentscheidungen. Ein zentrales Element derselben stellt die Situationsprädiktion[14] dar, die mögliche Entwicklungen einer wahrgenommenen Szene aus dem verkehrlichen Umfeld des Fahrzeugs innerhalb eines festgelegten Zeithorizonts von wenigen Sekunden vorausberechnet. Auf Basis dieser sogenannten Episoden werden sodann Aktionen geplant und schließlich korrespondierende Bewegungspfade – sogenannte Trajektorien – anhand spezifischer Kriterien errechnet (vgl. Dietmayer, 2015, S. 421). Problematisch ist hierbei nun, dass die Prädiktionsfähigkeit gegenwärtiger Wahrnehmungssysteme im Notfall nicht ausreicht: Im Fall von auftretenden Funktionseinschränkun-

14 Die Situationsprädiktion baut auf der Situationserkennung auf, bei der einzelne Komponenten des dynamischen Fahrzeug-Umfeldmodells zueinander in Beziehung gesetzt werden. Dietmayer (2015, S. 420–432) nennt Unsicherheiten hinsichtlich dreier Aspekte, durch welche die maschinelle Wahrnehmung beeinträchtigt wird: Zustandsunsicherheit, Existenzunsicherheit, Klassenunsicherheit.

gen bei Systemen auf Level 3 und 4 beträgt der kritische Zeithorizont, den das Fahrzeug für die Rückgabe an den Fahrer selbst überbrücken muss, ca. fünf bis zehn Sekunden; ebenso lange benötigt ein Fahrzeug auf Level 5 für die Erreichung eines eigensicheren Zustands.[15]

Trotz eingebauter sensorischer Redundanz können momentane Systeme eine zuverlässige Prozesskette bis hin zur sicheren Fahrzeugführung in Notsituationen nur unzureichend gewährleisten. Erforderliche Fortschritte im Bereich der maschinellen Perzeption und Kognition umfassen daher sowohl die Detektion von Objekten und ihre physikalische Vermessung als auch die Zuordnung korrekter semantischer Bedeutungen (vgl. Dietmayer, 2015, S. 420–426). Es reicht nicht aus, ein Verkehrsschild nur als solches zu erkennen; das Fahrzeug muss auch dessen Bedeutung, z. B. ›Stop‹ oder ›Vorfahrt gewähren‹, erfassen können. Gleiches gilt für die Gestenerkennung menschlicher Verkehrsteilnehmer, die einen wesentlichen Teil der Kommunikation im Verkehrsgeschehen ausmacht (vgl. Holzbock et al., 2023a). Hierfür ist die qualitative Weiterentwicklung von Kamera- und Sensortechnologien auf Hardwareseite ein kritischer Erfolgsfaktor.

Um Perzeption und Situationsverstehen auch in komplexen dynamischen Umgebungen zuverlässig sicherzustellen, spielt Künstliche Intelligenz mit zunehmendem Automatisierungsgrad, v. a. ab Level 3, eine bedeutende Rolle (vgl. Bengler et al., 2014, S. 12). Vor allem der Einsatz von *Deep-Learning*-Techniken ist kontinuierlich zu optimieren, mittels derer das Fahrzeugsystem zunächst in Simulationen und später in kontrollierten Feldversuchen ›trainiert‹ wird, um in realen Situationen zuverlässig den korrekten Output generieren zu können (vgl. Vallor & Bekey, 2017, S. 341). Zum Beispiel präsentiert ein neuerer Ansatz von Holzbock et al. (2023b) ein Fahrzeug-Umgebungsmodell, das die Körperhaltung von Fußgängern als Indiz für deren Bewegungsabsichten mithilfe neuronaler Netzwerke interpretiert. Auf Level 5 wirken intelligente, sensorgestützte Sicherheitssysteme schließlich mit vernetzter Fahrzeugtechnologie zusammen, um mit der verkehrlichen Infrastruktur zur Laufzeit zu kommunizie-

15 Eigensicherheit (*intrinsic safety*) wird einem System dann zugesprochen, wenn es über spezifische Konstruktionsmechanismen bzw. -prinzipien verfügt, die sicherstellen, dass auch im Fall einer Störung oder eines Systemversagens keine grundsätzliche Gefährdung vom System selbst ausgeht.

ren. Dies ermöglicht eine stetige Optimierung des Fahrverhaltens beispielsweise hinsichtlich Geschwindigkeit, Routenplanung oder Antizipation potenzieller Gefahren, wodurch Verkehrseffizienz und Fahrsicherheit erhöht werden. So schlagen Schumann et al. (2023) beispielsweise einen erweiterten Pfadplanungsalgorithmus für große Umgebungen vor, der eine neuartige Methode zur Erkennung signifikanter Umgebungsveränderungen beinhaltet und eine effiziente Nutzung der Manövrierfähigkeit zur Pfadplanung in engen Umgebungen ermöglicht.

Spezifische Herausforderungen auf dem Weg zur Marktreife des hoch- und vollautomatisierten Fahrens stellen sich im Hinblick auf die Konvergenz von Kommunikations- und Sensortechnologien sowie das Entwicklungspotenzial einzelner Komponenten. Konkrete Aufgabenstellungen sind z. B. die Optimierung der GPS-Technologie zur Positionsbestimmung, hochauflösende Kartierungsmethoden zur Verbesserung der Umfeldwahrnehmung, Bildverarbeitungsverfahren sowie zuverlässige und intuitive Mensch-Maschine-Schnittstellen für die Interaktion zwischen Fahrer und System (vgl. Bengler et al., 2014, S. 7). Neben der technischen Ausstattung der Fahrzeuge muss auch der Ausbau einer digitalisierten Verkehrsinfrastruktur vorangetrieben werden; dringlich ist hier vor allem die flächendeckende Ausstattung des Verkehrsnetzes mit dem neuen Mobilfunkstandard 5G (vgl. Verband der Automobilindustrie (VDA) e.V., 2022b).

Abseits von technologischen Durchbrüchen rückt auch die Entwicklung von Test- und Bewertungskonzepten künftig verstärkt in den Fokus; der »kritische Pfad zum autonomen Fahren« (Bengler et al., 2014, S. 16) führt über die Bereitstellung innovativer Metriken, mittels derer sich die Leistungsfähigkeit von Menschen und Fahrrobotern messen und vergleichen lässt, mittelfristig Testverfahren verbessert werden können und schließlich die Zuverlässigkeit der zur Reife gebrachten Systeme steigt. Letztere muss sich an den Anforderungen der funktionalen Sicherheit[16] orientieren; einschlägige

16 Die funktionale Sicherheit (*functional safety*) ist ein Teilbereich der allgemeinen Sicherheit, der sich auf die korrekte Funktionsweise eines Systems bezüglich seiner Eingabeverarbeitung bezieht. Sie ist dann gegeben, wenn jede spezifizierte Sicherheitsfunktion ausgeführt und die jeweilige Anforderungsstufe erreicht wird. Konkret bedeutet das, dass zur Erreichung der funktionalen Sicherheit alle

Referenzen für den Automobilbereich sind hierbei die Norm ISO 26262 (»Road vehicles – Functional Safety«)[17] und insbesondere deren Schlüsselkomponente, das Risiko-Klassifizierungsschema *ASIL* (*Automotive Safety Integrity Level*)[18], das für automatisierte Fahrzeuge die höchste Sicherheitsstufe (*ASIL D*) fordert.

Zweifellos ist die Entwicklung fortschrittlicher Technologien hin zur technischen Reife eine erfolgskritische Voraussetzung für die anvisierte Einführung des automatisierten Fahrens. Jedoch ist deren Gelingen auch abhängig von ökonomischen Faktoren. So sind automatisierte Fahrzeuge neben allen strategischen Zielen zur Stärkung des Technologie- und Automobilstandorts Deutschland schließlich auch ein Wirtschaftsgut, das den Gesetzmäßigkeiten der Marktwirtschaft unterliegt. Gemäß den Mechanismen von Angebot und Nachfrage ist eine großflächige Serienproduktion automatisierter Fahrfunktionen nur realistisch, wenn die entsprechende Nachfrage vorhanden ist. Auf diese Weise ließen sich zum einen die über einen langen Zeitraum notwendigen hohen Investitionen der Hersteller über den laufenden Markt refinanzieren. Zum anderen sind mittelfristig Netzwerkeffekte zu erwarten: Je mehr Menschen automatisierte Fahrzeuge nutzen, desto stärker wirken Skaleneffekte auf die Produktionskosten und damit auch den Preis, was wiederum die Attraktivität für weitere Nutzer erhöht (vgl. KPMG LLP, Center for Automotive Research, 2012, S. 20). Die Akzeptanz am Markt hat einen entscheidenden Einfluss auf die Durchdringung des Fahrzeug-

---

Maßnahmen gehören, die zur Fehlervermeidung dienen, ebenso wie Vorgänge, mit deren Hilfe während des Betriebs auftretende Fehler beherrscht werden können.

17 Die ISO 26262 stellt eine Abwandlung der IEC 61508 dar, die an die spezifischen Anforderungen im Automobilbereich angepasst ist: Lebenszyklus eines Fahrzeugs, Schnittstellen bei verschiedenen Zulieferaufträgen, konfigurierbare Software usw. Anwendbar ist die Norm auf alle Fahrzeugklassen bis 3500 kg (vgl. International Organisation for Standardisation, 2018).

18 Die Klassifizierung gemäß *ASIL* bezeichnet das Maß, in dem eine Fehlfunktion eines Systems relevant für dessen Sicherheit ist. Sie orientiert sich an drei Parametern: Häufigkeit der Situation mit Relevanz der jeweiligen Fehlfunktion (*exposure*), Kontrollierbarkeit der Fehlfunktion (*controllability*) und Schwere der Auswirkung bei geringer Kontrollierbarkeit (*severity*). Je nach Belegung der Parameter kann der *ASIL* auf einer Skala von A bis D bestimmt werden und die sich daraus ergebenden speziellen Anforderungen sind zusätzlich zur ISO 26262 zu erfüllen (vgl. International Organisation for Standardisation, 2018, Teil 9).

bestands mit automatisierten Fahrzeugen.[19] Diese wiederum stimuliert einerseits Umfang und Geschwindigkeit des technologischen Fortschritts und andererseits auch das Maß, in dem das autonome Fahren positive Wirkung entfaltet. Letztere ist skalierbar sowohl mit der Anzahl zugelassener Fahrzeuge als auch mit dem Grad der Automatisierung: Je mehr Fahraufgaben das System übernimmt, desto weniger verbleiben bei der fahrzeugführenden Person und desto seltener führen menschliche Fahrfehler zu Unfällen. Je mehr Fahraufgaben automatisiert ablaufen, desto mehr Szenarien können effizient koordiniert werden, z. B. zeit- und kraftstoffeffizientes Fahren im gebundenen Verkehr, im Stau oder bei der Parkplatzsuche. Für mobilitätseingeschränkte Personengruppen, die zum Führen konventioneller Autos nicht ermächtigt bzw. in der Lage sind, werden automatisierte Fahrzeuge zudem erst dann nutzbar, wenn kein menschliches Eingreifen mehr erforderlich ist, d. h. ab Level 4.

Wann können wir mit der Marktreife derartiger Systeme rechnen? Im nachfolgenden Unterkapitel wird der aktuelle Entwicklungsstand autonomer Fahrzeuge resümiert und sein Verhältnis zu entsprechenden gesetzlichen Bestimmungen ausgelotet.

2.2.2 Wo stehen wir heute? Aktueller technischer Stand und regulative Verordnungen

Im gegenwärtigen Realbetrieb auf öffentlichen Straßen sind Fahrzeuge mit teilautomatisierten Fahrfunktionen des Levels 2 weitgehend etabliert. In bestimmten Fahrsituationen, wie bei Stau- oder Autobahnfahrten, kann das System bereits die Fahrzeugkontrolle übernehmen, wobei die fahrzeugführende Person stets die Fahrumgebung überwachen und jederzeit zum Eingreifen bereit sein muss. Systeme, die sich am Übergang zwischen Level 2 und 3 befinden, stehen zur Serienproduktion bereit bzw. werden als Prototypen im (beschränkten) Realverkehr zur Weiterentwicklung der KI-gestützten Software erprobt. Automobilbauer setzen dabei verstärkt auf Ko-

19 Becker und Axhausen (2017) erarbeiten einen systematischen Überblick über Studien, die sich mit der Akzeptanz von automatisierten Fahrzeugen beschäftigen. Als relevante Faktoren identifizieren sie u. a. personenbezogene Charakteristika (z. B. Geschlecht, Alter), Präferenzen für Einsatzszenarien (z. B. Stadt, Autobahn) und eine grundsätzliche Affinität zu Systemen der Fahrerassistenz.

operationen mit Unternehmen außerhalb des Automobilbereichs; so entwickelt Daimler seine Mercedes-Benz-Modelle in enger Zusammenarbeit mit Grafikprozessor- und Chipsätze-Entwickler Nvidia, um KI-Komponenten zur Reife zu bringen. Auch Volkswagen testet die Komponenten seines Entwicklungspartners Argo AI auf Teststrecken in typischen Verkehrssituationen. Ein Durchbruch gelang kürzlich Mercedes-Benz, dem für den Staupiloten »Drive Pilot« als erstem deutschen Automobil die Zulassung für ein Level-3-System erteilt wurde; dieses fährt im Stau mit Geschwindigkeiten bis zu 60 km/h selbständig, ohne dass der Fahrer das System ständig überwachen muss (vgl. Rudschies & Kroher, 2024). Konkurrent BMW hingegen erhielt kürzlich vom Kraftfahrtbundesamt die Genehmigung für seinen Autobahnassistenten, der seit Oktober 2023 im neuen 5er-Modell erhältlich ist. Dabei handelt es sich um ein sogenanntes *Level-2-Hands-off*-System, d. h. das System fährt selbständig bis zu 130 km/h, der Fahrer muss allerdings jederzeit eingreifen können.

Kontinuierlich arbeitet die Automobilbranche in Deutschland an der Neuentwicklung höherstufiger Systeme. Wegbereiter ist hier hauptsächlich das automatisierte *Valet Parking*, mit der das Level 4 des hochautomatisierten Fahrens erstmals beschritten wird. Es stellt eine der ersten in Serie realisierten autonomen Fahrfunktionen dar, die bereits eine Zulassung für den Alltagsbetrieb erhalten haben. Als erster Automobilhersteller hat Mercedes-Benz die entsprechende Funktion in einem Serienfahrzeug verbaut, gemeinsam mit Partner Bosch im Parkhaus des Mercedes-Benz Museums in Stuttgart erprobt und darf diese nun in fest definierten Bereichen im Regelbetrieb einsetzen (vgl. Mercedes-Benz Group AG, 2020). Ende 2022 erteilte das Kraftfahrt-Bundesamt den beiden Stuttgarter Unternehmen die gemeinsame Zulassung für das weltweit erste zertifizierte *Automated-Valet-Parking*-System, den fahrerlosen »Intelligent Park Pilot«, der seitdem in einem speziell ausgerüsteten Parkhaus am Stuttgarter Flughafen operiert. Gesteuert durch das Fahrsystem und die von Bosch entwickelte intelligente Infrastruktur des Parkhauses erfolgt der Parkvorgang vollständig automatisiert, nachdem der Fahrer das Fahrzeug in einer definierten Übergabezone verlassen hat (vgl. Bosch Mobility Solutions, 2021; Mercedes-Benz Group AG, 2022). Weitere Meilensteine auf dem Weg zum autonomen Fahren aus Sicht deutscher Hersteller konnten im August 2024 vermeldet werden: In Kooperation mit dem chinesischen Unternehmen

WeRide schickte Mercedes-Benz erste Level-4-Testfahrzeuge auf die Straßen Pekings, zunächst nur auf Autobahnen und in spezifisch ausgewiesenen, stark frequentierten Bereichen in der Innenstadt. Erprobungsziel ist dabei u. a. die Umfelderkennung mittels verschiedener Sensoren (vgl. Köllner, 2024c). Zudem erhielt eine von der deutschen Firma Vector entwickelte *AUTOSAR*[20]-Basissoftware die Zertifizierung gemäß *ASIL D* und ist ab sofort für automatisierte Funktionen in der Fahrzeugentwicklung einsetzbar (vgl. Köllner, 2024b).

In den USA sind höherstufige Systeme bereits fester Bestandteil alltäglicher Mobilität. Allen voran Automobilriese General Motors investiert an verschiedenen Standorten weltweit hohe Summen in die Entwicklung automatisierter Fahrsysteme. Als Pionier im Bereich der Autopiloten agierte lange Zeit der von US-Milliardär Elon Musk geführte amerikanische Konzern Tesla, der bereits 2015 ein System präsentierte, das an der Grenze zu Level 3 auf Autobahnen operiert. Trotz seiner Innovationskraft – oder vielleicht gerade deswegen – wird Tesla regelmäßig mit negativen Schlagzeilen bedacht, die auf eine Vielzahl aufsehenerregender Unfälle und diverse laufende Klagen zurückzuführen sind. Unlängst entschied ein kalifornisches Gericht, dass das Autopilotsystem des Herstellers nicht für einen Unfall im Jahr 2019 verantwortlich zu machen ist, als ein Tesla-Fahrzeug von der Fahrbahn abgekommen und in Flammen aufgegangen war (vgl. ARD-Tagesschau, 2023). Dies könnte ein richtungsweisender Erfolg nicht nur für den Konzern, sondern auch für die US-Justiz sein. Diese wird sich in den kommenden Jahren mit zahlreichen ähnlichen Klagen befassen müssen, denn spektakuläre Unfälle im Zusammenhang mit automatisierten Fahrzeugen sind in den USA nicht selten. So stand auch die General-Motors-Tochter Cruise zeitweise in der Kritik, nachdem sie zuvor eine Vorreiterrolle eingenommen hatte. Als Folge eines fatalen Unfalls im vergangenen Jahr, an dem ein Robotaxi der Firma beteiligt war, wurde dieser vorerst die Lizenz entzogen. Inzwischen sind die Cruise-Fahrzeuge mit mittelfristigen Expansionsplänen auf die Straßen zurückgekehrt,

20 *AUTOSAR* (*AUTomotive Open System ARchitecture*) ist die Bezeichnung einer globalen Partnerschaft führender Unternehmen in der Automobil- und Softwareindustrie mit dem Ziel, standardisierte Software-Frameworks und -architekturen für zukünftige intelligente und sichere Mobilitätslösungen zu entwickeln und zu etablieren (vgl. AUTOSAR GbR, 2024).

werden zunächst allerdings von Menschen gesteuert, um Kartenmaterial zu aktualisieren und Vertrauen zurückzugewinnen (vgl. Kennemer, 2024).

Erfolgsgeschichte schreibt hingegen die Google-Schwestergesellschaft Waymo, deren Flotten von Robotaxis bereits seit Längerem in Phoenix, Arizona, San Francisco und Los Angeles unterwegs sind, teilweise ohne Überwachung durch einen Sicherheitsfahrer. Als weltweit erstes Unternehmen, das Robotaxis zur Marktreife gebracht hat, droht Waymo sogar Tesla bei möglichen Investoren den Rang abzulaufen (vgl. Göpfert, 2024). Bei allem ökonomischen Erfolg verbleiben allerdings auch hier Sicherheitsbedenken, die verstärkt ins Blickfeld der US-Verkehrsbehörde National Highway Traffic Safety Administration (NHTSA) geraten, wenngleich abgesehen von Situationen mit Blechschäden bislang keine folgenschweren Zwischenfälle zu vermelden waren (vgl. Ohnsman, 2024).

Es ist naheliegend, dass die jeweils gültige Rechtslage eine entscheidende Rolle dahingehend spielt, wo und in welchem Umfang selbstfahrende Fahrzeuge bereits praktisch zum Einsatz kommen. Einen Rechtsrahmen für das autonome Fahren und seine spezifischen ethischen und juristischen Herausforderungen zu schaffen, ist dabei eine rechtspolitische Aufgabe (vgl. Hilgendorf, 2019, S. 357–358). Regulatorische Instrumente haben sich in der Vergangenheit als effektive Stimuli erwiesen, um die Marktdurchdringung neuer Technologien im Automobilbereich zu erhöhen und auf diese Weise die Weiterentwicklung der Automatisierungsfunktionen entlang des Stufenmodells zu unterstützen. Umgekehrt können fehlende oder unzureichende gesetzliche Bestimmungen auch hemmend auf den Innovationsfortschritt und eine großflächige Markteinführung wirken (vgl. Hilgendorf, 2018a, S. 681). In den vergangenen Jahren stand die Gesetzeslage in Deutschland der Verkehrsautomatisierung noch eher konservativ gegenüber. So trat im Juni 2017 zunächst eine Änderung des Straßenverkehrsgesetzes in Kraft, wonach Systeme auf Level 3 zeitweise die Fahraufgabe übernehmen dürfen, wenn weiterhin eine fahrzeugführende Person im Notfall die Kontrolle übernehmen kann. Die Erprobung von Prototypen höherstufig automatisierter Fahrzeuge im Realverkehr wurde bis zu diesem Zeitpunkt noch durch gesetzliche Hürden ausgebremst; sie durften nur

mit Sondergenehmigungen auf fest definierten Betriebsflächen oder Teststrecken fahren.[21]

Diese über weite Strecken zögerliche Anpassung des Rechtsrahmens für den Regelbetrieb von autonom agierenden Fahrzeugen im öffentlichen Verkehr ist ursächlich dafür, dass Branchenexperten der deutschen Automobilindustrie trotz aller Investitionskraft einen Entwicklungsrückstand auf die USA attestieren. Dort sind die rechtlichen Rahmenbedingungen aufgrund des förderalen *Laissez-Faire*-Ansatzes günstiger; die Entscheidungsbefugnis über das Verkehrsrecht liegt bei den Bundesstaaten, sodass sich die jeweiligen Regelungen in verschiedenen Staaten unterscheiden. Einige erlauben den Regelbetrieb selbstfahrender Fahrzeuge bereits seit geraumer Zeit; die liberalsten Regularien bestehen in Arizona, Texas, Nevada und Michigan. Im Januar 2021 haben die USA zudem die Zulassungsanforderungen in Bezug auf die zu erfüllenden Sicherheitsnormen für Roboterautos gelockert, um die Einführung in den Realverkehr und damit mittelfristig auch die Markteinführung zu beschleunigen. Diese Regelung gilt allerdings zunächst nur für vollautomatisierte Fahrzeuge im Gütertransport, da sich hier keine schutzwürdigen fahrzeugführenden Personen mehr an Bord befinden (vgl. Schmidt, 2021). Im Herbst 2023 hat die US-amerikanische NHTSA neue regulative Entwürfe unter der Bezeichnung »AV STEP« (*ADS-equipped Vehicle Safety, Transparency, and Evaluation Program*) vorgelegt. Diese sollen die Einführung selbstfahrender Autos weiter beschleunigen, indem sie die Höchstgrenzen für die maximal zulässige Anzahl an Fahrzeugen mit Assistenz- und Automatisierungssystemen aufheben und zugleich eine vermehrte Datenpreisgabe von Seiten der Unternehmen fordern (vgl. Chasins, 2024; McElligott, 2023).

Andererseits können gesetzliche Vorschriften auch als »Mittel einer reflektierten Innovationsförderung« (Hilgendorf, 2018b, S. 93) wirken, beispielsweise indem sie die Serienausstattung von Neufahrzeugen mit bestimmten Fahrfunktionen verpflichtend vorschrei-

21 Dies erfolgte beispielsweise als innerbetriebliche Fahrten auf dem Firmengelände oder im Rahmen von Forschungsprojekten wie dem »EVA-Shuttle« für den Personentransport zwischen Haustür und Haltestelle des öffentlichen Nahverkehrs im Karlsruher Stadtgebiet oder als *Mobility-on-Demand*-System zur Reduzierung des Individualverkehrs am Touristenziel Hambacher Schloss im Projekt »Hambach-Shuttle« (vgl. Bundesministerium für Digitales und Verkehr, 2022a).

ben.[22] Ein zentrales Themenfeld, das kritisch für Investitionsrisiken und damit auch Investitionsanreize der Hersteller ist,[23] betrifft die Klärung haftungsrechtlicher Fragen (vgl. Borenstein et al., 2017, S. 67–68; Marchant & Lindor, 2012, S. 1337). Sollen Hersteller in vollem Umfang haften, wenn ein automatisiertes Fahrzeug in einen Unfall verwickelt ist? Die einschlägige juristische Fachliteratur wächst kontinuierlich; bisher wurden vor allem Ansätze einer angepassten und erweiterten Produkthaftung diskutiert, die Defizite traditioneller Regelungen ausgleichen und die spezifischen Anforderungen der neuen Technologien berücksichtigen sollen (vgl. Douma & Palodichuk, 2012; Gasser, 2015; Gurney, 2013, 2017; Hilgendorf, 2018a; Koch, 2022; Wu, 2015, 2020).[24]

Um die Konkurrenzfähigkeit der europäischen Automobilwirtschaft zu erhalten, haben die EU-Staaten eine Harmonisierung der Vorschriften auf internationaler (EU-)Ebene zum Ziel erklärt. Da dieses jedoch in seiner Ausformulierung bis auf Weiteres auf sich warten lässt, schuf die damalige Bundesregierung schließlich auf eigene Initiative einen Rechtsrahmen, um das automatisierte Fahren weiter voranzubringen. Im Mai 2021 beschloss der Deutsche Bundestag mit Zustimmung des Bundesrates schließlich das für das Bundesgebiet gültige sogenannte »Gesetz zum autonomen Fahren«. Damit war Deutschland das erste Land weltweit, das einen nationalen Rechtsrahmen für den Regelbetrieb hochautomatisierter Fahrzeuge (Stufe 4) in festgelegten Betriebsbereichen im öffentlichen Straßenverkehr bereits ab 2022 setzte. Ziel dieser Regelung ist es, Innovationen in der Fahrautomatisierung schnell zu etablieren und die Weiterentwicklung durch Erprobung unter Realbedingungen zu

22 So gehört etwa das Elektronische Stabilitätsprogramm seit 2014 zur Pflichtausstattung aller in der Europäischen Union zugelassenen Pkw (vgl. Bengler et al., 2014, S. 2–4).

23 Garza (2011, S. 616) erläutert, dass die Entwicklung autonomer Fahrzeuge sich trotz anfänglich hoher Haftungsrisiken mittelfristig für die Hersteller lohnt, da zu erwarten ist, dass sich langfristig Häufigkeit und Schwere negativer Vorfälle mit steigender Nutzerakzeptanz reduzieren.

24 Zu beachten ist, dass Lösungsansätze zu spezifischen rechtlichen Fragen wie der Unfallhaftung sich nur eingeschränkt verallgemeinern lassen, da sie sich zumeist auf spezifische staatliche Rechtsrahmen beziehen, wie beispielsweise die englischsprachige Literatur auf die Rechtsordnung der USA. In Bezug auf die deutsche Rechtsordnung sind hier vor allem die Ansätze von Hilgendorf (2018a, 2018b, 2019) einschlägig.

fördern (vgl. Verband der Automobilindustrie (VDA) e.V., 2022a). Als erste Einsatzszenarien sind u. a. ein Shuttlebetrieb, z. B. zur Aufwertung des ÖPNV in städtischen Randbereichen oder im ländlichen Raum, Betriebsfahrten (*Hub2Hub*), Logistik oder *Dual-Mode*-Fahrzeuge wie beim *Automated Valet Parking* vorgesehen (vgl. Bundesministerium für Digitales und Verkehr, 2021). Neben Regelungen für technische Anforderungen an das Fahrzeug, Kriterien für die Erteilung der Betriebserlaubnis und den Umgang mit Daten sowie Pflichten von Haltern und herstellenden Unternehmen führt das neue Gesetz die sogenannte »Technische Aufsicht« als menschliche Kontrollinstanz ein (Art. 1, § 1d, Abs. 3). Dabei handelt es sich um eine natürliche Person, die das Fahrzeug deaktivieren oder Fahrmanöver freigeben kann und durch eine Haftpflichtversicherung abgesichert ist. Experten aus Politik und Praxis bewerten das Gesetz zwar als wichtigen Schritt in die richtige Richtung, kritisieren aber u. a. die auf den ÖPNV und die gewerbliche Nutzung beschränkte Anwendbarkeit, die fehlende Regelungsklarheit hinsichtlich Datensicherheit und Datenschutz sowie Bestimmungen bezüglich der Haftungsproblematik, die zu stark zu Gunsten der Hersteller ausfällt (vgl. Deutscher Bundestag, 2021; Fahrenholz, 2021).[25] Zudem werden viele ethische Fragen nur vage thematisiert (vgl. Kriebitz et al., 2022).

Ein Jahr später folgte die »Verordnung zur Regelung des Betriebs von Kraftfahrzeugen mit automatisierter und autonomer Fahrfunktion und zur Änderung straßenverkehrsrechtlicher Vorschriften« (Bundesministerium für Digitales und Verkehr, 2022b), die den nationalen Rechtsrahmen zum autonomen Fahren vorerst vervollständigte.[26] Gemäß ihrer Digitalstrategie ist es das proklamierte Ziel der amtierenden Bundesregierung, das autonome Fahren bis 2025 vom Pilotprojekt zu einem festen Bestandteil der Praxis alltäglicher Mobilität voranzubringen (vgl. Bundesministerium für Digitales und Verkehr, 2023, S. 20). Zu diesem Zweck sollen zukünftig weitere

25 Für Anmerkungen und eine Übersicht über offene juristische Fragen das neue Gesetz betreffend siehe Hilgendorf (2017a, S. 226–229, 2019, S. 359–361).

26 Diese befasst sich mit den Forderungen von Verbänden und Unternehmen nach konkreten Vorgaben für Genehmigungsverfahren, Normung, Zertifizierung und Standards (vgl. Verband der Automobilindustrie (VDA) e.V., 2022a), der Sicherstellung von Kompatibilität (vgl. KPMG LLP, Center for Automotive Research, 2012, S. 15) und der Überwindung zulassungs- und haftungsrechtlicher Hürden.

Fördergelder in die ganzheitliche Entwicklung investiert werden, um die Produktionskosten angesichts des zeitversetzt zu erwartenden Refinanzierungseffekts über den Markt zu kompensieren.

Eine zentrale Bedeutung bei der Regulierung der Verkehrsautomatisierung kommt dem im Mai 2024 von den 27 Mitgliedsstaaten der EU verabschiedeten und wenige Monate später durch die Veröffentlichung im Amtsblatt der EU in Kraft getretenen prominenten Rechtsakt zur umfassenden Regulierung Künstlicher Intelligenz zu. Der sogenannte *AI Act* stellt einen horizontalen Regulierungsrahmen bereit, der eine Risikoklassifizierung von KI-Systemen vornimmt und spezifische Transparenzpflichten für vier definierte Risikokategorien vorgibt, die anhand von Kriterien wie Fairness, Transparenz, Robustheit und Zuverlässigkeit festgelegt werden. Während Systeme mit inakzeptablen Risiken[27] fortan verboten sind, werden Hochrisikoanwendungen mit weitreichenden Auflagen bedacht. Unter diese Risikokategorie fallen Systeme, die »eine erhebliche Bedrohung für die Gesundheit, Sicherheit oder die Grundrechte der Einzelnen darstellen« (Hengl, 2024) können; dies gilt neben beispielsweise Systembauteilen in der kritischen Infrastruktur oder medizinischen Geräten auch für autonome Fahrzeuge.

Nun ist der *AI Act* keine explizite Regelung für die Automobilindustrie, betrifft diese aufgrund der umfangreichen Risikoauflagen aber in besonderem Maße. So wird zukünftig vorgeschrieben, dass KI-Systeme selbstfahrender Fahrzeuge strenge Test- und Validierungsverfahren durchlaufen müssen, um zugelassen zu werden. Weiterhin ist auch die KI, die beim Test und der Validierung der Fahrzeugsoftware zum Einsatz kommt, Gegenstand neuer Regularien. Als besonders herausfordernd gestaltet sich dabei der Umstand, dass die Validierung im Hinblick auf dynamische und unvorhersehbare Umgebungen erfolgen muss, um als sicher und zuverlässig gelten zu können.[28]

---

27 Zu diesen zählen etwa kognitive Verhaltensmanipulation, Emotionserkennung oder Sozialkreditsysteme (vgl. Köllner, 2024a).

28 Darüber hinaus sieht die neue Verordnung die Einrichtung einer zweistufigen Governance-Architektur vor. Dabei obliegt nationalen Behörden die Durchsetzung der Vorschriften für KI-Systeme und deren Beobachtung, während KI-Modelle mit allgemeinem Verwendungszweck (*General Purpose AI*, *GPAI*) auf EU-Ebene beaufsichtigt werden. Diese Modelle »können für vielfältige Aufgaben eingesetzt werden und bilden die Grundlage für viele KI-Systeme in

Ist der *AI Act* nun eher als Innovationshemmer oder Erfolgsfaktor für die Entwicklung selbstfahrender Fahrzeuge zu bewerten? Der VDA befürchtet zunächst ein Ausbremsen von Innovation aufgrund des neuen KI-Gesetzes, denn die Regelungen sollen mittelfristig in die EU-Verordnung für die Genehmigung von Kraftfahrzeugen übernommen werden. Insbesondere die Regularien zur Typgenehmigung und Marktüberwachung von Kraftfahrzeugen sollen gezielt ergänzt und somit sektoriell reguliert werden, was zunächst mit hohen bürokratischen Belastungen verbunden sein wird. Grundsätzlich sei dieser delegierte Rechtsakt jedoch zu befürworten, um den spezifischen Herausforderungen der Automobilbranche gerecht zu werden (vgl. Köllner, 2024a). Zunächst ist allerdings ein verlangsamter Markteintritt aufgrund von hohen Compliance-Anforderungen zu erwarten. Mittelfristig bringen diese jedoch die Branche voran: Verbindliche und klare Regelungen schaffen Vertrauen und Akzeptanz, Rechtssicherheit minimiert das unternehmerische Risiko und beflügelt die Innovationsbereitschaft, die Förderung ethischer Standards führt zu nachhaltigen und verantwortungsvollen Innovationen, die europäischen Unternehmen einen Wettbewerbsvorteil verschaffen könnten.

Als allgemeiner Orientierungssatz für die Evaluierung bestehender und zukünftiger Rechtsrahmen gilt: Diese müssen so gestaltet sein, dass die Entwicklung des hoch- und vollautomatisierten Fahrens nicht behindert, sondern vielmehr gefördert und begleitet wird, wobei jedoch keine Kompromisse bei der Sicherheit zugelassener Systeme gemacht werden dürfen. Zielkonflikte scheinen hierbei vorprogrammiert. Inwiefern dies Abstriche hinsichtlich der Erwartungen an autonome Fahrzeuge bedeutet, wird im folgenden Unterkapitel diskutiert.

---

der EU. Einige davon könnten systemische Risiken bergen, wenn sie sich als besonders leistungsfähig erweisen oder eine weite Verbreitung finden.« (Europäische Kommission, 2024, o. S.) Das Europäische Gremium für Künstliche Intelligenz (KI-Gremium) soll die EU-weite Zusammenarbeit koordinieren. Verstöße gegen die Regularien werden künftig anhand von prozentualen Geldbußen sanktioniert. Das Testen von KI-Systemen unter realen Bedingungen soll ohne größere Hürden möglich werden, sofern gewisse Schutzvorkehrungen gewährleistet sind (vgl. ebd.).

2.2.3 Zwischen Utopie und Dystopie: Die Ambivalenz des autonomen Fahrens

Die erwarteten positiven Wirkungen erheben die Automatisierung der Mobilität zu einem ›digitalen Heilsversprechen‹, das Lösungen für viele unserer gegenwärtigen gesellschaftlichen Problematiken suggeriert. Sie gilt als zentrales Puzzleteil auf dem Weg zur Verwirklichung einer sicheren, effizienten und inklusiven Mobilität. Doch wie realistisch sind diese Erwartungen eigentlich? Wie wird die Zukunft der Mobilität tatsächlich aussehen, in der (voll-)automatisierte Fahrzeuge unsere Straßen bevölkern?[29]

Seit einiger Zeit mahnen kritische Stimmen vermehrt an, dass das weitgehend positive Bild vom autonomen Fahren wesentliche Aspekte vernachlässigt und sich mittelfristig ambivalente Folgeeffekte einstellen könnten. In der Forschungsliteratur finden mögliche sekundäre und tertiäre Effekte bisher noch wenig Beachtung. Eine weniger optimistische Haltung gegenüber der Einführung autonomer Fahrsysteme gründet sich im Wesentlichen auf Erfahrungen aus früheren Reduktionsstrategien. So waren Versuche, die Umweltbelastung durch technische Innovationen in der Energieeffizienz zu reduzieren, zu Beginn stets vielversprechend, haben sich jedoch in der Realität nur marginal ausgewirkt bzw. letztlich sogar eine weitere Verschlechterung der Umweltbilanz herbeigeführt:

> We have made these mistakes before, even for automobiles. When the automobile was first introduced, two of its primary purposes were to increase safety and protect the environment. Horses were dangerous and pollutive. The automobile was supposed to address these two problems and make roadways safer and cleaner. However, driver error replaced horse error. Carbon dioxide replaced horse manure, and its nonvisible nature and lack of smell make it a more challenging source of pollution. (Gurney, 2022, S. 147)

In der Vergangenheit waren sogenannte Reboundeffekte – neben unbeabsichtigten Externalitäten – u. a. auf erhöhte Konsumanreize (vgl. Jackson, 2009, S. 62–63) zurückzuführen, die durch neue Technologien geschaffen wurden. Eine ähnlich kontraintuitive Dynamik

29 Siehe hierzu auch den Beitrag von Ryan (2020), der anhand einer Szenarienanalyse mögliche Treiber und Hemmnisse der Entwicklung selbstfahrender Fahrzeuge bis zum Jahr 2025 untersucht.

wird auch für das autonome Fahren befürchtet: Durch die komfortable Form der Mobilität entstehen Kauf- und Konsumanreize für Privathaushalte und andere potenzielle Nutzer, deren Zahl sich um mobilitätseingeschränkte Personen erweitert. Es muss daher realistischerweise davon ausgegangen werden, dass sich die Zahl der gefahrenen Fahrzeugkilometer oder gar der Gesamtfahrzeugbestand mit der Einführung des autonomen Fahrens erhöht (vgl. Gogoll & Müller, 2017, S. 685). Damit sind weitere negative Externalitäten assoziiert, wie erhöhte Staugefahr oder die Untergrabung der Komfortvorteile des öffentlichen Nahverkehrs, die Emissionen weiter in die Höhe treiben. Aus ökonomischer Sicht drohen finanzielle Einbußen für die Kommunen, wenn gebührenpflichtige Parkräume wegfallen, und weniger Schadensfälle lassen die Gehaltsbudgets von Versicherungen und Werkstätten schrumpfen (vgl. Anderson et al., 2016, S. 38–40).

Auch die gesellschaftlichen Effekte lassen sich, zumindest in mittelfristiger Perspektive, nicht ausschließlich positiv bewerten. So ist anzunehmen, dass zunächst vor allem finanzstarke Individuen von den Vorzügen autonomer Fahrzeuge profitieren werden, wodurch soziale Ungleichheiten nicht reduziert, sondern verstärkt werden. Gurney (2022, S. 148–153) erläutert weitere Beispiele möglicher unbeabsichtigter sekundärer Folgen: Es ist zu befürchten, dass im Fall reduzierter Unfallzahlen die Verfügbarkeit von Spenderorganen zurückgeht, wodurch sich die Zahl der an Organversagen Verstorbenen erhöhen wird. Negative Effekte sind auch für Industrien alternativer Mobilitätsformen zu erwarten, beispielsweise für die Luftfahrt oder den Schienenverkehr, denen autonome Fahrzeuge durch ihre Attraktivität den Rang abzulaufen drohen.

Weitere Fragezeichen im Hinblick auf eine allzu rosige Vorstellung von der automatisierten Mobilität bestehen im Hinblick auf deren primären Motivator, das Sicherheitspotenzial. Gemäß DIN 31000 beschreibt der Begriff der Sicherheit eine Sachlage, in der das Risiko das Grenzrisiko nicht übersteigt. Entscheidend für die Sicherheit in einer Fahrsituation ist die Gesamtleistungsfähigkeit des Systems, welches das Fahrzeug im jeweiligen Fall steuert. Dieses besteht bis einschließlich Level 2 aus Fahrer und Assistenz- bzw. Teilautomatisierungsfunktionen. Analysen von Unfalldaten belegen, dass Letztere menschliche Fahrfehler gut kompensieren können, wodurch ihnen zumindest in Routinesituationen ein allgemeiner Sicherheitszuwachs

gegenüber Fahrzeugen ohne Automatisierungsfunktionen zugesprochen werden kann. Fahrsysteme der Hoch- und Vollautomatisierung müssen sich in Bezug auf ihre Leistungsfähigkeit allerdings dem Vergleich mit der Fahrfähigkeit eines menschlichen Fahrers stellen, dessen ›Bestleistung‹ sie nicht nur erreichen, sondern für ein erhöhtes Sicherheitspotenzial sogar übertreffen müssen (vgl. Winkle, 2015, S. 373).[30] Kahn (2022) stellt die berechtigte Frage, ab wann autonome Fahrsysteme denn als ›besser‹ als menschliche Fahrer gelten können – und welcher Bewertungsmaßstab dabei angelegt werden soll. Wenn immer mehr Fahraufgaben automatisiert ablaufen, droht langfristig sogar der Verlust der menschlichen Fahrkompetenz (vgl. Sparrow & Howard, 2017, S. 208). Dieses ›Paradoxon der Automatisierung‹ offenbart einen Zielkonflikt zwischen Sicherheit und Effizienz bzw. Fahrkomfort, der Systeme auf Level 2 und 3 wiederum als besonders kritisch im Hinblick auf ihre Fahrsicherheit erscheinen lässt.[31]

Beim teilautomatisierten Fahren (Level 2), das eine Überwachung und ein potenzielles Eingreifen durch die fahrzeugführende Person noch erforderlich macht, kommt zusätzlich ein weiterer, spezifischer Aspekt hinzu: Je geringer der Umfang, in dem Personen in die Fahraufgaben involviert sind, desto eher tendieren sie zur Unaufmerksamkeit, was dazu führt, dass sie in Notsituationen zu langsam, gar nicht oder nicht angemessen reagieren (können) (vgl. Köllner, 2021; Sparrow & Howard, 2017, S. 207–208).[32] Ein reales Beispiel in diesem Kontext ist der tödliche Unfall mit einem automatisierten

30 Systeme der bedingten Automatisierung (Level 3) stellen hier einen Grenzfall dar: Fahrer müssen zwar in Notfällen eingreifen können, sind also theoretisch Teil des Fahrsystems, aber in allen anderen Fällen irrelevant für dessen Gesamtleistungsfähigkeit.

31 Das Phänomen der übermäßigen Abhängigkeit vor allem von Autopiloten ist in der Luft- und Schifffahrt ein etabliertes Risiko, das auftritt, wenn Kontrollsysteme automatisiert werden und menschliches Eingreifen nur noch in Notfällen notwendig ist. Problematisch ist dabei insbesondere, dass die Aufgabenteilung mit dem System häufig unklar ist, was die Aufgaben des Piloten nicht einfacher, sondern komplexer macht (vgl. Wolmar, 2018, S. 42–47).

32 Dies genügt insbesondere nicht dem Konzept der *Meaningful Human Control*, das seit Kurzem in der Forschungsliteratur zur Kontrollierbarkeit und Verantwortungsfähigkeit autonomer Systeme diskutiert wird: »The concept of MHC appeals to the intuition that when autonomous systems are deployed in unstructured, dynamic and potentially unpredictable environments, simply having a human agent involved at some point in the decisional chain [...] may not

UBER-Fahrzeug, der sich 2018 im amerikanischen Arizona ereignete und infolge des großen Medieninteresses traurige Berühmtheit erlangte (vgl. Lee, 2019; Levin & Wong, 2018): Abgelenkt durch ihr Smartphone nahm die Sicherheitsfahrerin eine Fußgängerin, die die Straße überqueren wollte, zu spät wahr. Wie Untersuchungen bestätigten, war in diesem Fall ein Softwarefehler bzw. eine Voreinstellung ursächlich dafür, dass das automatisierte Fahrzeug die Fußgängerin fälschlicherweise als irrelevantes Objekt einstufte (vgl. Beutnagel, 2018). Dennoch wäre die Sicherheitsfahrerin trotz des aktivierten Autopiloten dazu verpflichtet gewesen, die Fahrumgebung stetig zu überwachen, und hätte so womöglich die Fehleinschätzung des Systems überstimmen können.

Weiterhin ist auch die zeitliche Perspektive des erwarteten Sicherheitszuwachses fraglich. Einer der wesentlichen Skalierungsfaktoren für die Größenordnung der Sicherheitswirkung, die autonome Fahrzeuge entfalten, ist der Grad der Marktdurchdringung.[33] So wird geschätzt, dass bei einem Marktanteil von 10 % bereits 1100 Leben pro Jahr gerettet werden könnten, während diese Zahl bei 90 % Anteil auf 21.700 Leben steigt. Nun bedeutet ein Fahrzeugkauf jedoch eine große Investition für private Haushalte, weshalb die Erneuerungszyklen in der Regel recht lang sind (vgl. Altenburg et al., 2018, S. 2). Es ist daher anzunehmen, dass sich neue Automatisierungstechnologien nicht disruptiv, sondern nur langsam im Bestand durchsetzen werden. Frühestens ab 2050 ist eine mengenmäßige Durchdringung im Gesamtfahrzeugbestand zu erwarten, die nennenswerte Sicher-

be sufficient to prevent unwanted mistakes and so-called accountability gaps; human persons must maintain a role that is as prominent as possible.« (Mecacci & Santoni de Sio, 2020, S. 104) Bisher wurde es primär im Kontext autonomer Waffensysteme betrachtet, lässt sich aber auch auf das automatisierte Fahren anwenden (vgl. Santoni de Sio & van den Hoven, 2018; Santoni de Sio, 2021, S. 720).

33 Dieser Zusammenhang gilt analog für die beiden Aspekte Effizienz und Mobilitätsbedürfnisse: Je höher der Anteil automatisierter bzw. autonomer Fahrzeuge am Verkehrsaufkommen ist, desto größer ist auch der Anteil des effizient steuerbaren Verkehrsflusses und des umweltfreundlichen Fahrverhaltens am gesamten Verkehrsaufkommen. Die gesamte Zeitersparnis durch effizienten Verkehrsfluss, insbesondere Stauvermeidung, wird bei einem Marktanteil von 10 % auf 756 Millionen Stunden pro Jahr geschätzt, bei 90 % Marktanteil auf 2772 Millionen Stunden pro Jahr (vgl. Fagnant & Kockelman, 2015, S. 172–175). Je mehr automatisierte Fahrzeuge zugelassen werden, desto mehr Personen können von neuen bzw. wiedergewonnenen Mobilitätschancen profitieren.

heitseffekte mit sich bringt. Diese werden sich erwartungsgemäß zunächst in einer Reduzierung der Sachschäden äußern. Auf signifikante Auswirkungen auf die Zahl der Personenschäden wird man noch länger warten müssen, denn schwere Unfälle mit Todesfolge passieren besonders häufig auf Landstraßen, wo auf absehbare Zeit noch kaum Automatisierungsfunktionen greifen werden. Demnach kann realistischerweise erst langfristig – im Zeithorizont von frühestens dreißig Jahren – mit einem relevanten Sicherheitszuwachs gerechnet werden (vgl. ebd., S. 40–47). Ebenfalls fragwürdig ist, inwiefern die Vision einer vernetzten und zentral gesteuerten Mobilität ein realistisches Ziel darstellt, solange Sicherheitsbedenken in Bezug auf Datenschutz und Cyberattacken fortbestehen:

> Ob in Zukunft eine dem Bahn- und Luftverkehr entsprechende vollständige Vernetzung und zentrale Steuerung sämtlicher Kraftfahrzeuge im Kontext einer digitalen Verkehrsinfrastruktur möglich und sinnvoll sein wird, lässt sich heute nicht abschätzen. Eine vollständige Vernetzung und zentrale Steuerung sämtlicher Fahrzeuge im Kontext einer digitalen Verkehrsinfrastruktur ist ethisch bedenklich, wenn und soweit sie Risiken einer totalen Überwachung der Verkehrsteilnehmer und der Manipulation der Fahrzeugsteuerung nicht sicher auszuschließen vermag. (Di Fabio et al., 2017, Regel Nr. 13)

Für höherstufig automatisierte Systeme gilt ferner, dass es aufgrund ihrer eingeschränkten maschinellen Wahrnehmungsfähigkeit und daraus resultierender Fehleinschätzung von Situationen zu neuen, bislang unbekannten Unfallkonstellationen kommen kann, die in einem Verkehrsgeschehen ohne Automatisierung nicht auftreten würden (vgl. Brändle & Grunwald, 2019, S. 289). Dies wäre beispielsweise dann der Fall, wenn eine Programmierung auf Schadensminimierung dazu führt, dass andere Verkehrsteilnehmer die Risikoaversion autonomer Fahrsysteme ausnutzen und sich bewusst unachtsam und rücksichtslos verhalten (vgl. Millard-Ball, 2018, S. 10–11). Dieser Aspekt wirft die Frage auf, ob autonome Fahrsysteme letztendlich überhaupt einen Mehrwert liefern. Sofern sie tatsächlich ein signifikantes Sicherheitspotenzial besitzen, ist ihre flächendeckende Einführung streng genommen ein moralischer und gesellschaftlicher Imperativ (vgl. Hevelke & Nida-Rümelin, 2015b, S. 621; Nyholm,

2018c, S. 6).[34] Konsequent zu Ende gedacht würde ein solcher auch das Verbot konventioneller Fahrzeuge implizieren, die dem durch Fahrautomatisierung gesetzten Sicherheitsstandard nicht mehr gewachsen sind: »Once they are safer than human drivers when it comes to risks to 3rd parties, then it should be illegal to drive them: at that point human drivers will be the moral equivalent of drunk robots.« (Sparrow & Howard, 2017, S. 206) In diesem Sinne bekräftigen auch Müller und Gogoll (2020, S. 1550–1561), dass konventionelle Fahrzeuge ein ungerechtfertigtes Risiko für andere Verkehrsteilnehmer in sich bergen und daher aus moralischen Gründen verboten werden müssten: »To put it more drastically: In a nutshell, manual driving in a scenario in which autonomous cars are affordable is the moral equivalent of running around with a hand grenade for pleasure.«[35]

Tatsächlich stellt gerade der Mischverkehr eine große Herausforderung in puncto Sicherheit dar. Dies ist u. a. darauf zurückzuführen, dass sich der menschliche Fahrstil und die Funktionsweise eines Fahrroboters grundlegend unterscheiden:

> Self-driving cars have optimizing driving styles and are very strict rule followers. Human drivers, in contrast, are typically satisficers: They drive just well enough to satisfy their driving goals. And humans are more flexible in their attitudes to traffic rules. (Nyholm, 2018c, S. 7)

In der Folge kann es zu Problemen in der Abstimmung und Erwartungsbildung zwischen Mensch und Maschine kommen, aus denen eine erhöhte Unfallgefahr resultiert. So können autonome Fahrzeuge menschliches Fahrverhalten nur sehr schwer antizipieren, während ihre mangelnde Fähigkeit zu nonverbaler und informeller Kommunikation beispielsweise durch Gestik es wiederum menschlichen Fahrern erschwert, das Verhalten autonomer Fahrzeuge korrekt zu deuten (vgl. Färber, 2015, S. 137–143; Sparrow & Howard, 2017, S. 2011).

---

34 Ein ähnliches Argument verwendet John Harris in seiner berühmten *Survival Lottery* (1975).

35 Müller und Gogoll beziehen sich in ihrer Begründung auf ein Argument von Sven Ove Hansson (2003), der Risikoübertragungen genau dann als gerechtfertigt betrachtet, wenn diese ihrerseits Vorteile für alle Beteiligten mit sich bringen. Dieses Argument wird in Kap. 7.3.3.4 näher betrachtet.

Ein mögliches rigoroses Verbot konventioneller Fahrzeuge tangiert jedoch nicht zu vernachlässigende soziale und kulturelle Aspekte. So wird das eigenhändige Führen eines Fahrzeugs häufig als »Ausdruck eines besonderen Lebensgefühls […], in dem sich die Dokumentation des eigenen Status, Sehnsucht nach Freiheit, Freude an Sport und gelegentlich auch Lust auf Abenteuer verbinden« (Hilgendorf, 2017a, S. 225), angesehen. Entsprechende Restriktionen würden die individuelle Selbstverwirklichung und Autonomie empfindlich beschneiden (vgl. Borenstein et al., 2019, S. 392; Hansson et al., 2021, S. 1402; Moor, 2016). Wie die Ethik-Kommission (Di Fabio et al., 2017, S. 11, Regel 6) konstatiert, wäre ein solcher Eingriff in die Selbstbestimmung ethisch nicht vertretbar:

> Umgekehrt ist eine gesetzlich auferlegte Pflicht zur Nutzung vollautomatisierter Verkehrssysteme oder die Herbeiführung einer praktischen Unentrinnbarkeit ethisch bedenklich, wenn damit die Unterwerfung unter technische Imperative verbunden ist (Verbot der Degradierung des Subjekts zum bloßen Netzwerkelement).

Bisher sind einige alternative Ansätze vorgeschlagen worden, um ein striktes Verbot konventioneller Fahrzeuge zu umgehen und gleichzeitig negative Effekte des Mischverkehrs abzuschwächen. Müller und Gogoll (2020, S. 1563–1564) entwerfen das visionäre Szenario eines *AI-Supervised Human Driving*, bei dem der Mensch die dynamische Fahraufgabe ausführt, dabei jedoch von einem intelligenten künstlichen System überwacht wird, das im Notfall den Fahrer überstimmen kann. Hingegen plädieren Nyholm und Smids (2020, S. 339–340) dafür, bestimmte Aspekte des menschlichen Fahrens an eine robotergestützte Fahrweise anzupassen, um auf diese Weise einerseits die Koordinationsprobleme zwischen Mensch und Maschine zu überwinden und andererseits potenzielle Risikoquellen für menschliche Fahrfehler einzudämmen.[36] Praktisch umsetzbar wäre dies beispielsweise über restriktive Anpassungen der Verkehrsvorschriften oder mithilfe bestimmter Technologien, z. B. durch geschwindigkeitsregulierende Komponenten, Frühwarnsysteme bei erhöhter Kollisionsgefahr oder Alkohol-Interlocks. Roy (2016, o. S.) argumentiert in eine ähnliche Richtung: »[…] if autonomous cars

36 Nyholm und Smids (2020, S. 338–339) liefern verschiedene Argumente, weshalb eine umgekehrte Anpassung automatisierten Fahrverhaltens an das menschliche wenig zielführend ist.

can set a certain, provable safety standard, it might make sense that the licensing requirements would therefore require a human driver to prove a similar competency—which raises the safety bar for human-driven cars, as well.«

Schlussendlich verbleiben mögliche Effekte, die sich heute noch gar nicht absehen lassen. Eine gewisse Ambivalenz ihrer Wirkungen ist innovativen Technologien inhärent; sie können sowohl unbeabsichtigte positive als auch negative Konsequenzen nach sich ziehen und langfristig in Weisen wirken, die nicht vorhersehbar sind (vgl. Lin, 2013a). Dieser unter dem Begriff ›Schmetterlingseffekt‹ (*butterfly effect*) bekannte Mechanismus erschwert es, zum aktuellen Entwicklungszeitpunkt eine umfassende Bewertung des automatisierten Fahrens in langfristiger Perspektive vorzunehmen. Der Schwerpunkt aktueller Bemühungen sollte daher primär auf den generellen Wirkungen der Fahrautomatisierung liegen, die kurz- und mittelfristig zu erwarten sind – und auf Maßnahmen, die dazu dienen, mögliche negative Effekte so gut wie möglich zu antizipieren und abzuschwächen. Um dieses Ziel zu erreichen, ist eine Auseinandersetzung mit ethischen Fragen rund um autonome Fahrzeuge unerlässlich. Im folgenden dritten Kapitel wird nun zunächst der bestehende ethische Diskurs rekonstruiert, wobei insbesondere moralische Dilemma-Szenarien in den Blick genommen werden.

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 ausweglöser 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.

II.
Problemzugänge in zwei Diskursen: Darstellung und Kritik

In diesem zweiten Teil des Buches wird, im Anschluss an die einschlägigen Diskurse in der Angewandten Ethik bzw. Technikethik einerseits und der Metaethik andererseits, in zwei Kapiteln das zentrale Argument entwickelt, welches die erste These der Forschungsarbeit begründet: Ein Zugang zur Gestaltung von Unfallalgorithmen, der weiten Teilen des einschlägigen Forschungsdiskurses zugrunde liegt, weist sowohl methodische als auch inhaltliche und problemstrukturelle Schwächen auf, die (zu) viele Fragen offenlassen.

In Kap. 4 wird zunächst das zweite Teilziel erarbeitet, indem dominante Ansätze bisheriger Forschung, welche die Programmierung von Unfallalgorithmen als Problematik moralischer Designentscheidungen begreifen, anhand relevanter Literatur rekonstruiert und einer kritischen Betrachtung unterzogen werden. Eine differenzierte Auseinandersetzung wird zeigen, dass eine (ausschließlich) moralphilosophische Konzeption von Unfallalgorithmen mit zahlreichen Schwierigkeiten behaftet ist, die alternative Zugänge erforderlich machen.

Der argumentative Gedankengang des vierten Kapitels lässt sich im Einzelnen wie folgt wiedergeben: Als Hinführung zur Thematik wird in Kap. 4.1 skizziert, weshalb sich eine algorithmische Ent-

scheidungsfindung in dilemmatischen Fahrsituationen durch eine dezidiert ethische Dimension auszeichnet. Unter Bezugnahme auf den einschlägigen maschinenethischen Diskurs werden spezifische Herausforderungen bei der Implementierung maschineller Moral konkretisiert, wobei insbesondere auf Methoden maschinellen Lernens eingegangen wird. Im Anschluss an eine systematische Problematisierung repräsentativer Dilemma-Szenarien wird thematisiert, inwiefern diese sich plausibel als Instanzen eines modifizierten, angewandten Trolley-Problems darstellen lassen. In Kap. 4.2 wird sich sodann dem lebensweltlichen Kontext zugewandt, in dem das Anwendungsproblem zu verorten ist. Es wird zunächst die gesellschaftlich-soziale Tragweite von Unfalldilemmata ergründet, bevor die spezifische Problematik ihrer Regulierung in pluralistisch geprägten Gesellschaften diskutiert wird. Zudem wird die These vertreten, dass Designstrategien für Unfallalgorithmen notwendigerweise als Entscheidungen unter Risiko zu konzipieren sind. In den nachfolgenden beiden Unterkapiteln werden schließlich konkrete Ansätze evaluiert, die im Rahmen des dominanten Forschungszugangs vorgeschlagen wurden: Zunächst werden in Kap. 4.3 deskriptive Ansätze kritisch beleuchtet, die mittels Methoden experimenteller Ethik versuchen, sich dem Anwendungsproblem anzunähern. In Kap. 4.4 wird nachgewiesen, dass normative Begründungsversuche aus der philosophischen Ethik insbesondere hinsichtlich der konkreten Implementierung eines Prinzips der Schadensminimierung an strukturelle und praktische Grenzen stoßen. Abschließend werden in Kap. 4.5 die Ergebnisse des vierten Kapitels in einer zentralen Schlussfolgerung zusammengeführt (zweites Zwischenergebnis).

Kap. 5 ist der Erarbeitung des dritten Teilziels der Forschungsarbeit gewidmet. Dieses besteht darin, die Betrachtung des praktischen Anwendungsproblems um eine theoretisch-formale Komponente zu erweitern, indem die zugrundeliegende Problemstellung moralischer Dilemmata aus metaethischer Perspektive erörtert wird. Eine Vorgehensweise, die anwendungsorientierte Aspekte mit formaler Analyse verbindet, wird im Rahmen dieser Arbeit als essenziell betrachtet, um eine differenziertere Perspektive auf den Problemkomplex entwickeln zu können, als sie bisher im Rahmen des dominanten Forschungszugangs vorhanden ist. In diesem Sinne wird anhand einer metaethischen Rekonstruktion der relevanten Wertekonflikte, die in Dilemma-Szenarien des autonomen Fahrens zum Tragen kommen,

eine ganzheitliche Problembetrachtung ermöglicht, die das in Kap. 4 entwickelte Argument zur Verifizierung der ersten These ergänzt. Es wird gezeigt, dass nicht nur im Hinblick auf (praktische) Aspekte des Anwendungskontextes essenzielle Fragen offenbleiben, sondern auch hinsichtlich der praktischen und theoretisch-formalen Implikationen, die sich aus der metaethischen Struktur des zugrundeliegenden Entscheidungsproblems ergeben. Gleichzeitig wird durch das Eruieren spezifischer Charakteristika von Unfalldilemmata die Grundlage für einen pragmatisch orientierten,[77] risikoethischen Ansatz gelegt, der im dritten Teil des Buches mit Blick auf die zweite These diskutiert wird. Damit liefert das fünfte Kapitel kein eigenes Argument im engeren Sinne, sondern fungiert quasi als Bindeglied zwischen den beiden zentralen Thesen der Arbeit.

Die Struktur des fünften Kapitels gliedert sich wie folgt: Zunächst wird in Kap. 5.1 in die Thematik eingeführt, indem Entscheidungsdilemmata als Grenzsituationen moralischen Handelns im Lichte von tradierten Beispielen und Narrativen geschildert werden. Als Einstieg in die abstrakt-formale Betrachtung von Dilemma-Strukturen werden Kriterien für das Vorliegen echter Dilemmata erörtert und zu einer anspruchsvollen Definition zusammengeführt. Daraufhin erfolgt in Kap. 5.2 eine kursorische Darstellung der zentralen Argumente der einschlägigen metaethischen Debatte, die sich mit der Frage beschäftigt, inwiefern echte Dilemmata überhaupt möglich sind bzw. ob sie tatsächlich existieren. In Kap. 5.3 wird sodann untersucht, welche definitorische Rolle die Unlösbarkeit von Konfliktsituationen spielt und welche Entscheidungsstrategien sich daraus im metaethischen Dilemma-Diskurs ergeben. Zudem wird die Inkommensurabilität spezifischer moralischer Werte als zentrales Argument zur Begründung der Abwesenheit systematischer Ansätze zur Entscheidung derartiger Dilemma-Situationen diskutiert. Im Zuge des Versuchs einer Erklärung, weshalb Akteure angesichts von moralischen Dilemmata unvermeidlich scheitern, werden zwei Konzeptionen erläutert. Diese argumentieren mittels der Nicht-Einlösbarkeit spezifischer Werte bzw. der Nicht-Verhandelbarkeit entsprechender moralischer Gebote. Schließlich wird in Kap. 5.4 die metae-

77 ›Pragmatisch‹ wird im Kontext dieser Forschungsarbeit v. a. im Anschluss an Habermas' Verwendung des Begriffs im Sinne einer wertorientierten Zweckrationalität verstanden, siehe Kap. 5.4.2.2.

thische Ebene zugunsten einer stärkeren Anwendungsorientierung verlassen. Vor dem Hintergrund der zuvor erlangten Erkenntnisse wird begründet, warum eine pragmatische Herangehensweise an mögliche Entscheidungsstrategien für Unfalldilemmata vielversprechend erscheint (drittes Zwischenergebnis), wie sie unter Bezugnahme auf risikoethische Konzepte im dritten Teil des Buches herausgearbeitet wird.

4. Unfallalgorithmen als moralisches Designproblem: Rekonstruktion und Kritik

4.1 Entscheidungsalgorithmen, Dilemma-Szenarien und vermeintliche Trolley-Analogien

4.1.1 Die ethische Dimension von Entscheidungsalgorithmen

Algorithmen treffen Entscheidungen grundsätzlich anders als Menschen; das gilt sowohl für Entscheidungen im Allgemeinen als auch für solche ethischer Natur. Für eine fundierte ethische Auseinandersetzung mit dilemmatischen Entscheidungssituationen, denen sich künstliche Systeme gegenübersehen, ist zumindest ein Grundverständnis dessen notwendig, was im spezifischen Anwendungskontext eine algorithmische von einer menschlichen Entscheidungsfindung unterscheidet. Zentrale Kriterien werden im Folgenden kurz dargestellt.

In kritischen Unfallsituationen, die im dynamischen Verkehrsgeschehen meist plötzlich und unvorhergesehen auftreten, reagieren menschliche Fahrer gewöhnlicherweise spontan und intuitiv in Sekundenbruchteilen (vgl. Dilich et al., 2002, S. 239–240). Da sie weder in der Lage sind, Alternativen zu prüfen, noch überlegte Entscheidungen anhand moralischer Kriterien zu treffen, kann in diesen Fällen eigentlich nicht von Situationen moralischen Entscheidens bzw. Handelns gesprochen werden. Letztere entstehen erst dadurch, dass die Umstände eine (moralisch) begründete Entscheidungsfindung erlauben; erst durch die Möglichkeit einer überlegten Reaktion werden unvermeidbare Unfallsituationen zu einer moralischen Problemstellung.[78] Wenn algorithmische Fahrentscheidungen durch Fahrroboter ausgeführt werden, sind Menschen jedoch weiterhin

78 Die Diskrepanzen, die sich zwischen konkreten Entscheidungen in Situationen mit unterschiedlichen zeitlichen Restriktionen ergeben, untersuchen Lucifora et al. (2021) anhand eines Laborexperiments.

wesentlich beteiligt. Menschliche Entscheidungen werden sozusagen vorverlagert, indem sie zum Zeitpunkt der Programmierung die ethischen Kriterien vorgeben, an denen sich das System in konkreten Dilemma-Situationen orientieren soll. Weber und Zoglauer (2019, S. 158) präzisieren in diesem Sinne, dass es streng genommen nicht die autonomen Fahrzeuge sind, die sich im Dilemma befinden, sondern die Menschen, die Designentscheidungen über die Gestaltung von Unfallalgorithmen treffen.

Dabei ist die Entscheidungssituation jedoch grundlegend anders als in Unfallszenarien mit konventionellen Fahrzeugen; moralisch relevant sind vor allem drei wesentliche Unterschiede. Erstens verfügt ein autonomes Fahrzeug zur Laufzeit über mehr Informationen über seine Umgebung und kann diese wesentlich schneller verarbeiten als ein menschlicher Fahrer (vgl. Nyholm & Smids, 2016, S. 1278). Aufgrund der verkürzten Reaktionszeit sowie der sensorgestützten räumlichen Umfeldwahrnehmung stehen einem autonomen Fahrsystem in Notsituationen mehr Optionen zur Verfügung als einem Menschen, dessen Sicht z. B. auf den nachfolgenden Verkehr naturgemäß eingeschränkt ist. Es kann daher angenommen werden, dass einige Dilemma-Situationen durch den Einsatz autonomer Fahrsysteme erst entstehen. So würde sich das folgende Entscheidungsdilemma nur stellen, wenn der Motorradfahrer im Rückspiegel rechtzeitig wahrgenommen wird, was für ein mit Sensoren ausgestattetes selbstfahrendes Fahrzeug eher realisierbar ist als für einen menschlichen Fahrer:

> A self-driving car finds itself with the following dilemma: either it brakes to avoid running over a careless pedestrian who crosses the road suddenly, but the motorcyclist behind who is following too closely will die in the crash against the rear window; or the car does not brake and runs over the pedestrian but saves the life of the motorcyclist behind. (Coca-Vila, 2018, S. 62)

Zweitens wird die Entscheidung darüber, welche Handlung ein autonomes Fahrzeug im konkreten Dilemma-Fall ausführen soll, antizipatorisch lange vor dem Zeitpunkt getroffen und implementiert, in dem sich die entsprechende lebensweltliche Situation potenziell manifestiert (vgl. Brändle & Grunwald, 2019, S. 286; Faulhaber et al., 2019, S. 400; Hevelke & Nida-Rümelin, 2015c, S. 8–9; Nyholm & Smids, 2016, S. 1280–1281). Die Entscheidungsträger sind dabei frei von situativem, unmittelbarem Handlungsdruck und psycholo-

gischem Stress, sie können moralische Argumente prüfen und zu einer überlegten Entscheidung gelangen. Goodall (2016a, S. 813) beschreibt dies wie folgt: »This may have been an instinctual response from the driver, but in the days of vehicle automation, instinct will be replaced by decisions and logic encoded in software, sometimes programmed years before the crash.« Eine Programmierung, die sich rein am intuitiven menschlichen Reaktionsverhalten orientiert, wäre deshalb ethisch nur schwer zu rechtfertigen:

> But the programmer and OEM do not operate under the sanctuary of reasonable instincts; they make potentially life-and-death decisions under no truly urgent time-constraint and therefore incur the responsibility of making better decisions than human drivers reacting reflexively in surprise situations. (Lin, 2015, S. 75)

Drittens zeichnen sich die Aktionen von Fahrrobotern aufgrund der Implementierung vorab festgelegter Kriterien durch eine generelle Wiederholbarkeit bei gleichen Eingangsdaten (vgl. Siegel & Pappas, 2023, S. 218) – und damit Konsistenz auch in ethischer Hinsicht – aus. Allerdings sind ihre Handlungsmuster auch systematisch. In der Folge fallen individuelle, unabhängige Entscheidungen weg; Instinkt und Impulsivität werden durch eine Algorithmenlogik ersetzt, die zu einer systemischen Verzerrung gemeinhin akzeptierter Risiken des Straßenverkehrs führen kann. Diese Algorithmenlogik kommt nicht nur in einer einzigen, sondern einer Vielzahl von ähnlichen Situationen zur Anwendung (vgl. Himmelreich, 2018, S. 678). Für die Programmierung ethischer Unfallalgorithmen ergibt sich daher eine besondere Verantwortung hinsichtlich kontrollierter und aggregierter Effekte:

> But our technologies are powerful: they give us increasing omniscience and control to bring order to the chaos. When we introduce control to what used to be only instinctive or random – when we put God in the machine – we create new responsibility for ourselves to get it right. (Lin, 2014b, o. S.)

Dies gilt insbesondere für das Problem, dass zentralisierte und algorithmische Entscheidungsprozesse verzerrt sein und z. B. diskriminierende Effekte hervorrufen können:

> [...] the prospect for the same algorithmic preferences controlling the vehicles to be replicated across any number of such vehicles leads to the possibility for identical responses that are governed by the same

> rule-structure. This in turn creates a systemic and collective dimension whereby the generated outcomes will be reliably and systematically skewed according to the coded preferences, whether intentional or not. The crucial differentiator is thus the removal of hitherto discrete and independent actions undertaken by individuals and the range and diversity of available responses that flow as a result. The subsequent harmonization in accumulating these responses skews together results in systematic biases in relation to certain sets of characteristics. If the preferred or penalized preferences map onto individual or group characteristics for which discriminating based on those characteristics is impermissible, these structured biases have been translated into systematic discrimination. (Liu, 2018, S. 160–161)

Die jeweiligen Entscheidungssituationen im Kontext autonomer und konventioneller Fahrzeuge unterscheiden sich demnach wesentlich in problemstruktureller und epistemischer Hinsicht: Aus einer intuitiv-situativen Reaktion wird eine überlegte, bewusste Entscheidung (vgl. Lin, 2015, S. 74; Nyholm & Smids, 2016, S. 1278–1279), die es notwendig macht, spezifische Entscheidungsdilemmata nicht nur ethisch, sondern auch juristisch neu zu bewerten. So schlussfolgern Dilich et al. (2002, S. 246), dass das Resultat von Notsituationen weniger von den Fahrfähigkeiten des Fahrers abhängt als vom Zufall der Umstände. Folgerichtig ist ein rechtswidriges Verhalten wie die Tötung eines Menschen nicht zwangsläufig auch als schuldhaft anzusehen (vgl. Contissa et al., 2017, S. 368); die entsprechende Person ist zwar juristisch verantwortlich, aber nicht moralisch schuldig. Diese in der bisherigen Rechtsprechung weitgehend etablierte Diskrepanz zwischen gesetzlichen und moralischen Wertungen ist im Fall von Algorithmen nur begrenzt anwendbar, wie die Ethik-Kommission (Di Fabio et al., 2017, S. 11, Regel Nr. 8) anmerkt: »Derartige in der Rückschau angestellte und besondere Umstände würdigende Urteile des Rechts lassen sich nicht ohne weiteres in abstrakt-generelle Ex-Ante-Beurteilungen und damit auch nicht in entsprechende Programmierungen umwandeln.« Algorithmen können sich nicht auf psychologische Stressfaktoren oder andere besondere Umstände der Entscheidungssituation berufen; ihre Reaktionen sind systematisch und (moralisch) unentschuldbar (vgl. Birnbacher & Birnbacher, 2016, S. 8; Lin, 2013b; Trappl, 2016, S. 745–746).

Hinzu kommt, dass das deutsche Strafrecht auf einer normativen Unterscheidung fußt, die zwischen aktiver Verursachung von Schä-

den einerseits und passiver Schädigung durch Unterlassen einer Handlung andererseits differenziert. Im Fall konventioneller Autos würde lediglich das bewusste Ändern einer bereits aktivierten Trajektorie, z. B. durch ein Ausweichmanöver, als aktiver Akt und daher strafrechtlich haftbar gelten. Bei autonomen Fahrsystemen hingegen ist diese Differenzierung wenig sinnvoll, denn ein Fahrroboter hat keine eigenen Absichten. Vielmehr ist jede automatisierte Trajektorienwahl – das Spurhalten wie auch das Ausweichen – eine bewusste, intentionale Designentscheidung menschlicher Entscheidungsträger (vgl. Contissa et al., 2017, S. 368); es gibt für selbstfahrende Fahrzeuge kein (passives) Standardverhalten (vgl. Birnbacher & Birnbacher, 2016, S. 14). Inwiefern sich daraus juristische und moralische Probleme ergeben, wird besonders deutlich im Hinblick auf die absichtliche Auswahl von Zielobjekten (vgl. Lin, 2015, S. 72–73). Für die Optimierung des Unfallverhaltens müssten Kostenfunktionen implementiert werden, die bereits zum Zeitpunkt der Programmierung festlegen, welche Ziele das Fahrzeug angesichts einer unvermeidbaren Kollision ansteuern soll. Aus juristischer Sicht würde dies mit vorsätzlichem Töten gleichgesetzt.[79]

Unfallalgorithmen weisen also eine hohe ethische Relevanz auf. Wie können selbstfahrende Fahrzeuge dieser Bürde gerecht werden? Die Aufgabe, Maschinen in moralischen Situationen zu vertretbaren Aktionen zu befähigen, tangiert grundlegende maschinenethische Fragen: Sind Maschinen prinzipiell handlungsfähig? Und falls nicht, wie können sie dennoch in moralischen Entscheidungssituationen bestehen? Im folgenden Unterkapitel werden konzeptionelle Ent-

79 Eine Gegenposition hierzu bezieht Gasser (2015, S. 557), der betont, dass implementierte Entscheidungen noch immer so allgemein sind, dass sich in Bezug auf die Grundrechte kein relevanter Unterschied zwischen maschineller und menschlicher Fahrzeugsteuerung ergibt: »Zwar ist […] jede Steuerungsentscheidung durch die Programmierung des entsprechenden Systems unter bestimmten Randbedingungen vorgegeben und somit letztlich nicht zufällig, allerdings handelt es sich dabei gerade nicht um die Konkretisierung eines *bestimmten* Handlungsablaufes. Die Programmierung einer autonomen Fahrfunktion gibt vielmehr (nur) vor, welche Gesichtspunkte zu berücksichtigen sind, sodass hieraus unter mehreren Alternativen diejenige gewählt werden kann, die einen Schaden nach Möglichkeit ganz vermeidet oder den geringsten Schaden verursacht. […] Damit werden aber im Rahmen der Programmierung keine Steuerungsentscheidungen getroffen, sondern (nur) abstrakte Kriterien für die einzelfallbezogene Steuerungsentscheidung vorgegeben.«

würfe und Herausforderungen erläutert, die sich im Hinblick auf die Implementierung maschineller Moral ergeben.

4.1.2 Maschinelle Moral, kontextsensitive Systeme und maschinelles Lernen

Autonome Systeme werden auf unseren Straßen unweigerlich mit moralisch relevanten Entscheidungssituationen konfrontiert werden. Doch inwiefern sind Maschinen überhaupt in der Lage, diesen angemessen zu begegnen? Die Gestaltung eines möglichen Zusammenlebens und -wirkens von Menschen und Robotern zählt zu den epochalen interdisziplinären Fragestellungen der Gegenwart.[80] Die maschinenethische Forschung geht dabei traditionell der Frage nach, inwiefern Roboter und autonome Systeme als Subjekte moralischen Handelns – sogenannte *Artificial Moral Agents* (*AMAs*) – angesehen werden können. Dabei basiert die üblicherweise referenzierte Vorstellung künstlicher Moralität auf einem reduzierten Moralverständnis, welches den Anspruch hat, lediglich bestimmte Grundzüge menschlicher Moral nachzubilden, wie beispielsweise das Befolgen bestimmter Prinzipien (vgl. Bendel, 2018, S. 35). Auf der Grundlage von James H. Moors (2006, S. 19–21) bedeutendem hierarchischen Schema zur Klassifikation moralischer Akteure liegt der Fokus gegenwärtiger Diskurse in der Maschinenethik vor allem auf der Untersuchung einer expliziten ethischen Handlungsfähigkeit (*explicit ethical agency*); entsprechende Systeme sollen in der Lage sein, ethische Urteile anhand explizit implementierter ethischer Regeln zu fällen und zu begründen.

Doch werden sie damit schon zu handelnden Akteuren? In Übereinstimmung mit Moors Ansatz zeigen zahlreiche maschinenethische Positionen, dass künstliche Systeme grundsätzlich nicht an die Komplexität einer vollwertigen ethischen Handlungsfähigkeit (*full ethical agency*) heranreichen. Relevante Argumentationen beziehen sich neben der Fähigkeit zu ethischem Urteilen, Reflektieren und

80 Eine viel beachtete systematische Abhandlung über das Verhältnis von Mensch und Roboter und damit verbundene ethische Fragestellungen stammt von Nyholm (2020a). Eine Sammlung verschiedener maschinenethischer Forschungsbeiträge, die einen Überblick über die Vielschichtigkeit des Forschungsfelds und seiner Themenstellungen liefern, wurde von Rath et al. (2019) herausgegeben.

Begründen vor allem auf das Fehlen innerer Zustände, kognitiver Kapazitäten sowie metaphysischer Eigenschaften wie (phänomenales) Bewusstsein, Intentionalität und Willensfreiheit.[81] Diese gelten als Voraussetzungen dafür, dass Akteure für ihre Handlungen moralisch verantwortlich sein können (vgl. Misselhorn, 2018b, S. 123–126; Moor, 2006, S. 20–21; Searle, 1980, S. 450–454).[82] Moor (2006) spricht in diesem Kontext von einer unüberwindbaren ontologischen Differenz zwischen Mensch und Maschine:

> Many believe a bright line exists between the senses of machine ethics discussed so far and a full ethical agent. For them, a machine can't cross this line. The bright line marks a crucial ontological difference between humans and whatever machines might be in the future. (Ebd., S. 20)

Auch Kamm (2020, S. 89–91) weist darauf hin, dass es Prinzipien gibt, die für Menschen Gültigkeit besitzen, möglicherweise aber nicht für Maschinen; diese können weder aus Gründen handeln noch verfügen sie über eine akteurszentrierte Sicht oder eine emotionale Beziehung zu ihrem eigenen Verhalten. Vor diesem theoretischen Hintergrund folgern auch Hevelke und Nida-Rümelin (2015c) aus anwendungsnaher Perspektive, dass Handlungen sich im Kern durch intentional gesteuertes Verhalten auszeichnen:

> Handeln ist ein von Intentionen motiviertes und kontrolliertes Verhalten. Nur Bereiche des Verhaltens, die intentional kontrolliert und motiviert sind, haben den Status von Handlungen. Handlungen sind dementsprechend die Bereiche menschlichen Verhaltens, für die wir verantwortlich sind, da sie unserer intentionalen Kontrolle unterliegen. Diese wird aber (zumindest bei halbwegs vernünftigen Akteuren) von

81 In einem neueren Beitrag stellt beispielsweise Véliz (2021) die mangelnde Empfindungsfähigkeit algorithmischer Systeme heraus, die sie plakativ als »a kind of functional moral zombie« (ebd., S. 487) bezeichnet.

82 Eine Gegenposition vertreten Floridi und Sanders (2004, S. 366–376). Sie argumentieren, dass Maschinen zwar nicht verantwortlich im Sinne von *responsibility*, aber im Sinne von *accountability* (jemand ist die Quelle eines moralisch schlechten Ereignisses) sein können. Dafür müssen moralische Akteure nicht notwendigerweise einen freien Willen oder mentale Zustände besitzen. Eine Zusammenstellung verschiedener Beiträge zur Verantwortungsproblematik aus maschinenethischer Sicht bietet der Sammelband von Rath et al. (2019, Teil II). Insbesondere Wölm (2019, S. 183–188) geht in seinem Aufsatz auf die Möglichkeit einer geteilten Verantwortlichkeit ein, die er anhand einer Priorisierung der technischen über die ethische Vollkommenheit als Endziel der Entwicklung autonomer Fahrsysteme skizziert.

> Gründen geleitet oder zumindest beeinflusst. […] Solange ein Mensch keine Kontrolle über sein Verhalten hat (er etwa schlafwandelt) macht eine moralische Bewertung seines Verhaltens ebenso wenig Sinn, wie wenn ihm grundsätzlich die Fähigkeit abgeht, moralische Gründe zu verstehen und sich von ihnen affizieren zu lassen. Solange autonome Fahrzeuge schlicht ihrer Programmierung folgen und nicht in der Lage sind, Überzeugungen auszubilden und sich dabei von Gründen beeinflussen zu lassen, macht es keinen Sinn, sie als moralische Akteure wahrzunehmen. (Ebd., S. 9)

In der Konsequenz stimmt die zeitgenössische maschinenethische Forschung um die Pioniere der frühen Maschinenethik wie Anderson und Anderson (2011) oder Wallach und Allen (Allen et al., 2005; Wallach & Allen, 2008) bzw. im deutschsprachigen Raum Misselhorn (2018a, 2018b, 2019) und Bendel (2016, 2018, 2019) weitgehend überein, dass Maschinen keine moralischen Agenten im Sinne handelnder und entscheidender Subjekte sind. Vielmehr sind sie lediglich *beschränkte* moralische Akteure, die Aspekte menschlicher Moral simulieren. Begründet wird dies mehrheitlich mit der Diskrepanz zwischen maschineller und moralischer Autonomie: Nur weil eine Maschine mechanisch in der Lage ist, ohne menschliches Eingreifen die ihr vorgegebenen Aufgaben zu erfüllen, bedeutet das noch nicht, dass sie dies aufgrund eigener moralischer Überzeugungen (vgl. Lucas Jr., 2015, S. 2871–2872) bzw. eigenständig durch Selbstreflexion gewonnener ethischer Kompetenz (vgl. Miller et al., 2017, S. 392–400) tut. Anders ausgedrückt:

> Selbst wenn programmierte Maschinen, die zunehmend unseren Alltag (mit-) bestimmen, den Eindruck erwecken, dass sie moralische Agenten seien, folgen sie doch nur den durch Menschen vorgegebenen Regeln und können sich grundsätzlich nicht von diesen befreien. Im besten Fall wurden diese Regeln wohlbedacht, im schlechtesten Fall spiegeln sie die Vorurteile und normativen Schwächen ihrer Schöpfer wider. Bisher gibt es aber keinen Anlass, davon zu sprechen, dass Maschinen selbst ein moralisches Urteil gefällt hätten. Existierende Maschinen sind keine moralische Agenten und entwickeln keine eigene Moral. (Weber & Zoglauer, 2019, S. 159)

Aus der weitreichenden moralischen Handlungsunfähigkeit von Maschinen ergeben sich schließlich bedeutende Implikationen für die Entwicklung autonomer Fahrzeuge. Zum einen folgt in direkter Konsequenz aus der maschinenethischen Auseinandersetzung, dass

die Nutzer autonomer Fahrsysteme zwingend eine (Teil-)Verantwortung tragen (vgl. Wölm, 2019, S. 179–182). Zum anderen müssen die Systeme entweder entsprechend programmiert oder via Techniken maschinellen Lernens trainiert werden, um in moralischen Konfliktsituationen bestehen zu können. Die Gestaltung von Unfallalgorithmen wird im laufenden Diskurs daher überwiegend als ethisch geleitete *Designproblematik* aufgefasst, die Entwürfe *maschineller Moral* fokussiert: einer Moral, die in der Maschine wirkt, und die Frage in den Blickpunkt rückt, welche spezifischen ethischen Prinzipien in implementierten Entscheidungsalgorithmen zur Anwendung kommen sollen.[83]

Die Maschinenethik kennt gegenwärtig drei dominante Konzepte zur Entwicklung moralischer Maschinen,[84] die auf unterschiedlichen Heuristiken für die Festlegung von Entscheidungsnormen basieren. Beim sogenannten *Top-Down*-Ansatz werden explizit formulierte, normative Prinzipien – wie beispielsweise Kants kategorischer Imperativ oder das utilitaristische Nutzenkalkül – aus tradierten ethischen Theorien in das Steuerungssystem einer Maschine eingebaut. Der *Bottom-Up*-Ansatz hingegen geht eher induktiv vor. Er stellt die Entwicklung moralischer Sensibilität in den Mittelpunkt, wobei maschinelles Handeln in moralischen Entscheidungssituationen ohne explizit vorgegebene Regeln durch einen konnektionistischen Algorithmus[85] abgebildet wird. Unter Anwendung von Techniken ma-

83 Es sei darauf hingewiesen, dass die Verwendung von Begriffen wie ›ethical crash algorithms‹, ›ethical self-driving cars‹ oder deutschsprachigen Pendants, wie sie im Diskurs oft erfolgt, vor diesem Hintergrund irreführend ist. Fahrsysteme verfügen *nicht* über ethische Fähigkeiten im Sinne einer Reflexionsfähigkeit auf moralische Fragen. Entsprechende Termini sind daher metaphorisch zu verstehen; in der vorliegenden Arbeit wird auf deren Verwendung ausdrücklich verzichtet.

84 Die im Rahmen des maschinenethischen Diskurses häufig verwendete Bezeichnung ›moralische Maschinen‹ ist als Terminus technicus zu verstehen und nicht als attributive Charakterisierung.

85 Der Konnektionismus ist eine Forschungsrichtung in der KI-Forschung, welche die Grundlage des maschinellen Lernens bildet. Im Zentrum steht die Entwicklung von Softwarearchitekturen, deren Prozesse der Informationsverarbeitung der Funktionsweise des menschlichen Gehirns nachempfunden sind und sich anhand sogenannter künstlicher neuronaler Netze vollziehen.

schinellen Lernens[86] werden kontinuierlich Datensätze analysiert, die menschliches Handeln, das für moralisch korrekt befunden wurde, in real-lebensweltlichen Einzelfallsituationen dokumentieren (vgl. Birnbacher & Birnbacher, 2016, S. 13). Aus erkannten Verhaltensmustern werden sodann implizite Entscheidungskriterien abgeleitet, die von den künstlichen Systemen systematisiert und auf neue Fälle angewandt werden: »In bottom-up approaches to machine morality, the emphasis is placed on creating an environment where an agent explores courses of action and learns and is rewarded for behavior that is morally praiseworthy.« (Wallach & Allen, 2008, S. 80)

Sowohl *Top-Down-* als auch *Bottom-Up*-Ansätze sind nicht nur hinsichtlich ihrer technischen Implementierung herausfordernd (vgl. Misselhorn, 2018a, S. 165–166), sondern sie beruhen auch auf spezifischen metaethischen Grundannahmen über das Wesen der Moral und moralische Urteile, von denen ihre Plausibilität entscheidend abhängt. So geht der *Top-Down*-Ansatz von der generellen Begründbarkeit universaler Moralprinzipien aus, die unabhängig von Spezifika konkreter Situationen moralische Urteile ermöglichen und sich sodann in Form eines Systems abgeleiteter Handlungsprinzipien bzw. -regeln auf konkrete Fälle anwenden lassen (vgl. Filipović, 2016, S. 44; Misselhorn, 2018b, S. 96). Der *Bottom-Up*-Ansatz hingegen gründet sich auf ein partikularistisches Moralverständnis, das sich gegen eine Reduzierung der Moral auf theoriegeleitete Prinzipien wendet. Er betont, dass moralisches Handeln stets situatives Urteilsvermögen erfordert. Moralische Werte sind nicht universal begründbar, sondern kontextabhängig, indem sie immer schon implizit durch tatsächliches Handeln Ausdruck finden (vgl. Dancy, 2017; Wallach & Allen, 2008, S. 80). So ist unter besonderen Umständen ein Abweichen von ansonsten anerkannten moralischen Wertvorstellungen nicht nur akzeptabel, sondern sogar wünschenswert. Auch tatsächlich beobachtbares menschliches Verhalten gibt Anlass zu der Annahme, dass Menschen ihre moralischen Werte nicht immer streng nach Theorien ausrichten. Vielmehr scheinen sie diese

86 Vallor und Bekey (2017, S. 240) definieren maschinelles Lernen als »[…] a developmental process in which repeated exposures of a system to an information-rich environment gradually produce, expand, enhance, or reinforce that system's behavioral and cognitive competence in that environment or relevantly similar ones.«

im Laufe ihres Lebens unter verschiedenen Einflüssen zu kultivieren (vgl. Etzioni & Etzioni, 2017, S. 406–407).

In der softwaretechnischen Praxis erweisen sich beide Ansätze aufgrund ihrer starken Orientierung an explizit gegebenen bzw. erlernten Regelkatalogen für komplexe Anwendungsprobleme häufig als ungeeignet. Zum einen sind die Regeln, denen sie folgen, meist zu allgemein, um für eine ausreichende Zahl denkbarer Fälle zuverlässige Handlungsvorgaben liefern zu können.[87] Zum anderen sind bei der Anwendung ethischer Prinzipien hohe Rechenleistungen zur Informationsverarbeitung erforderlich; so müssen bei konsequentialistischen Kriterien sämtliche möglichen Konsequenzen für alle Handlungsoptionen berechnet werden. Aus diesen Gründen werden in der Praxis meist hybride Designansätze[88] gewählt, welche die Potenziale beider Konzepte vereinen (vgl. Allen et al., 2005, S. 151–154; Wallach & Allen, 2008, S. 80–81):

> Sie operieren mit einem vorgegebenen Rahmen moralischer Werte, der dann durch Lernprozesse an spezifische Kontexte angepasst und verfeinert werden kann. […] Um von einem hybriden Modell sprechen zu können, muss das künstliche System einen Spielraum zur Verfügung haben, innerhalb dessen es auf moralische Wertvorstellungen kontextsensitiv reagieren kann. (Misselhorn, 2019, S. 51)

Doch wie ›gut‹ ist die resultierende maschinelle Moral im Vergleich zur menschlichen? Auf der Basis von Techniken maschinellen Lernens konstruierte Maschinen bieten grundsätzlich den Vorteil, dass sie im Vergleich zu einer starren Programmierung flexibler sind, da sich ihr Handlungsraum nicht nur auf Situationen beschränkt, die fest einprogrammiert sind (vgl. Etzioni & Etzioni, 2017, S. 408–409).

---

87 Reed et al. (2021, S. 778) zeigen, dass dies unabhängig von moralisch brisanten Situationen bereits für Verkehrsregeln im Allgemeinen gilt.

88 Hinsichtlich der inhaltlichen Konkretisierung von Designansätzen beschreiben Pan et al. (2016) zwei Methoden, mittels derer sich die Verhaltenssteuerung automatisierter Systeme im Rahmen der sogenannten moralischen Regulierung umsetzen lässt. Während proskriptive Fahrstrategien den Fokus auf die Konformität mit Regeln sowie deren Übertretungen richten und eher Vermeidungsstrategien implizieren, betonen präskriptive Fahrstrategien das Erreichen bestimmter Ziele. Siegel und Pappas (2023, S. 217–220) evaluieren Techniken der Algorithmenimplementierung, die angesichts der durch Dilemma-Szenarien gegebenen komplexen praktischen Bedingungen jedoch allesamt nicht robust genug sind.

Jedoch sind Prozesse maschinellen Lernens nur dann zielführend, wenn das System auch die Möglichkeit hat, entsprechend dem Trial-and-Error-Prinzip aus nicht erfolgreichen Strategien zu lernen (vgl. Allen et al., 2005, S. 151). Eine gewisse Toleranz für fehlerhaftes Verhalten muss also vorhanden sein, kann allerdings im Fall moralischer Entscheidungssituationen zu kritischen Resultaten führen (vgl. Metz, 2016). Hier ist es zentral, einen kategorischen Unterschied zwischen moralischen und nicht-moralischen Handlungen anzuerkennen:

> However, […] [this] is to presume that there is no significant difference between learning to respond differently, say, to green, red, and yellow traffic lights, and—learning to understand and appreciate the moral imperative to take special care not to hit a bicyclist traveling in the same lane as the car, let alone not to harass or deliberately hit the cyclist out of road rage. (Etzioni & Etzioni, 2017, S. 407)

Eine unreflektierte Anwendung maschinellen Lernens auf moralisches Handeln wäre in zweierlei Hinsicht fragwürdig. Zum einen sind algorithmische Entscheidungen, die auf Lernen basieren, in ihrer Kausalität und Begründung häufig schwer nachvollziehbar und werfen u. a. im Hinblick auf Fragen der Verantwortungszuschreibung erhebliche Probleme auf. Insbesondere bei mehrschichtigen neuronalen Netzen ist mangelnde Transparenz problematisch:

> A major shortcoming of a neural network is its incapability to explain its decision. Unlike a decision tree, in which the logic can be traced back over several steps to its source, a neural network is not easily reverse-engineered, and it can be difficult to determine how it arrived at its decision. In an automated vehicle crash, an understanding of the logic behind an automated vehicle's actions is critical, particularly if the vehicle did not behave as expected. […] Without the knowledge of why an automated vehicle behaves a certain way, there is no way to fix the problem to ensure that it will not happen again. (Goodall, 2014a, S. 63)

Zum anderen ist fraglich, an welchem ›Vorbild‹ sich die Systeme beim Lernen orientieren sollen. Wie bereits erläutert, sind menschliche Reaktionen in Dilemma-Situationen aus moralischer Sicht nicht ausgereift; eine Orientierung an diesen würde keine ›Verbesserung‹ bewirken:

> Humans tend to react very slowly and badly in car crash situations; they can even kill entire families instinctively (though unintentionally)

> rather than run over a squirrel. Instructing smart cars to act in the same way would amount not only to wasting the potential of such cars, but also effectively randomising crash outcomes, because humans often effectively randomise such results through intuitive or slow decision making. (Shaw & Schneble, 2021, S. 75)

Wenn menschliche Fahrer in kritischen Situationen also gar nicht moralisch handeln, sondern lediglich instinktiv reagieren, dann enthalten Trainingsdaten, die menschliches Verhalten abbilden, nicht das, was richtig, sondern was üblich ist. Es wäre erforderlich, dass eine menschliche Kontrollinstanz die von der Maschine im Rahmen des Lernprozesses extrahierten Entscheidungskriterien nochmals evaluiert und auf ihre Akzeptabilität hin prüft,[89] bevor das System sie übernimmt (vgl. Brändle & Grunwald, 2019, S. 287) – ein klassischer Anwendungsfall für den *Human-in-the-Loop*-Ansatz.[90] Geschieht dies nicht, käme es zu einem naturalistischen Fehlschluss, indem aus beobachtbarem bzw. gegebenem Verhalten ein normativer Geltungsanspruch gefolgert wird. So würde auch moralisch falsches Verhalten erlernt, wobei sich beispielsweise in den Trainingsdaten vorhandene Diskriminierungseffekte in Algorithmen verfestigen können. Damit autonome Systeme sinnvolle Entscheidungsstrategien für Unfalldilemmata erlernen, sollten die Trainingsdaten nur moralisch wünschenswertes Verhalten beinhalten:[91] »Ethics addres-

89 Ein solches Prüfverfahren existiert momentan bereits im Hinblick auf nicht-moralische Lerninhalte; eine unabhängige Drittprüfung durch technische Dienste, die das Erlernte verifizieren, ist erforderlich, bevor angelernte Systeme eingesetzt werden dürfen. Der TÜV-Verband fordert, diese Vorgehensweise auch im Rahmen der Anpassungen an die Anforderungen des *AI Act* beizubehalten und KI-Systeme im Automobilbereich auf diese Weise gesetzeskonform abzusichern (vgl. TÜV-Verband e. V., 2024).

90 Der Begriff des *Human in the Loop* (*HITL*) bezeichnet ein verbreitetes Konzept des Softwareengineering, bei dem Mensch und Maschine gemeinsam daran arbeiten, optimale Ergebnisse zu erzielen. Es handelt sich um eine Form der Interaktion, die den Menschen aktiv in den Entscheidungszyklus eines Systems einbindet. Durch kontinuierliche Überwachung und mögliche Eingriffe werden die vom System erzeugten Ergebnisse dabei durch menschliches Urteilsvermögen validiert.

91 Alternative Positionen weisen in diesem Zusammenhang auf strukturelle Ähnlichkeiten zwischen maschinellem und moralischem Lernen, z. B. in Bezug auf das Erlernen von Normen, hin. So erläutert Wolkenstein (2018, S. 169–170): »[…] consider that the (split-second) decisions in a typical TD is based on a history of moral education that resembles the history of ›moral education‹ an

ses how humans ought or want to behave, rather than how they actually behave, and artificial intelligence techniques should capture ideal behavior.« (Goodall, 2014a, S. 62)

Die beschriebenen Schwierigkeiten, welche beim Einsatz maschinellen Lernens für moralische Entscheidungen entstehen, treten umso stärker hervor, wenn es um Entscheidungsstrategien für Einzelfallsituationen geht. Diese stellen vor allem im Hinblick auf die erforderliche Kontextsensitivität[92] hohe Anforderungen an die entsprechenden Systeme. Aus softwaretechnischer Sicht verfügen kontextsensitive Systeme über sogenannte Kontextmodelle, in denen die zu erfassenden Situationsparameter definiert sind. Dies können beispielsweise die Umgebungstemperatur oder Spezifika der Objekterkennung (Masse, Geschwindigkeit) in einem bestimmten definierten Umfeld sein. Da die Generierung dieses Modells eine hohe Rechenleistung erfordert, geschieht dies bei *Machine-Learning*-Systemen allerdings nicht zur Laufzeit, sondern bereits in der Lernphase. In Bezug auf moralische Entscheidungssituationen gilt dann, dass die als moralisch relevant identifizierten Kontextvariablen sowohl für die Herausbildung von Entscheidungskriterien aus den Lerndatensätzen als auch bei der Entscheidung von entsprechenden realen Situationen herangezogen werden.

Dies erscheint so lange unproblematisch, wie nur Standardsituationen auftreten, bei denen Lern- und Anwendungskontext übereinstimmen (vgl. Bendel, 2016, S. 65) und für die es viele Beispieldatensätze gibt. Nun werden autonome Systeme aber mit der voranschreitenden technologischen Entwicklung in immer komplexeren Bereichen eingesetzt, in denen die Zahl möglicher, hochspezifischer Handlungsszenarien potenziell unbegrenzt ist. Soll eine Maschine

---

algorithm has experienced. The intuitions people have in a TD are not merely spontaneous reactions, but are based on the working mechanism or morality, just as the algorithm is based on the working mechanism of morality, including the intuitive reactions to a TD that engineers and programmers have.«

92 In der Informationstechnik beschreibt der Begriff der Kontextsensitivität die Fähigkeit von Systemen, ihr Verhalten in Abhängigkeit von Informationen über ihre Umgebung bzw. ihren Kontext zu regulieren. In der Perzeptionsphase werden Kontextinformationen dabei meist über entsprechende Sensoren erfasst und im Rahmen von parametrisierten Funktionen in Systemalgorithmen verwertet. Ein Beispiel für ein kontextsensitives System sind ortsabhängige Dienste, die den anhand von GPS-Daten ermittelten Standort ihrer Nutzer berücksichtigen.

eine Situation auf der Basis von erlerntem Verhalten bewältigen, so kann sie nur diejenigen Fälle ›korrekt‹ entscheiden, die in den entsprechenden Trainingsdaten abgebildet sind. Diese Datenbasis kann hinsichtlich Quantität und Qualität abgebildeter Szenarien zwar stetig weiterentwickelt werden, jedoch nie alle denkbaren Szenarien, ihre spezifischen Besonderheiten und »kritischen Situationsentwicklungen« (Dietmayer, 2015, S. 435) vollständig abdecken (vgl. Reed et al., 2021, S. 778). Das trifft insbesondere im Hinblick auf dynamische Kontextvariablen zu, die ihre Zustände verändern können, z. B. Lichtsignalanlagen, Licht- und Wetterbedingungen oder andere Verkehrsteilnehmer (vgl. Geyer et al., 2014, S. 185). Es können immer Situationen auftreten, die entgegen aller Voraussicht ein wenig anders sind, unbekannte Objekte enthalten oder ungewöhnlich komplexe Konstellationen aufweisen (vgl. Reed et al., 2021, S. 784). Auch bei der korrekten Erfassung und Klassifizierung von Objekten mangelt es an technischer Präzision, sodass z. B. der Typ eines Objekts oder die Anzahl potenziell involvierter Personen nicht zuverlässig bestimmt werden können (vgl. Kirkpatrick, 2015, S. 19). Unfallalgorithmen beziehen sich daher tendenziell auf eine definierte Klasse von Szenarien, nicht aber auf jedes denkbare spezifische Einzelszenario. Bendel (2018, S. 35) spricht in diesem Zusammenhang von Unschärfen, die sich zwischen Moral einerseits und Anwendungsfall der Moral andererseits ergeben.

In der Praxis bedeutet das, dass eine Maschine tatsächliche Situationen in der realen Lebenswelt anhand ihres zugrundeliegenden Kontextmodells u. U. nicht korrekt klassifizieren kann (vgl. LaCroix, 2022). Den trainierten Systemen fehlen zudem die kognitiven Kapazitäten, um zu erkennen, weshalb ein bestimmtes erlerntes Verhalten ethisch wünschenswert ist. So sind sie weder in der Lage, selbst ethische Prinzipien zu entwickeln, um diese in neuen, unbekannten Situationen anzuwenden, noch können sie ihre eigenen Handlungen begründen. Dies ist besonders problematisch, wenn es um Fragen der (Hersteller-)Haftung für ein bestimmtes Fahrzeugverhalten geht (vgl. Reed et al., 2021, S. 778). Techniken maschinellen Lernens unterliegen damit naturgemäß konzeptionellen bzw. informationstechnischen Grenzen, die entscheidend dafür verantwortlich sind, dass autonome Fahrsysteme sich nicht in dem Maße kontextsensitiv konstruieren lassen, wie es nötig wäre, um Entscheidungsdilemmata

adäquat zu bewältigen.[93] Moralische Entscheidungsprobleme mit dilemmatischen Strukturen sind immer Einzelfälle, die Fingerspitzengefühl und eine Würdigung der spezifischen Umstände erfordern.

Nachdem bisher in weitgehend abstrakter Weise von Unfallszenarien die Rede war, sollen diese im Folgenden näher konkretisiert werden. Im Rahmen des nachfolgenden Unterkapitels wird daher eine Übersicht über mögliche Szenarienkonstellationen und ihre spezifischen ethischen Problematiken präsentiert, auf die im weiteren Argumentationsgang immer wieder rekurriert wird.

4.1.3 Systematisierung repräsentativer Dilemma-Szenarien und ihre moralphilosophische Problematisierung

Der wissenschaftliche Diskurs um Unfallalgorithmen verdankt seine Praxisnähe und Lebendigkeit nicht zuletzt einer Vielfalt anschaulicher Beispielszenarien, welche die vielschichtigen moralischen Problemkomplexe relevanter Unfallkonstellationen illustrieren. Im Anschluss an den lebhaft geführten Diskurs wird im weiteren Verlauf dieses Buches zu Veranschaulichungszwecken auf nachfolgend aufgeführte repräsentative Szenarien stellenweise Bezug genommen.

---

93 Was folgt daraus für die weitere maschinenethische Forschung und den praktischen Einsatz von *AMAs*? Für gegenwärtige Systeme müssen Lösungsansätze entwickelt werden, die auf realistischen Ansprüchen an moralische Maschinen basieren und zugleich sicherstellen, dass beim Einsatz von beschränkt moralisch handlungsfähigen Systemen keine ethischen Konflikte auftreten. Hilfreich erscheint in diesem Zusammenhang eine inkrementelle Vorgehensweise im Sinne des dynamischen Klassifizierungskonzepts moralischer Handlungsfähigkeit von Wallach und Allen (2008, S. 25–33). Diesem zufolge entwickeln sich komplexe *AMAs* aus primitiveren Formen der Technologie im Zuge der Interaktion von steigender Autonomie und Wertesensitivität von *operational* über *functional morality* hin zu *responsible/full moral agency*. Für konkrete praktische Zusammenhänge kommen bereits heute verschiedene Realisierungskonzepte in Frage. So ist es denkbar, die Autonomie künstlicher Systeme einzuschränken und diese als moralische Ratgeber einzusetzen, die moralische Entscheidungen lediglich für einen menschlichen Entscheidungsträger vorbereiten (vgl. Misselhorn, 2018b, S. 72–74). Alternativ könnten Maschinen zunächst als einfache moralische Akteure für beschränkte Einsatzbereiche konstruiert werden, sodass sie auf Basis weniger Regeln nur Standardsituationen entscheiden müssen (vgl. Bendel, 2016, S. 65). Auf der Grundlage derartiger Ansätze ließen sich künstliche Systeme sodann unter begleitender Berücksichtigung ethischer Aspekte schrittweise weiterentwickeln.

Beispielszenario 1 ›Großmutter versus Kind‹: Ein autonomes Fahrzeug steuert unaufhaltsam auf eine Großmutter und ihr Enkelkind zu, die innerhalb der Bremsdistanz die Straße überqueren. Linksseitiges Ausweichen wäre gleichbedeutend mit einer Kollision mit dem Kind, wobei die Großmutter unverletzt bliebe; rechtsseitiges Ausweichen würde dagegen die Großmutter verletzen und das Kind verschonen. Ohne Ausweichmanöver würden beide durch den Frontalaufprall schwer verletzt (vgl. Lin, 2015, S. 70).

Beispielszenario 2 ›Einzelperson versus Gruppe‹: Ein autonomes Fahrzeug fährt unaufhaltsam auf eine vierköpfige Gruppe von Personen zu, die innerhalb der Bremsdistanz die Straße überqueren. Bei linksseitigem Ausweichen würde das Fahrzeug mit nur einer der Personen kollidieren; bei Ausweichen nach rechts würden dagegen die anderen drei Personen verletzt. Ohne Ausweichmanöver würden alle Beteiligten durch den Frontalaufprall schwer verletzt (vgl. ebd., S. 70).

Beispielszenario 3 ›Rote Ampel‹: Ein autonomes Fahrzeug steuert auf eine Person zu, die eine Fußgängerampel bei Rot verkehrswidrig überquert. Durch eine abrupte Notbremsung bliebe die Person unverletzt, jedoch würde ein nachfolgender Motorradfahrer durch den resultierenden Aufprall schwer verletzt (vgl. Coca-Vila, 2018, S. 62).

Beispielszenario 4 ›Motorradfahrer mit/ohne Helm‹: Ein autonomes Fahrzeug befindet sich auf der rechten Fahrspur einer Autobahn. Durch ein plötzlich auf der Fahrbahn auftauchendes Hindernis kann ein Zusammenstoß mit einem vor dem Fahrzeug fahrenden Motorradfahrer, welcher keinen Helm trägt, nur durch einen Wechsel in die benachbarte Spur vermieden werden, wo sich ein zweiter Motorradfahrer befindet, der die vorgeschriebene Schutzausrüstung trägt. Während eine Kollision für den vorausfahrenden Motorradfahrer ohne Helm tödlich enden würde, würde der zweite lediglich leicht verletzt werden (vgl. Coca-Vila, 2018, S. 62–63; Goodall, 2014a, S. 62; Lin, 2014a, 2015, S. 73).

Beispielszenario 5 ›Unbeteiligte auf Bürgersteig‹: Aufgrund eines spontan auftretenden Bremsversagens steuert ein autonomes Fahrzeug ungebremst auf eine Person zu, welche die Straße an

einem Fußgängerüberweg vorschriftsmäßig überquert. Die Person kann nur gerettet werden, indem das Auto auf den Bürgersteig ausweicht, wo sich eine unbeteiligte Fußgängerin befindet.

Beispielszenario 6 ›Tunnel‹: Ein autonomes Fahrzeug nähert sich einem Tunnel, als das vorausfahrende Fahrzeug im gebundenen Verkehr plötzlich abrupt abbremst. Eine Kollision kann nur vermieden werden, indem das Fahrzeug ausweicht und in die Tunnelwand steuert, wodurch die Insassen verletzt würden.

Beispielszenario 7 ›Klippe‹: Ein autonomes Fahrzeug fährt auf einer schmalen Straße entlang einer Klippe oder eines Grabens. In einer unübersichtlichen Kurve kommt ihm ein voll besetzter Schulbus entgegen, der verkehrswidrig die Kurve schneidet. Ein Frontalzusammenstoß ließe sich nur vermeiden, indem das Fahrzeug in Richtung des steilen Abhangs ausweicht (vgl. Lin, 2013a, 2015, S. 76).[94]

Beispielszenario 8 ›Herannahender LKW‹: An einer Kreuzung wartet ein automatisiertes Fahrzeug an einer roten Fußgängerampel. Die Sensoren des Fahrzeugs erkennen einen sich von hinten nähernden LKW, der mit ungebremster Geschwindigkeit auf die Kreuzung zufährt. Die einzige Möglichkeit, einen Auffahrunfall und damit einen erheblichen Schaden für die Insassen beider Fahrzeuge zu vermeiden, besteht darin, dass das autonome Fahrzeug die rote Ampel überfährt und in einem Ausweichmanöver nach rechts abbiegt, wobei es allerdings einige Kinder leicht verletzen würde, die gerade die Straße überqueren (vgl. Lin, 2015, S. 78).[95]

Wie aus dieser Zusammenstellung von Beispielen ersichtlich wird, lassen sich Dilemma-Szenarien hinsichtlich der ethischen Problemstellungen, die sie jeweils tangieren, systematisieren und voneinander abgrenzen. Die beiden zentralen Kategorien bilden dabei moralische Quantifizierungs- und Qualifizierungsprobleme. Welche Rolle spielt die Anzahl potenziell betroffener Personen bzw. die Höhe des zu erwartenden Schadens für die Entscheidungsfindung (siehe Bei-

94 In ihrem Grundaufbau ähnliche, jedoch modifizierte Szenarien finden sich u. a. bei Gogoll und Müller (2017, S. 683), Goodall (2016a, S. 810), Himmelreich (2018, S. 669) und Marcus (2012, o. S.).

95 Geringfügig variierte Szenarien finden sich bei Goodall (2014a, S. 59, 2020, S. 3).

spielszenario 2 ›Einzelperson versus Gruppe‹)? Während sich Quantifizierungsprobleme mit der simplen Formel ›Do numbers count?‹ beschreiben lassen, können sich Qualifizierungsfragen auf verschiedene Aspekte der Szenarienkonstellation beziehen: Inwiefern sollen persönliche oder soziale Merkmale der Betroffenen, z. B. Alter, Geschlecht oder sozialer Status, besondere Berücksichtigung finden (siehe Beispielszenario 1 ›Großmutter versus Kind‹)? Qualifizieren findet immer da statt, wo Lebenswerte gegeneinander abgewogen werden. Im weiteren Sinne fallen in diese Kategorie zudem spezifische Szenarien, die thematisieren, inwiefern die Insassen des betreffenden autonomen Fahrzeugs besonders schutzwürdig sind bzw. ob sich diese in altruistischer Weise selbst opfern sollten, um andere Parteien zu schützen (siehe Beispielszenario 6 ›Tunnel‹). Sollen Unfallalgorithmen egoistisch oder altruistisch eingestellt sein?

Neben klassischen Quantifizierungs- und Qualifizierungsproblemen adressieren einige Szenarien eine weitere komplexe Fragestellung, die sich auf den Grad moralisch notwendiger Intervention bezieht. Relevant ist hierbei vor allem die Unterscheidung von Handlungen aktiven Tötens und passiven Sterbenlassens sowie die daraus folgende Klassifizierung von Schäden als Handlungs- oder Unterlassungsfolgen. Klassischer Anwendungsbereich der sogenannten *Killing-versus-Letting-Die*-Debatte, in der derartige Überlegungen primär zum Tragen kommen, sind medizinethische Diskurse, beispielsweise um die ethische Beurteilung aktiver Euthanasie. Zunehmend sind sie jedoch auch in nicht-medizinischen Handlungsbereichen involviert. In dilemmatischen Fahrsituationen entfalten sie in Form einer möglichen Unterscheidung zwischen beteiligten und unbeteiligten Personen und deren Rolle für ethisches Urteilen eine nicht unerhebliche Bedeutung. Dabei ist auf der einen Seite relevant, ob die potenziell Betroffenen an der Entstehung der Situation im Sinne eines schuldhaften Verhaltens beteiligt sind. Dies wäre etwa dann der Fall, wenn sie grob fahrlässig handeln bzw. bewusst gegen Verkehrsregeln verstoßen (siehe Beispielszenario 3 ›Rote Ampel‹). In diesem Zusammenhang wird häufig auch problematisiert, dass unter gewissen Umständen – insbesondere durch das Postulat der Schadensminimierung – implizite Fehlanreize gesetzt werden, die das Unterlassen von Schutzmaßnahmen zur persönlichen Risikominimierung motivieren, wie im Beispielszenario 4 ›Motorradfahrer mit/ohne Helm‹ (vgl. Motwani et al., 2021, S. 53). Auf der anderen

Seite wird in zahlreichen Forschungsbeiträgen diskutiert, inwiefern Unbeteiligten ein legitimes Eigeninteresse und ein daraus abgeleitetes Recht zugesprochen werden sollte, nicht ungerechtfertigt in eine Unfallsituation verwickelt zu werden (siehe Beispielszenario 5 ›Unbeteiligte auf Bürgersteig‹).

Auf der Komplexitätsachse lassen sich Dilemma-Szenarien zudem entsprechend ihrer moralphilosophischen Vielschichtigkeit verorten. Triviale Szenarien wie die oben beschriebenen Beispiele 1 bis 6 zeichnen sich dadurch aus, dass sie stets *eine* spezifische Problemstellung in den Vordergrund rücken. So geht es in reinen Qualifizierungsszenarien wie dem Beispielszenario 1 ›Großmutter versus Kind‹ um die Frage, wie sich Unterschiede bei persönlichen oder sozialen Merkmalen auf die Entscheidungsfindung auswirken, während andere Aspekte wie die Anzahl der betroffenen Personen nicht relevant sind bzw. im Sinne von Kontrollvariablen konstant gehalten werden. Komplexere Szenarien dagegen integrieren mehrere ethische Problemstellungen mit der Absicht, mögliche Hierarchien zwischen Prinzipien bzw. Ansätzen zu untersuchen. Welchen ethischen Aspekten soll bei der Entscheidungsfindung Priorität eingeräumt werden? Soll das Fahrzeug im Konfliktfall vorranging quantifizieren oder qualifizieren? Welche Rolle spielt die Beteiligung an der Unfallentstehung? Ein Vergleich der beiden Szenarien ›Tunnel‹ und ›Klippe‹ macht exemplarisch deutlich, welch erheblichen Einfluss kleine Veränderungen in der Konstellation der dilemmatischen Situation auf die ethische Beurteilung haben können, beispielsweise eine Skalierung der Anzahl betroffener Personen, deren Alter oder ein mögliches schuldhaftes Fehlverhalten. In ähnlicher Weise stellt das Beispielszenario 8 ›Herannahender LKW‹ die Frage zur Diskussion, ob eine mögliche leichte Verletzung der angefahrenen Kinder einer möglichen schweren Schädigung der beiden Fahrzeuginsassen vorzuziehen wäre, also ob die Schwere der zu erwartenden Personenschäden im Rahmen der ethischen Entscheidungsfindung berücksichtigt werden sollte.

Nachdem geklärt ist, welche moralphilosophischen Problematiken in Dilemma-Szenarien auftreten können, stellt sich nun die Frage, welche Strategien sich aus ethischer Sicht zu deren Bewältigung anbieten. Vor diesem Hintergrund wird nun in einem nächsten Schritt das dominante Forschungsframework, das sich am berühmten Trolley-Problem orientiert, eingeführt und kritisch beleuchtet.

4.1.4 Dilemma-Szenarien als angewandtes Trolley-Problem? Von Diskrepanzen und Disanalogien

Den Beginn des ethischen Diskurses um Unfallalgorithmen markierte Mitte der 2010er-Jahre die Auseinandersetzung mit spezifischen Dilemma-Szenarien vor dem Hintergrund des prominenten Trolley-Problems. Dabei handelt es sich um ein philosophisches Gedankenexperiment mit dem Ziel, moralische Intuitionen anhand von Entscheidungen in konstruierten Szenarien mit unvermeidbaren negativen Konsequenzen zu analysieren und Begründungen für normative Schlussfolgerungen darzulegen. In seiner klassischen Form geht das Trolley-Problem auf die Philosophin Philippa Foot (1978) zurück. Es versetzt die Teilnehmer in die Situation eines Fahrers einer außer Kontrolle geratenen Straßenbahn, die mit hoher Geschwindigkeit auf fünf Personen zusteuert, welche auf den Gleisen arbeiten. Um die Gleisarbeiter vor dem sicheren Tod zu retten, hat der Fahrer die Option, die Straßenbahn auf ein anderes Gleis umzuleiten, auf dem sich eine Person befindet, die ebenfalls durch den Aufprall getötet würde. Was soll der Fahrer tun?

Foots klassische Version wurde vielfach aufgegriffen, variiert und für verschiedene Anwendungskontexte modifiziert. Eine prominente Weiterentwicklung und heute zugleich die am häufigsten referenzierte Variante stammt von Judith Jarvis Thomson (1976, 1985b). Sie ersetzt den Fahrer in seiner Funktion als Entscheidungsträger durch einen ansonsten unbeteiligten Zuschauer, der durch einen Schalter die Straßenbahn umleiten kann:

> Let us begin by looking at a case that is in some ways like Mrs. Foot's story of the trolley driver. I will call her case Trolley Driver; let us now consider a case I will call Bystander at the Switch. In that case you have been strolling by the trolley track, and you can see the situation at a glance: The driver saw the five on the track ahead, he stamped on the brakes, the brakes failed, so he fainted. What to do? Well, here is the switch, which you can throw, thereby turning the trolley yourself. Of course you will kill one if you do. (Thomson, 1985b, S. 1397)

Ferner ergänzt Thomson das Gedankenexperiment um eine weitere Variante, die als ›Fetter-Mann-Problem‹ (*fat man problem*) bekannt ist: Die handelnde Person befindet sich auf einer Brücke über den Gleisen und erkennt, dass die Straßenbahn nur aufgehalten werden kann, wenn ein dicker Mann, der sich ebenfalls auf der Brücke

befindet, auf diese Gleise gestoßen wird, um die Bahn auf diese Weise zum Stehen zu bringen (vgl. Thomson, 1976, S. 207–208).

Der moralische Konflikt, der dem Gedankenexperiment zugrunde liegt, bewegt sich in zwei Dimensionen, die eng miteinander zusammenhängen und somit die Komplexität der Problemstellung erhöhen: Zum einen geht es um die Problematik des Aufwiegens von Menschenleben, die sich in der Frage konkretisiert, ob der Tod Weniger in Kauf genommen werden darf bzw. sollte, um Viele zu retten. Zum anderen thematisiert das Trolley-Problem insbesondere eine kritische Unterscheidung in Bezug auf ethische Entscheidungen und deren zugrundeliegende Intentionen: Ist (aktives) Töten aus moralischer Sicht schlechter als (passives) Sterbenlassen? Die Untersuchung unserer moralischen Intuition hinsichtlich einer möglichen ethisch relevanten Unterscheidung zwischen Handlungen aktiven Handelns einerseits und passiven Unterlassens andererseits steht im Zentrum der sogenannten *Doing-versus-Allowing*-Problematik.[96] Wie empirische Studien nahelegen, bestehen Wechselwirkungen zwischen beiden Dimensionen, sodass diese nicht gänzlich unabhängig voneinander betrachtet werden können (vgl. Greene, 2013).

Für den Forschungsdiskurs um Unfallalgorithmen ist das Trolley-Problem von großer Bedeutung. Bisher wurde es in weiten Teilen der Forschungsliteratur als dominantes Framework verwendet, um Dilemma-Szenarien des autonomen Fahrens zu adressieren. Entsprechende, vermehrt kritische Auseinandersetzungen stellen noch heute einen wesentlichen Anteil der einschlägigen Neupublikationen zum Thema; es gibt kaum einen Artikel, der nicht in der einen oder anderen Weise Bezug auf das Trolley-Problem nimmt (vgl. Santoni de Sio, 2021, S. 715).[97] Im Anschluss an Thomsons Modifikationen wurden im Rahmen des Diskurses um Unfallalgorithmen verschiedene Varianten von Trolley-Fällen[98] als pointierte Repräsentationen dilemmatischer Unfallszenarien vorgeschlagen. Wie soll sich ein au-

96 Alternative Bezeichnungen des entsprechenden ethischen Diskurses sind ›*Killing versus Letting Die*‹ bzw. ›*Intending versus Foreseeing*‹.

97 Eines der wenigen alternativen Gedankenexperimente entwerfen Kumfer und Burgess (2015). Dabei werden Szenarien mittels eines MATLAB-Programms simuliert, um die Implikationen verschiedener ethischer Theorien für Unfallalgorithmen zu untersuchen.

98 Zur begrifflichen Unterscheidung zwischen ›Trolley-Problem‹ und ›Trolley-Fällen‹ siehe Himmelreich (2018, S. 669–670).

tonomes Fahrzeug angesichts einer drohenden Kollision mit fünf Personen verhalten, die die Straße überqueren? Soll es ausweichen und stattdessen auf einen einzelnen Fußgänger zusteuern, oder sollte es gar seine Insassen opfern, indem es z. B. auf ein schweres Hindernis auffährt? Diese und ähnliche Fragestellungen scheinen auf den ersten Blick auf eine praxisnahe Reformulierung des Trolley-Problems hinzudeuten. Deren spezifische Instanzen zeichnen sich durch bestimmte Eigenschaften aus:

> (1) the AV must choose one of two actions; (2) the AV knows what the consequences of each action will be; (3) each action imposes a distribution of benefits and burdens over at least two affected parties; and (4) the interests of these parties are jointly unsatisfiable. (Keeling, 2020, S. 294)

Im Zentrum des Interesses des ursprünglichen Trolley-Problems steht der Versuch, Veränderungen in moralischen Intuitionen zwischen verschiedenen Trolley-Fällen zu begründen, die sich geringfügig in moralisch relevanten Gesichtspunkten unterscheiden. Die Debatte über Unfallalgorithmen bezieht sich dagegen nicht auf das Gedankenexperiment in diesem engeren, klassischen Sinne, sondern versteht es als philosophische Methode, die idealisierte Fälle gebraucht, um moralisch relevante Merkmale zu identifizieren und zu untersuchen. Unfallalgorithmen werden im Kern als Programmier- bzw. Designentscheidung maschineller Moral aufgefasst, die die Ausgestaltung konkreter algorithmischer Steuerungsaktionen als Antwort auf modifizierte Trolley-Szenarien betrifft. Dilemma-Szenarien lassen sich dabei als spezifische moralische Entscheidungsprobleme hinsichtlich der Frage beschreiben, welche ethischen Handlungsprinzipien bzw. moralischen Werte in Notsituationen zur Anwendung kommen sollen. Wie soll ein Fahrzeug in der jeweiligen Situation agieren?[99]

Obwohl ein auf Instanzen modifizierter Trolley-Fälle basierender, moralphilosophischer Zugang die Forschung zu Unfallalgorithmen weitgehend dominiert, stehen viele Wissenschaftler, die sich mit der Ethik autonomen Fahrens auseinandersetzen, dieser vermeintlichen

99 Einige Argumente, die sowohl für eine direkte als auch indirekte (normative) Relevanz des Trolley-Problems für Unfallalgorithmen sprechen, werden von Paulo (2023) skizziert.

Analogie zunehmend kritisch gegenüber. Sie merken an, dass es sich bei Trolley-Fällen im Wesentlichen um theoretisch konstruierte Szenarien handelt, die auf Annahmen beruhen, welche grundlegend verschieden sind von moralischen Entscheidungen in real-lebensweltlichen Kontexten.[100] Wie Lawlor (2022, S. 207–214) ausführt, ist das Trolley-Problem in seinem Zweck und Wesen innerhalb des Forschungsdiskurses um Unfallalgorithmen vielfach fehlinterpretiert worden. Tatsächlich ähnelt es eher einem Laborexperiment als einer Modellvorlage für angewandte Probleme. So sind Trolley-Fälle geeignet, individuelle moralische Intuitionen durch die Konfrontation mit Extremfällen zu offenbaren und kritisch zu hinterfragen sowie einige der zentralen moralphilosophischen Problematiken der jeweiligen Entscheidungssituation aufzudecken. Als philosophisches Gedankenexperiment ist das Trolley-Problem jedoch als isoliertes Entscheidungsproblem konzipiert, das von jeglicher Kontexteinbettung abstrahiert. Es ›existiert‹ nur innerhalb der Experimentumgebung und konstruiert Trolley-Fälle als binäre *Single-Choice*-Entscheidungen in einer vollständig kontrollierbaren Umgebung (vgl. Goodall, 2014b, S. 96, 2016a, S. 812, 2017, S. 496), die keine Abhängigkeiten zu externen Faktoren aufweist.[101] Betroffene Personen werden weitgehend als unpersönliche Entitäten und Entscheidungsträger als Unbeteiligte modelliert (vgl. Hübner & White, 2018, S. 688; Liu, 2017, S. 202), Verantwortungsaspekte bleiben unberücksichtigt.

Im Gegensatz dazu sind Unfallalgorithmen als praktisches Problem an real-lebensweltliche Kontexte geknüpft. Smilansky (2022, S. 118–122) erläutert, dass das binäre ›Entweder–Oder‹-Design paradigmatischer Trolley-Fälle die Vielfalt flexibler Handlungsmöglichkeiten, die sich autonomen Fahrsystemen bietet, nicht adäquat widerspiegelt. Entscheidungen in realen Kontexten sind nicht isoliert, auch ein Vorher und ein Nachher entfalten moralische Relevanz.

100 Eine umfangreiche Diskussion der Unterschiede zwischen Trolley-Problem einerseits und Entscheidungsdilemmata im Kontext des autonomen Fahrens andererseits wurde bereits an anderer Stelle von der Autorin publiziert (vgl. Schäffner, 2021). Weitere vertiefende Auseinandersetzungen und Systematisierungen zur Thematik finden sich z. B. bei Bruers und Braeckman (2014), Fossa (2023), Himmelreich (2018), Nyholm und Smids (2016), Wolkenstein (2018) und Wu (2020).

101 Siehe auch Kap. 4.3.2 für weitere Kritikpunkte an einem trolley-basierten Design experimenteller Studien.

Vor allem was ihre Entstehung angeht, sind Unfalldilemmata in hohem Maße kontextualisiert; beispielsweise ist es für Haftungsfragen höchst relevant, wie es zu einem entstandenen Schaden gekommen ist (vgl. Kauppinen, 2021, S. 630–631). Als Entscheidungsträger können wir uns den Situationen, in denen unsere Entscheidungen zum Tragen kommen, nicht entziehen: »Yet, ethical situations are not snapshots frozen in time but uncertain and living movements. We are not engaged with them as outside judges but as ethical characters.« (JafariNaimi, 2018, S. 309) Aspekte moralischer und rechtlicher Verantwortung sind stets mit getroffenen Entscheidungen verwoben (vgl. Nyholm & Smids, 2016, S. 1283–1284; Santoni de Sio, 2017, S. 420). Dies gilt umso mehr im Hinblick darauf, dass für autonome Fahrsysteme keine Grundeinstellung hinsichtlich aktivierter Trajektorien festgelegt ist, sodass jede Aktion vom System berechnet und daher als aktive Handlung gedeutet werden muss. Weitere Diskrepanzen zwischen Trolley-Problem einerseits und Dilemma-Szenarien andererseits ergeben sich bei der Konzeption des jeweils zugrundeliegenden Entscheidungsproblems. Trolley-Fälle fragen nach den moralischen Präferenzen bezogen auf eine konkrete Situation; als moralphilosophische Entscheidungsprobleme werden sie den gesellschaftlichen Effekten,[102] die ein spezifisches Design von Unfallalgorithmen mit sich bringt, nicht gerecht.[103]

Auch die eingeschränkte Implementierbarkeit des Trolley-Problems in Algorithmen mittels Konzepten maschinellen Designs stellt eine gravierende Limitation der Vorgehensweise dar, Dilemma-Szenarien auf Instanzen eines angewandten Trolley-Problems zu reduzieren (vgl. Himmelreich, 2018, S. 675; Keeling, 2020, S. 301). Die spezifische Struktur des Gedankenexperiments erfordert Antworten in Form expliziter ethischer Handlungen oder Handlungsprinzipien, die aus technischer Sicht einem *Top-Down*-Ansatz entsprächen. Nun stützt sich die technische Realisierung hochkomplexer autonomer Systeme jedoch zu einem großen Teil auf verhaltenssteuernde Komponenten, die auf Methoden maschinellen Lernens bzw. neuronaler Netze basieren. Das Verhalten autonomer Fahrzeuge in Unfalldilem-

102 Dieser Aspekt wird im nachfolgenden Kap. 4.2 näher ausgeführt.

103 Für eine argumentative Auseinandersetzung hierzu siehe Smith (2022, S. 286–289), der die Rolle und Plausibilität eines trolley-basierten Designs von Dilemma-Szenarien vor dem Hintergrund eines institutionalistischen Verständnisses diskutiert.

mata wird von Algorithmen bestimmt, die nicht von Ingenieuren direkt programmiert werden, sondern die sich das System auf Basis von Trainingsdaten selbst generiert hat. Diese erlauben es nicht, die Reaktion autonomer Fahrzeuge in spezifischen Unfallszenarien in stets konsistenter Weise vorherzubestimmen. Daraus folgern Behrends und Basl (2022), dass das Design des Trolley-Problems als Entscheidungsproblem mit klar definierten, eindeutigen Antworten nicht geeignet ist, um direkte Implikationen für die technische Implementierung von Unfallalgorithmen abzuleiten.

Um es kurz zu sagen: Im Gegensatz zum Trolley-Problem sind Unfalldilemmata ›mitten aus dem Leben gegriffen‹; der Kontext, in dem sie stehen, ist essenziell, um sie vollumfänglich begreifen und bewältigen zu können. Das nächste Unterkapitel ist daher einer ausführlichen Untersuchung des praktischen Problemkontextes gewidmet, der bedeutende Implikationen für die Konzeption des zu entscheidenden Problems hat.

4.2 Praktische Kontexteinbettung: Politisch-soziale Dimension und Entscheidungen unter Risiko

4.2.1 Die gesellschaftlich-soziale Dimension von Dilemma-Szenarien

Die Motivatoren, die hinter der anvisierten (Voll-)Automatisierung des Verkehrs stehen, beziehen sich primär auf die Schwächen und Probleme gegenwärtiger Mobilität im Hinblick auf Sicherheits- und Effizienzdefizite sowie veränderte Mobilitätsbedürfnisse der Gesellschaft (siehe Kap. 2.1.2). Die Vision des autonomen Fahrens ist allerdings nicht nur eine Antwort auf bestehende Probleme; sie ist auch disruptiv und bringt ihrerseits weitreichende Wirkungen auf gesellschaftlicher Ebene hervor, die eine mögliche Mobilitätsrevolution nachhaltig mitbestimmen werden:

> […] the current vision of SDV technology […] neglects a range of societal dimensions of technology. It fails to recognize interdependencies between societal dimensions. It does not account for the broader picture of why and how society uses technology in the first place, and how technology continues to influence and shape societal structures and relations. (Blyth et al., 2016, S. 48)

Es ist festzustellen, dass die gesellschaftlich-soziale Dimension autonomer Fahrsysteme im Forschungskontext bisher unzureichend thematisiert worden ist. Dies ist zu einem großen Teil darauf zurückzuführen, dass sich der dominante Forschungszugang auf trolleybasierte Frameworks als Designgrundlage für Dilemma-Szenarien fokussiert. Wie in Kap. 4.1.4 gezeigt, blendet das Trolley-Problem entscheidungstheoretische Aspekte von erheblicher moralischer Relevanz für real-lebensweltliche Unfallsituationen aus, z. B. moralische und rechtliche Verantwortung, (strategische) Interaktion oder die Komplexität des Entscheidungskontextes (vgl. Gogoll & Müller, 2017, S. 690; JafariNaimi, 2018, S. 306; Nyholm, 2018b, S. 5). Ein häufig zitiertes Paradoxon aus den prominenten Studien von Bonnefon et al. (2015, 2016) liefert Hinweise darauf, dass insbesondere das moralphilosophische Design des Trolley-Problems der tatsächlichen Problemstruktur praktischer Unfalldilemmata nicht gerecht wird. Im Rahmen der genannten Untersuchungen wird eine allgemeine Präferenz für eine utilitaristische Programmierung autonomer Fahrzeuge konstatiert, welche sich am Prinzip der Schadensminimierung orientiert. Zugleich würde aber die Mehrheit der Nutzer Fahrzeuge bevorzugen, die sie selbst als Insassen im Notfall schützen (vgl. Bonnefon et al., 2015, S. 5–8).

Diese auf den ersten Blick widersprüchlich anmutende Einstellung wird dann plausibel, wenn man sich vergegenwärtigt, dass es bei der Frage nach der Programmierung von Unfallalgorithmen – anders als beim Trolley-Problem – nicht um ein individuelles Entscheidungsproblem geht. Die Konzeption des klassischen Trolley-Experiments erfordert eine individuelle Entscheidung, indem man sich entweder in den Fahrer (in Foots Version) oder den Zuschauer (in Thomsons Version) hineinversetzt. Die zugrundeliegende Fragestellung impliziert ein moralphilosophisches Entscheidungsproblem; sie fragt danach, wie ein Individuum in der spezifischen Situation entscheiden würde bzw. soll. Aufgrund der großen Resonanz, die auf Instanzen modifizierter Trolley-Szenarien beruhende empirische Studien experimenteller Ethik in den wissenschaftlichen Untersuchungen zu Unfallalgorithmen kontinuierlich erfahren (siehe Kap. 4.3), werden Unfalldilemmata im Forschungsdiskurs mehrheitlich als rein moralphilosophische Entscheidungsprobleme thematisiert.

Dies wird dem real-lebensweltlichen Kontext, in den die Problematik eingebettet ist, allerdings nicht gerecht. Im praktischen Ein-

satz sind autonome Systeme mit komplexen Situationen konfrontiert, die nicht nur von einzelnen Entscheidungsträgern abhängen, sondern mehrere miteinander verflochtene Ebenen sozialer und strategischer Interaktion zwischen verschiedenen beteiligten Parteien beinhalten:

> Technologies are not introduced in a vacuum […] They exist within a vast network of incentives among people, industries, lawmakers, and so on. Harnessing emerging technologies for the benefit of humanity requires a vivid and active imagination, for understanding where the technology will fit into this network and disrupt its existing incentive structure. (Jenkins, 2022, S. 143–144)

Die Verhaltenssteuerung autonomer Fahrzeuge kann nicht als isoliertes Problem betrachtet werden, sondern ist stets Teil eines Systemkonzepts vernetzter Infrastruktur (vgl. Borenstein et al., 2019, S. 386–394; Lundgren, 2021, S. 409) einerseits und gesamtgesellschaftlicher Wirkungen andererseits (vgl. Smith, 2022, S. 279–286). Als sozio-technische Systeme stehen selbstfahrende Fahrzeuge in vielschichtigen gesellschaftlichen Zusammenhängen, in denen Strategien für real-lebensweltliche Entscheidungsdilemmata durch die Implementierung in Algorithmen systemischen Charakter erhalten. Programmierentscheidungen haben Auswirkungen nicht nur auf eine individuelle Situation, sondern auf das Verhalten einer Vielzahl analog implementierter Fahrzeuge; Himmelreich (2018, S. 678) spricht in diesem Zusammenhang von einem »large-scale problem«. Algorithmen können auf einen kontinuierlichen Lösungsraum von Trajektorien zurückgreifen, der ihnen eine größere Handlungsflexibilität ermöglicht (vgl. Geisslinger et al., 2021, S. 1035). Bei bestimmten Situationskonstellationen können so beispielsweise kumulative Effekte hervorgerufen werden (vgl. Liu, 2017, S. 202). Es bleibt unklar, inwiefern sich moralische Entscheidungspräferenzen, die in vereinfachten Szenarien erhoben werden, auf moralische Entscheidungen in komplexen Situationen übertragen lassen (vgl. Lundgren, 2021, S. 408).

Wie das zitierte Paradoxon von Bonnefon et al. (2015, 2016) zeigt, werden utilitaristische Autos zwar als das Mittel der Wahl zur Förderung des Wohls der Allgemeinheit angesehen, jedoch hat jeder Einzelne einen Anreiz, davon abzuweichen, solange er sich durch egoistisches Verhalten besser stellen kann (vgl. Bonnefon et

al., 2016, S. 1575).[104] Dies offenbart ein klassisches soziales Dilemma: Offensichtlich besteht eine Diskrepanz zwischen dem, was Individuen grundsätzlich im Rahmen ihrer persönlichen Wertvorstellungen in einer spezifischen Situation präferieren, und dem, was diese als Grundlage der Programmierung autonomer Systeme im Hinblick auf viele vergleichbare Situationen für sich selbst und andere als wünschenswert erachten. In diesem Sinne führt Černý (2022) aus, dass Entscheidungsstrategien für dilemmatische Situationen weder in einem grundlegenden Widerspruch zu wichtigen moralischen Intuitionen potenzieller Nutzer stehen noch die Grundwerte außer Acht lassen dürfen, die auf der normativen Gleichheit aller Menschen beruhen. Unfalldilemmata stellen daher ein soziales Entscheidungsproblem von gesellschaftlicher Dimension dar. Individuelle Präferenzen sind in diesem spezifischen Kontext zwar nicht irrelevant, als alleinige moralische Orientierung jedoch fragwürdig, da stets Entscheidungen anderer bzw. die Implikationen für andere mitberücksichtigt werden müssen.[105] Vielmehr manifestiert sich die spezifische Problematik von Unfallalgorithmen in einem Spannungsfeld von drei potenziell inkompatiblen Zielsetzungen, denen adäquate Entscheidungsstrategien ganzheitlich entsprechen müssen:

> Not discouraging buyers is a commercial necessity—but it is also in itself a moral imperative, given the social and safety benefits AVs provide over conventional cars. Meanwhile, avoiding public outrage, that is, adopting moral algorithms that align with human moral attitudes, is key to fostering public comfort with allowing the broad use of AVs in the first place. However, to pursue these two objectives simultaneously may lead to moral inconsistencies. (Bonnefon et al., 2015, S. 2)

Wie lassen sich Kompromisse bei zu erwartenden Zielkonflikten finden? Im Folgenden werden Herausforderungen spezifischer regu-

104 Bonnefon et al. (2016, S. 1575–1576) gehen davon aus, dass sich eine Regulierung in dem Sinne, dass eine prinzipiell utilitaristische Programmierung gesetzlich vorgeschrieben wird, kontraproduktiv auswirken würde, da das generelle Sicherheitspotenzial des autonomen Fahrens sich durch die verminderte Kaufbereitschaft nur verzögert entfalten würde.

105 Diesen Aspekt können insbesondere Trolley-Szenarien als rein moralphilosophisch konzipierte Entscheidungsprobleme nicht abbilden; sie sind »merely the unrealistic discrete version of a very real dilemma that emerges at a statistical level.« (Bonnefon et al., 2019, S. 504)

lativer Perspektiven unter Beachtung der pluralistischen Prägung moderner Gesellschaften erörtert.

4.2.2 Politische Regulierung: Unfallalgorithmen im Spannungsfeld zwischen individuellen Präferenzen und pluralistischen Wertvorstellungen

Aus dem Blickwinkel des gesellschaftlichen Kontextes, in den Unfallalgorithmen eingebettet sind, werden in unvermeidbaren Notsituationen grundrechtlich sensible Probleme aufgeworfen, die fundamentale Individualrechte als zentrale politische Werte tangieren. Diese zu schützen, ist Aufgabe politischer Regulierung, nicht moralphilosophischer Überlegungen. Technologische Innovationen regulativ zu begleiten, ist ein anspruchsvoller Auftrag an politische Instanzen, die verschiedene regulatorische und normative Systeme umfassen:

> All in all, the politics of self-driving cars thus concerns a total of five different domains of an upgraded version of ›practical reason‹: politics (in the narrow sense), ethics, law, economics, social norms, and technology. These latter regulatory systems may either reinforce or undermine one another, or even render the claims of another regulatory system superfluous. Depending on how these normative systems interact within a given domain of technological innovation, different observables and variables of the analysis may result. In the field of AVs, existing institutional initiatives and amendments to current AV laws have given rise to four different categories of normative issues vis-à-vis the five different domains of politics (in the broader sense). Such observables of the analysis with their variables regard (1) the ethics of AVs; (2) law and business; (3) the role of social norms; and (4) the governance of technological innovation in the field of AVs. (Pagallo, 2022, S. 163)

Im Rahmen von kollektiven Entscheidungen müssen die Interessen unterschiedlicher Anspruchsgruppen bei der Programmierung von Unfallalgorithmen Berücksichtigung finden.[106] Dabei erweist sich als problematisch, dass liberale Gesellschaften von pluralistischen Wertvorstellungen geprägt sind, die sich in einer weitreichenden

106 Millar et al. (2020) stellen in einem empirisch gestützten Entwurf mögliche Anspruchsgruppen und deren Wertvorstellungen zusammen und bereiten die Ergebnisse für ingenieurtechnische Designaufgaben auf.

moralischen Uneinigkeit bezüglich akzeptierter moralischer Kriterien widerspiegeln. Es gibt keine universale Vorstellung von Moral, die sich in global akzeptierte Maschinenalgorithmen übersetzen ließe (vgl. Maxmen, 2018, S. 469).

Die Anerkennung eines ›vernünftigen Pluralismus‹ (*reasonable pluralism*)[107] und der daraus abgeleiteten Entscheidungsautonomie des Einzelnen bildet den Ausgangspunkt der politischen Philosophie. Vor deren Hintergrund wurden in den letzten Jahren verstärkt alternative Ansätze gewählt, um Dilemma-Szenarien jenseits des Trolley-Problems als Frage legitimierter politischer Regulierung zu konzipieren. Die Problematik der Gestaltung von Unfallalgorithmen bewegt sich als ethisches, aber zugleich auch politisches Entscheidungsproblem im Spannungsfeld von (individueller) Autonomie, (sozialer) Akzeptanz und (moralischer) Akzeptabilität, was den Forschungsdiskurs vor komplexe Herausforderungen stellt. Bisher gehen einige wenige Artikel explizit auf heuristische Ansätze ein, die jeweils unterschiedliche Schwerpunkte bei der Priorisierung der einzelnen Aspekte setzen. Zum gegenwärtigen Zeitpunkt sind diese jedoch eher als Impulse für weitere Forschung denn als ausgearbeitete Konzepte zu werten.

Eine der am häufigsten referenzierten und zugleich umstrittensten Heuristiken in diesem Kontext stellt eine mögliche Personalisierung der ethischen Einstellungen selbstfahrender Fahrzeuge dar, welche die Entscheidungsautonomie der Individuen in den Vordergrund rückt. Dabei wird von einem etablierten Ansatz aus der politischen Philosophie ausgegangen, der dem Problem normativer Inkonsistenzen begegnet, indem er auf eine universale Regelung verzichtet und stattdessen den moralischen Entscheidungsraum aufteilt, sodass der Einzelne die Möglichkeit erhält, nach seinen eigenen normativen Standards zu handeln (vgl. Gogoll & Müller, 2017, S. 687). Im Kontext von Unfallalgorithmen würde dies implizieren, dass jeder Nutzer selbst darüber entscheiden kann, wie sich sein Fahrzeug in einer

107 *Reasonable pluralism* ist ein Begriff, den John Rawls (1993) in seinem späteren Werk zum politischen Liberalismus geprägt hat. Er erkennt an, dass es für Individuen gute Gründe gibt, unterschiedliche Meinungen und Werte zu vertreten. Auf dieser Basis beschreibt er die Existenz und Persistenz einer Diversität unvereinbarer, aber dennoch legitimer moralischer, religiöser oder philosophischer Weltanschauungen als Merkmal moderner demokratischer Gesellschaften.

Notsituation verhalten soll. Das Design entsprechender autonomer Fahrzeuge müsste dann über vorinstallierte Ethik-Module verfügen, die sich an verschiedenen ethischen Theorien bzw. Prinzipien orientieren. Aus diesen können die jeweiligen Nutzer diejenige personalisierte Einstellung (*personalized ethics setting*, *PES*) wählen, die ihren persönlichen moralischen Präferenzen am besten entspricht. In der praktischen Umsetzung könnte dies z. B. durch einen Fragenkatalog erfolgen, um das gewünschte Verhalten für Klassen von Szenarien zu ermitteln (vgl. Fournier, 2016, S. 44). Eine solche Frage könnte lauten, ob die Person ihr eigenes Leben stets priorisieren oder ob sie sich zugunsten einer bestimmten Anzahl an Betroffenen opfern wollen würde. Contissa et al. (2017, S. 371–375) beschreiben ein beispielhaftes mathematisches Modell zur Bestimmung der utilitaristisch optimalen Handlungsoption, das den jeweils individuell gesetzten relativen Wert des eigenen Lebens im Vergleich zu dem anderer berücksichtigt.

Der zentrale Vorteil eines solchen *PES* wäre, dass autonome Fahrzeuge mit individualisierbaren Einstellungen eine höhere Nutzerakzeptanz aufweisen (vgl. Sütfeld et al., 2019, S. 8–9); Formosa (2022, S. 181) bezeichnet dies als das »popularity argument«. Auf diese Weise würde nicht nur paternalistischen Konstrukten eine Absage erteilt (vgl. Millar, 2014c), sondern vielmehr die Entscheidungsautonomie gewahrt und individuellen moralischen Präferenzen entsprochen (vgl. Himmelreich, 2019, S. 35). Wer, wenn nicht der Nutzer selbst, kann und sollte über sein Leben und damit verbundene Risiken entscheiden? Speziell im Kontext möglicher Selbstopferungshandlungen sind fremdbestimmte Entscheidungen problematisch (vgl. Lin, 2013a). Autonome Fahrzeuge würden im Idealfall lediglich als moralische Vertreter (*moral proxies*) fungieren (vgl. Millar, 2015, S. 53–54).[108] Zudem stünde ein solches Vorgehen in Einklang mit den Grundwerten einer liberalen Gesellschaft (vgl. Gogoll & Müller, 2017, S. 688).

108 Ferner zeigen Zhang et al. (2023) in einer empirischen Studie, dass moralische Urteile über die Eignung KI-gestützter Systeme, in dilemmatischen Szenarien moralische Entscheidungen zu treffen, eng mit den kognitiven und emotionalen Prozessen zusammenhängen, die bei unterschiedlichen Szenarientypen im menschlichen Gehirn aktiviert werden.

Einen argumentativ anders begründeten, in der praktischen Realisierung aber letztlich ähnlichen Ansatz entwickeln Shaw und Schneble (2021). Aus der Feststellung, dass sich Stärken und Schwächen von Mensch und Maschine jeweils komplementär ergänzen, folgern sie, dass eine gemeinsame Entscheidungsfindung eine optimale Strategie für unvermeidbare Unfallsituationen darstellt: Maschinen besitzen höhere Kapazitäten zur Datenverarbeitung und haben bessere Reaktionszeiten, wohingegen Menschen prinzipiell ethisch handlungsfähig sind. Nach dem Vorbild der im medizinischen Bereich etablierten Vorgehensweise der gesundheitlichen Versorgungsplanung (*advance care planning*)[109] plädieren sie dafür, dass Nutzer vorab ihre generellen Wertepräferenzen für mögliches Kollisionsverhalten, insbesondere im Hinblick auf Selbstschutz, an das Fahrzeug übermitteln und dieses dann in konkreten Situationen darauf zurückgreift.

Auch wenn Ansätze einer möglichen Personalisierung von Unfallalgorithmen zunächst vielversprechend klingen, gibt es diesbezüglich viele kritische bzw. skeptische Stimmen. Entsprechende Argumentationen lassen sich im Spannungsfeld eines sozialen Zielkonflikts zwischen Selbstbestimmung und Sicherheit verorten, mit dem sich die Nutzer konfrontiert sehen. Das Argument, durch ein *PES* würde die Entscheidungsautonomie der Nutzer gewahrt, lässt sich auch anders interpretieren – nämlich als das Verbot, anderen Personen Schaden zuzufügen und auf diese Weise ihre Autonomie zu beeinträchtigen (vgl. Formosa, 2022, S. 181). Weiterhin lässt sich bemängeln, dass das Verhalten autonomer Fahrzeuge weniger vorhersehbar würde, falls die individuelle Entscheidung über deren Handlungsnormen von Nutzern getroffen wird (vgl. Birnbacher & Birnbacher, 2016, S. 9), was in der Folge zu einer erhöhten Unfallgefahr führt. Millar (2014c) weist auf weitere praktische Limitationen einer vollständigen Personalisierung von Unfallalgorithmen hin:

109 Die gesundheitliche Versorgungsplanung ist ein Beratungskonzept der Gesundheitsvorsorge. Im Kontext des Ansatzes von Shaw und Schneble (2021) dient es als Beispiel für eine gemeinsame Entscheidungsfindung zwischen medizinischem Personal und Patienten. Dabei dokumentieren Patienten ihre persönlichen Werte, Lebensziele und Präferenzen hinsichtlich zukünftiger medizinischer Versorgung. Diese sollen als Orientierung herangezogen werden, falls in zukünftigen Situationen die Entscheidungsfähigkeit der Patienten nicht mehr gegeben oder eingeschränkt ist.

> Yes, we must recognize the importance of letting drivers autonomously express certain preferences. But we must also balance the need for personal autonomy with the severity of the ethical problem posed by the design decision. Asking for driver input in all scenarios would create unreasonable barriers to design and would prevent society from realizing the many other benefits posed by the technology, such as a reduction in overall crashes. Some driver preferences will not be ›serious‹ enough to warrant their input. (Ebd., o. S.)

Weiterhin bringt ein *PES* nicht unerhebliche Herausforderungen für die koordinierte Steuerung mit sich: Wenn für jedes Fahrzeug separat eine personalisierte ethische Einstellung gewählt wird, kann es zu ineffizienten Resultaten kommen; so ist unklar, wie eine optimierte Koordination erfolgen kann, wenn beide Fahrzeugführenden altruistische Präferenzen haben. Lin (2014b) schließlich bewertet ein mögliches *PES* vor dem Hintergrund von Haftung und Verantwortung kritisch. Formosa (2022, S. 182–183) stellt angesichts der Komplexität moralischer Dilemma-Szenarien in Frage, ob im Fall eines *PES* stets von informierten Entscheidungen gesprochen werden kann. Um zu vermeiden, dass die Hersteller für potenziell diskriminierende oder anderweitig moralisch fragwürdige Einstellungen der Nutzer haftbar gemacht werden, müsste sichergestellt werden, dass diese nur innerhalb bestimmter, mit moralischen und rechtlichen Grundsätzen konformer Grenzen personalisierbar sind (vgl. ebd., S. 182; Millar, 2014b). Alternativ wäre auch denkbar, stattdessen die Nutzer selbst in die Verantwortung zu nehmen, was allerdings eine erhebliche rechtliche und moralische Bürde für diese bedeuten würde.[110] Oder sollte gar den Herstellern die Entscheidungshoheit über individualisierte Algorithmen ihrer Produktserien überlassen werden? Anhand einer empirischen Studie demonstrieren Inoue et

110 Hinsichtlich der Entscheidungssituation wäre ein *PES* vergleichbar mit der Trolley-Variante von Foot, in dem der Straßenbahnfahrer als unmittelbar involvierte Person die Entscheidung treffen muss. JafariNaimi (2018, S. 307) betont, dass die relationalen Bindungen, welche Entscheidungsträger mit der Situation und den potenziellen Opfern verbinden, ein wichtiger Faktor bei der Bestimmung der Entscheidungsumstände sind: »There is a difference practically, emotionally, and intellectually to being in charge of the trolley and knowing firsthand about the brakes, the tracks, the terrain, the number of the passengers, and other specifics of the situation as opposed to being a bystander who is making inferences about the situation from a distance.«

al. (2022), dass Individuen dazu tendieren, von etablierten sozialen Normen abzuweichen, wenn ihre Entscheidungen der Öffentlichkeit nicht zugänglich sind. Dies würde Herstellern einen Anreiz bieten, egoistischen Tendenzen durch ihr Produktdesign zu entsprechen, sofern die Gesetzeslage hier einen Spielraum lässt.[111]

Gogoll und Müller (2017) äußern zudem Zweifel, inwiefern ein *PES* aus Sicht eines Individuums überhaupt erstrebenswert ist. Das von Bonnefon et al. (2015, 2016) in utilitaristisch geprägten Nutzerpräferenzen aufgezeigte Paradoxon legt nahe, dass der Einzelne bei einer schadensminimierenden Ausrichtung seine persönliche Sicherheit durch die Wahl einer egoistischen Einstellung maximieren kann, wenn alle anderen altruistische bzw. moralische Einstellungen wählen. Unter der Annahme rationaler Agenten lässt sich dies jedoch auf alle Individuen einer Gesellschaft ausweiten, sodass daraus letztlich ein sozial unerwünschtes Ergebnis resultiert: »[…] there is good reason to believe that morality will become crowded out in a world where people can choose their own ethics setting.« (Gogoll & Müller, 2017, S. 694) Mittels einer spieltheoretischen Analyse zeigen sie, dass eine obligatorische ethische Einstellung (*mandatory ethics setting*, *MES*), die ein egoistisches Abweichen unmöglich macht, tatsächlich nicht nur im Interesse der Gesamtgesellschaft, sondern auch jedes Einzelnen ist (vgl. ebd., S. 689–695).[112] Sobald Individuen zu dieser Erkenntnis gelangen, greift das »popularity argument« nicht mehr zugunsten eines *PES*, sondern vielmehr eines *MES* (vgl. Formosa, 2022, S. 181). Zurückzuführen ist dies vor allem darauf, dass das Resultat einer Situation nicht allein von der Aktion eines einzelnen Fahrzeugs abhängt, sondern diese vielmehr eingebettet ist in einen Kontext strategischer Interaktion. Besonders

111 Siehe hierzu auch den Beitrag von Martin (2017), der verschiedene Antworten auf die Frage diskutiert, wer Beschlüsse über die Programmierung autonomer Systeme für moralische Entscheidungssituationen treffen sollte.

112 Dazu modellieren sie Entscheidungsdilemmata im Stil des bekannten Gefangenendilemmas. Ihr zentrales Argument lautet, dass es nicht nur sozial unerwünscht ist, wenn alle ein egoistisches Setting wählen, sondern auch suboptimal für den Einzelnen, dessen erwarteter Nutzen umso höher ist, je weniger Personen ein egoistisches Setting wählen. Gogoll und Müller (2017, S. 694–695) veranschaulichen sodann entlang vertragstheoretischer Argumentationslinien, dass eine staatliche Regulierung in Form eines verpflichtenden *MES* dem Problem begegnen kann, sofern es auf einer Maxime der Schadensminimierung für alle Beteiligten basiert.

bei Szenarien, in denen eine mögliche Selbstopferung der Insassen relevant wird, sind die strategischen Absichten anderer involvierter Fahrzeuge maßgeblich dafür, ob eine altruistische oder egoistische Einstellung gewählt wird. Formosa (2022, S. 183–187) schlägt einen hybriden Ansatz vor, der sowohl *PES* als auch *MES* integriert und auf diese Weise die Vorteile beider Ansätze bietet. Die Grundlage ist die Implementierung einer Entscheidungsarchitektur, welche ein tendenziell altruistisches *MES* als Standardeinstellung vorgibt. Diese kann durch die Nutzer an individuelle, tendenziell egoistische Präferenzen angepasst werden. Von diesem spezifischen Design verspricht man sich, dass zumindest einige die Standardeinstellung beibehalten werden, sodass die von Gogoll und Müller prophezeite Situation, in der alle die egoistische Variante wählen, vermieden wird.

Das stärkste Argument zugunsten eines *MES* ergibt sich schließlich aus gerechtigkeitsethischen Überlegungen:

> The main argument in favor of an MES is the justice argument, which says that serious calculated harms to others are collective political or justice issues requiring mandated solutions, not personal ethical ones to be left up to each individual to decide. (Ebd., S. 183)

Angesichts dessen erscheint es höchst implausibel, potenzielle Nutzer selbst die ethische Ausrichtung der Unfallalgorithmen ihres Fahrzeugs wählen zu lassen. Vielmehr sollte dies Gegenstand eines demokratischen Entscheidungsprozesses sein, der auf politischer – und nicht individueller – Ebene geführt wird:

> We need government leadership and laws if we are to solve global collective action problems, such as reducing carbon emissions, but also if we are to introduce driverless cars. Humans have a tendency to free ride on the sacrifices of others. We should not let the market decide. If the public is less likely to buy a more ethical driverless car, they can be incentivized or even coerced. Laws and policies are required to prevent the tragedy of the commons and to ensure that risk of harm is minimized to reasonable levels. (Savulescu et al., 2021, S. 656)

Doch wie kann bzw. soll dies praktisch vonstattengehen? Eine alternative, bisher noch unterrepräsentierte Position verfolgt das Ziel einer gesellschaftlichen Konsensfindung über allgemein akzeptierte Werte. Hierzu legt Himmelreich (2018, 2019) einen kontrovers diskutierten Ansatz vor, indem er feststellt, dass Unfallalgorithmen

weniger eine Frage der richtigen (individuellen) Entscheidung als vielmehr eine Frage der richtigen Politik sind:

> A major problem with such trolley cases and other such dilemmas is that they look at these choices as if they were exclusively a moral problem even though they raise a distinctively political problem. Trolley cases ask: What is the right thing to do? What would you do? What should the car do? But instead we need to think more broadly about value pluralism, individual agency, and political legitimacy when developing self-driving cars. Self-driving cars—whether it is about trolley cases or left turns—raise the question of how we get along as a community or people. (Himmelreich, 2019, S. 35)

Vor diesem Hintergrund fordert Himmelreich, dass Unfallalgorithmen primär als Frage einer politischen Regulierung anerkannt werden sollten, die nicht notwendigerweise auf moralische Antworten angewiesen ist.[113]

> A trolley case prompts us to make an individual choice when what we in fact face is a social choice. What seems needed is a kind of compromise to overcome disagreements over issues of value. Insofar as we value the moral diversity of our political community, it should be recognized that autonomous vehicles pose primarily a political problem, not a moral one. (Himmelreich, 2018, S. 676)

Eine Gegenposition hierzu bezieht Keeling (2020). An Himmelreichs Argumentation kritisiert er, dass soziale Akzeptanz zwar eine notwendige, aber nicht hinreichende Bedingung für eine begründete und akzeptable Bewältigung von Dilemma-Situationen ist. Stattdessen geben normethische Argumente letztlich den entscheidenden Ausschlag; z. B. sollten unmoralische Prinzipien nicht implementiert werden, nur weil sie akzeptiert sind:

---

113 Auch Rodríguez-Alcázar et al. (2021, S. 814) bemängeln, dass bisherige Betrachtungen es versäumen, die politische Dimension der Problemstellung anzuerkennen: »[…] although there is room for an ethics of AVs (and even for the discussion of trolley cases within it) which is related not only to the individual decisions of all the relevant actors (software engineers, consumers, lawmakers, carmakers, and others) but also to the elucidation of people's moral intuitions to make them compatible with AVs behavior, the question of the values that ought to guide the design of AVs algorithms and the question of how to adjudicate the unavoidable tradeoffs among them are political questions that are better addressed using political instead of moral criteria.«

> Himmelreich claims that broad societal acceptance is a necessary condition for a successful answer to the moral design problem. This might be true. But it does not follow that our problem is essentially one of aggregating individual tastes, preferences or values. This is true only if broad societal acceptance is both a necessary and sufficient condition for a successful answer to the problem. And there are reasons to accept or reject solutions to the moral design problem which do not pertain to social choice. On one hand, if a collective judgement holds that AVs should act in accordance with immoral principles, then there is a moral reason to reject that solution to the moral design problem. On the other hand, if there is a moral difference between, for example, killing and letting die, then there is a pro tanto reason for this distinction to be reflected in AV decision-making algorithms. This reason has genuine weight irrespective of whether the killing and letting die distinction is reflected in the values of society taken as a whole. (Ebd., S. 304)

Einen vielversprechenden Mittelweg bzw. Kompromiss zwischen den zuvor beschriebenen Standpunkten erarbeiten Brändle und Schmidt (2021). Sie gehen in ihrer Argumentation davon aus, dass normethische Begründbarkeit und gesellschaftliche Akzeptanz nicht als trennscharfe Prozesse zu verstehen sind, sondern vielmehr in einer Weise zusammenwirken, die einer öffentlichen Vernunft (*public reason*)[114] entspricht. Der einschlägige Diskurs, welcher sich mit Unfallalgorithmen aus Sicht der politischen Philosophie beschäftigt, stützt sich vor allem auf die Überlegungen von John Rawls (1971),[115] der zu den führenden zeitgenössischen Theoretikern der öffentlichen Vernunft zählt. Unter Bezugnahme auf dessen Konzepte argumentieren Brändle und Schmidt u. a., dass Entscheidungsstrategien innerhalb der öffentlichen Vernunft als begründet gelten können, wenn sie Teil eines übergreifenden Konsenses moralischer Zugeständnisse sind, die bereits in der politischen Sphäre etabliert und akzeptiert sind:

---

114 Die öffentliche Vernunft stellt in der politischen Philosophie ein moralisches Ideal dar, welches verlangt, dass politische Entscheidungen aus der Sicht jedes Einzelnen vernünftig zu rechtfertigen bzw. zu akzeptieren sind. Sie versteht sich als Versuch, angesichts eines bestehenden *reasonable pluralism* einen gemeinsamen Rahmen für eine legitimierte politische Regulierung zu entwickeln.

115 Zur Diskussion thematisch einschlägiger Fragestellungen sozialer und distributiver Gerechtigkeit auf der Grundlage von Rawls' politischer Philosophie siehe auch Dubljevic und Bauer (2022) sowie Smith (2022).

> Rather, the political justification begins with shared moral (and non-moral) commitments — commitments that are already accepted in the political sphere by all reasonable citizens. Of course, this is only the starting point of the justificatory process, as these commitments still require systematization — only if a given commitment (e.g., a principle of justice or a solution to an AD challenge) is publicly shown to be part of the most plausible system of shared moral commitments is it shown to be justified. This presupposes that there are certain justified political propositions that form a subset of justified moral propositions, namely, shared and coherently systematized moral commitments concerning issues of justice. Solutions to AD challenges, then, must be justified by showing that they belong to this set of justified moral commitments that are shared as such by the members of a specific democratic society. In Rawlsian terms: solutions to AD challenges must be justified within the scope of public reason by showing that they are part of the set of considered judgements that form a full reflective equilibrium based on an overlapping consensus of a specific liberal society. We contend that this means the justification for the solution to AD challenges has been shown to be appealing and acceptable to every reasonable citizen. (Brändle & Schmidt, 2021, S. 1480)

Ebenso wie menschliche Lebensformen an sich unterliegen auch moralische Praktiken und Präferenzen einer kulturellen und zeitlichen Dynamik, wodurch sie im Laufe der Zeit kontinuierlichen Veränderungen unterworfen sind (vgl. Bergmann et al., 2018, S. 4). Deshalb können angewandte Problemstellungen profitieren, wenn der Prozess iterativer Reflexion mit empirischen Daten gestützt wird. Zur Bewältigung normativer Probleme ist weniger relevant, was (soziologisch) akzeptiert, sondern was (ethisch) akzeptabel ist. Daher erscheint es notwendig, gut begründete Positionen auf der Grundlage plausibler Argumentationen zu entwickeln und Intuitionen in einer Weise an neue Umstände anzupassen, die diese mit sozio-politischen Normen und validen ethischen Prinzipien in Einklang bringt (vgl. ebd., S. 2). Eine empirisch informierte, kritische Auseinandersetzung mit moralischen Wertvorstellungen erscheint geeignet, um den Diskurs allgemein akzeptierter Kriterien für die Bewältigung von Dilemma-Szenarien effektiv voranzubringen. Einen entsprechenden Ansatz präsentieren Awad et al. (2020) mit ihrem entwickelten Framework; es integriert die Beteiligung der Öffentlichkeit an Strategien zur Entscheidung moralischer Konflikte im Kontext

der politischen Regulierung von Maschinenalgorithmen als gleichwertiges Instrument neben fundierten Expertenanalysen.

Eine traditionelle, in diesem Kontext relevante Methode aus der politischen Philosophie ist das prominente und zugleich umstrittene Rawls'sche *Überlegungsgleichgewicht* (*reflective equilibrium*). Dieses besagt, dass intuitive moralische Urteile entweder über allgemeine moralische Moralprinzipien oder bestimmte relevante Fälle so lange reflektiert und überarbeitet werden, bis sich systematische Prinzipien ergeben, die sich für die Praxis als moralisch gültige Begründungen erweisen. In diesem Gleichgewichtszustand werden die abgeleiteten Urteile schließlich als stabil und konfliktfrei betrachtet und bieten eine konsistente praktische Orientierung:

> When a person is presented with an intuitively appealing account of his sense of justice (one, say, which embodies various reasonable and natural presumptions), he may well revise his judgments to conform to its principles even though the theory does not fit his existing judgments exactly. He is especially likely to do this if he can find an explanation for the deviations which undermines his confidence in his original judgments and if the conception presented yields a judgment which he finds he can now accept. From the standpoint of moral philosophy, the best account of a person's sense of justice is not the one which fits his judgments prior to his examining any conception of justice, but rather the one which matches his judgments in reflective equilibrium. As we have seen, this state is one reached after a person has weighed various proposed conceptions and he has either revised his judgments to accord with one of them or held fast to his initial convictions (and the corresponding conception). (Rawls, 1971, S. 48)

Allerdings wird Rawls' Konzept primär als ›Rezept‹ für eine individuelle Entscheidungsfindung interpretiert (vgl. Daniels, 1979), das im Hinblick auf praktische Anwendungsprobleme mit sozialer Dimension erweiterungsbedürftig ist. Savulescu et al. (2021) demonstrieren, dass der Kern des Rawls'schen Ansatzes für Unfallalgorithmen fruchtbar gemacht werden kann, wenn empirische Daten über öffentliche Präferenzen anstelle individueller Intuitionen als Ausgangspunkt dienen. Sie entwerfen einen adaptierten Prozess, den sie »Collective Reflective Equilibrium in Practice« nennen und der eine politische Regulierungsentscheidung ermöglicht, die sowohl ethisch vertretbar als auch politisch legitimiert ist.

Der Versuch, die Komplexität real-lebensweltlicher Situationen auf isolierte Entscheidungsprobleme zu reduzieren, welche unter

experimentellen Laborbedingungen kontrollierbar sind, stellt eine fragwürdige Idealisierung dar. Dies gilt nicht nur – wie zuvor gezeigt – in Hinsicht auf den gesellschaftlich-sozialen Kontext, in den das praktische Problem eingebettet ist, sondern auch in Bezug auf epistemische Aspekte der Probleminterpretation selbst. Diese werden nun im Folgenden näher untersucht.

4.2.3 Epistemische Diskrepanzen: Sicherheit, Unsicherheit und Risiko im Kontext von Unfallszenarien

Eine der essenziellsten Fehlinterpretationen, die dominanten Forschungszugängen aufgrund ihrer Fokussierung auf das Trolley-Problem zugrunde liegen, besteht in der Annahme sicherer Handlungskonsequenzen bezüglich des Eintretens bestimmter Umweltzustände. Während das Trolley-Problem auf einem deterministischen Entscheidungsmodell beruht, bei dem die Folgen der jeweiligen Handlungsoptionen als sicher gelten oder es zumindest suggerieren (vgl. Goodall, 2014b, S. 96; JafariNaimi, 2018, S. 306–307), kann dies im Fall von Dilemma-Szenarien hingegen nicht angenommen werden. Zum Zeitpunkt der softwaretechnischen Implementierung besteht Unsicherheit hinsichtlich der zu erwartenden Folgen, die aus einem spezifischen Design der Entscheidungsalgorithmen in der realen Lebenswelt potenziell resultieren werden. Diese Unsicherheit entsteht dabei nicht aus der zeitlichen Entkoppelung von Entscheidungsfindung und späterer Manifestation, denn die Zukunft ist immer unsicher. Vielmehr sind es Unsicherheitsfaktoren des spezifischen Kontextes von Unfallalgorithmen, die in der Design- und Implementierungsphase mittels Wahrscheinlichkeitsprognosen und Schätzungen von erwarteten Schadenshöhen antizipiert werden.

Die Abschätzung wird dabei durch diverse Faktoren erschwert. Zum einen operieren autonome Fahrzeuge in einer dynamischen Umgebung, die durch schwer antizipierbares Verhalten anderer Verkehrsteilnehmer und variable Umweltzustände bestimmt wird. Borenstein et al. (2019, S. 387–390) merken an, dass sich lediglich das Verhalten autonomer Fahrzeuge durch Vernetzungsmechanismen bis zu einem gewissen Grade koordinieren lässt. Im Mischverkehr dagegen hängt das Ergebnis einer Situation wesentlich von der Interaktion zwischen verschiedenen Verkehrsteilnehmern ab. In diesem Sinne verweisen Dilich et al. (2002, S. 245) auf eine ältere Studie von

Lechner und Malaterre (1991), die zu dem Schluss kommen, dass das Ergebnis einer Situation vor allem in Notsituationen völlig ungewiss ist und das Verhalten der zu schützenden Verkehrsteilnehmer die Konsequenzen erheblich mitbeeinflusst. Extreme Witterungsbedingungen können zusätzlich die Qualität und Zuverlässigkeit generierter Sensordaten einschränken.

Zum anderen sind aus technischer Sicht schlichtweg Grenzen hinsichtlich der Berechenbarkeit möglicher Handlungsfolgen gesetzt. Eine niemals vollständig eliminierbare Fehleranfälligkeit des technischen Systems des Fahrzeugs und eventuelle Programmierfehler stellen potenzielle Unsicherheitsfaktoren dar. Keeling (2019, S. 51) weist darüber hinaus auf qualitative Mängel in der Objekterkennung gegenwärtig verfügbarer Systeme hin. Sie führen dazu, dass autonome Systeme bislang weder Objekte eindeutig klassifizieren noch die Anzahl involvierter Personen, z. B. der Insassen eines Unfallfahrzeugs, zuverlässig bestimmen können. Aufgrund von qualitativ unzureichenden Perzeptionstechnologien verfügen autonome Fahrsysteme nicht nur über unvollständige Informationen hinsichtlich möglicher Folgen, sondern auch in Bezug auf die Situation selbst, in der sie eine bestimmte Aktion ausführen sollen (vgl. Keeling, 2020, S. 300). Aus technischer Sicht kommt es bei der Klassifizierung erfasster Objekte zu einer Abwägung moralisch relevanter Ziele hinsichtlich Sicherheit einerseits und Zeiteffizienz andererseits. Nur wenn das System ein erkanntes Objekt mit hinreichender Wahrscheinlichkeit z. B. als Fußgänger klassifiziert, ist die Priorisierung von Sicherheits- über Effizienzaspekte und damit eine Notbremsung im Hinblick auf den Zielkonflikt moralisch gerechtfertigt.

Unklar ist jedoch, was aus moralischer Sicht als hinreichend wahrscheinlich gelten kann (vgl. Keeling, 2022, S. 47–53). In dieser Hinsicht unausgereifte technische Komponenten waren beispielsweise mitverantwortlich für den tragischen Unfall im März 2018 in Arizona, als das System die tödlich verunglückte Elaine Herzberg zunächst falsch klassifizierte und deshalb zu spät eine Notbremsung einleitete. Unsicherheiten in der korrekten Interpretation von Umgebungsobjekten sind moralisch signifikant, wenn es darum geht, die Folgen gewählter Trajektorien zu bewerten. Auch Wahrscheinlichkeiten und die Schwere der zu erwartenden physischen Schäden lassen sich unter Bezugnahme auf sensorisch erfasste Situationsmerkmale nicht zuverlässig berechnen:

> AVs have fallible sensors. From the AV's point of view, there are different ways the world might be, and the outcome of a collision depends on both the AV's action and the true state of the world. For example, the AV might be uncertain about the behaviors of pedestrians or it might be uncertain about morally relevant characteristics of the affected parties such as age and physical condition. (Keeling et al., 2019, S. 50)

Grundsätzlich sind dabei sowohl Entscheidungen unter *Risiko*[116] (bekannte Eintrittswahrscheinlichkeiten) als auch unter *Ungewissheit* (unbekannte Eintrittswahrscheinlichkeiten) denkbar. Gegenwärtige technische Systeme ermöglichen es, Szenarien mittels probabilistischer Bewertungsmethoden und stochastischer Entscheidungsmodelle in Simulationsumgebungen zu erproben. Daher wird im Forschungsdiskurs mehrheitlich davon ausgegangen, dass die Eintrittswahrscheinlichkeiten grob bestimmbar sind; Programmierentscheidungen über Unfallalgorithmen lassen sich somit als risikobehaftet klassifizieren. In der Folge kann berechtigterweise davon ausgegangen werden, dass Entscheidungen in Dilemma-Situationen stets unter Bedingungen unvollständiger Information und Unsicherheit getroffen werden (vgl. Goodall, 2016a, S. 813; Wolkenstein, 2018, S. 168). Diese Feststellung hat bedeutende Implikationen für die Auseinandersetzung mit der ethischen Thematik rund um Unfallalgorithmen: Unsicherheiten bezüglich des Eintretens möglicher Umweltzustände bzw. konkreter Handlungsfolgen sind moralisch relevante entscheidungstheoretische Charakteristika, die nicht ausgeblendet werden dürfen.

Mit Unsicherheiten behaftete Entscheidungsprobleme sind aus normativer Sicht grundsätzlich anders zu bewerten als solche, die von sicheren Handlungsfolgen ausgehen; es besteht ein kategorischer Unterschied in der normativen Betrachtung entsprechender Situationen:

> […] the moral reasoning that somebody facing a trolley case uses is not about risks and how to respond to different risks. Nor is it about how to make decisions in the face of uncertainty. This is a categorical difference between trolley-ethics and the ethics of accident-algorithms for selfdriving cars. Reasoning about risks and uncertainty is categorically different from reasoning about known facts and certain outcomes. The key concepts used differ drastically in what inferences they warrant.

116 Auf den Begriff des Risikos wird in Kap. 6.2.1 näher eingegangen.

> And what we pick out using these concepts are things within different metaphysical categories, with different modal status (e.g. risks of harm, on one side, versus actual harms, on the other). (Nyholm & Smids, 2016, S. 1286)

Empirische Studien stützen diese These. So stellen Meder et al. (2019) fest, dass Studienteilnehmer tendenziell ein standardisiertes Verhalten wie eine Notbremsung bei gleichzeitigem Spurhalten bevorzugen, welches mit geltenden Verkehrsregeln konform ist, wenn Unsicherheiten über mögliche Folgen bestehen. Dies gilt auch, wenn alternative Trajektorien einen geringeren Schaden versprechen. Melo et al. (2020) heben hervor, dass Verletzungsrisiken sowohl für Fahrzeuginsassen als auch für externe Personen eine große Rolle für die jeweiligen moralischen Präferenzen spielen. Auch individuelle Einstellungen zur Risikoaffinität bzw. Risikoaversion fließen ein. Daraus lässt sich für die zu untersuchende Problemstellung folgern, dass die ethische Bewertung von risikobehafteten Entscheidungen nicht auf idealisierte, unter Annahme von Sicherheiten konzipierte Beispielfälle reduziert werden darf, wie sie Instanzen des Trolley-Problems darstellen:

> Of the various moral principles that have emerged from the now four-decades-long preoccupation with trolley problems, none can handle the problem of garden-variety risk. As a result, trolleyology is at best engaged in what amounts to a moral sideshow. (Fried, 2012, S. 506)

Hinzu kommt ein bekanntes und oftmals kritisiertes Problem der Angewandten Ethik als solche: Die moralische Relevanz von Risiko und Unsicherheit findet in der klassischen Ethik und Moralphilosophie keine systematische Berücksichtigung; durch Handlungen kausal verursachte Folgen und Umweltzustände werden vielmehr als vollständig bekannt und eindeutig angesehen. Streng genommen sind jedoch real-lebensweltliche Entscheidungen immer mit Unsicherheiten bezüglich ihrer Folgen behaftet.[117] Bei Problemstellungen, bei denen Risiken und Unsicherheiten eine zentrale ethische Relevanz besitzen – wie bei der Erprobung innovativer, risikobehafteter Technologien – stoßen etablierte Moraltheorien jedoch an ihre Grenzen; es ist unklar, ob und wie sich traditionelle normative An-

117 Ein weiterer Aspekt in diesem Kontext ist, dass zukünftige Ereignisse selbst dann nicht zwangsläufig für Akteure zu erkennen sind, wenn sie bereits als eindeutig (kausal) festgelegt betrachtet werden.

sätze auf unsichere bzw. risikobehaftete Entscheidungen übertragen lassen (vgl. Himmelreich, 2018, S. 677; Lundgren, 2021, S. 406–407; Nyholm & Smids, 2016, S. 1284–1286). »Uncertainty is a ubiquitous feature of life. Decisions that are easy to make under certainty can become much more difficult and morally fraught under uncertainty«, konstatieren Bjorndahl et al. (2017, S. 2). Ein prominenter und auch im Kontext von Unfallalgorithmen häufig herangezogener Ankerpunkt ist hierbei Sven Ove Hanssons Kritik an der Standardmoralphilosophie, die unfähig ist, mit den Risiken und Unwägbarkeiten vieler ethischer Fragen der Lebenswelt angemessen umzugehen. Hansson bemängelt in diesem Zusammenhang u. a. die defizitäre Konzeption menschlichen Handelns (vgl. 2013, Kap. 3) und deren deterministische Annahmen (vgl. 2003, S. 291–292, 2013, Kap. 2). Die daraus erwachsenden Schwächen traditioneller Moraltheorien sind im Hinblick auf praktische Anwendungsfragen gravierend:

> To someone whose focus is set on the moral problems served at the philosophy department's seminar table, the inability of common moral theories to deal with risk and uncertainty may seem like one of the many small failures that keep the philosophical discussion alive and well. Unfortunately it is much worse than that. Outside of the seminar room, uncertainty about the effects of one's actions is a ubiquitous and often dominant element in the moral problems that we face in both private and public life. In order to make moral theory practically useful, we need to develop workable methods to analyse the ethical aspects of decisions under risk and uncertainty. (Hansson, 2013, S. 43)

Vor diesem Hintergrund erscheint es höchst fragwürdig, die Programmierung von Unfallalgorithmen unreflektiert mittels spezifischer, für Entscheidungen unter Sicherheit konzipierter Entscheidungsansätze vorzunehmen, wie es im Rahmen von trolley-basierten Frameworks regelmäßig geschieht.

An dieser Stelle sind die Determinanten des Kontextes, in dem das Anwendungsproblem steht, nun hinreichend expliziert worden. Im Rahmen der folgenden beiden Unterkapitel werden sodann Ansätze kritisch reflektiert, die unter bisher dominanten Forschungszugängen vorgelegt worden sind. Dabei werden zunächst deskriptive Ansätze in den Blick genommen.

4.3 Deskriptive Ansätze: Perspektiven aus der Moralpsychologie

4.3.1 Moralische Präferenzen der Öffentlichkeit im Fokus einer experimentellen Ethik

Die Reaktionen autonomer Fahrzeuge, die in dilemmatischen Unfallsituationen aktiviert werden, stehen im Zusammenhang mit einer technologischen Innovation, die in ausweichlichen Unfallsituationen eine situativ-menschliche durch eine algorithmische Entscheidung ersetzt. Von großer Bedeutung für die Entwicklung und Einführung dieser Technologie ist die Akzeptanz, die potenzielle Nutzer ihr entgegenbringen (vgl. Karnouskos, 2020). Um diesem Umstand Rechnung zu tragen, wurden Dilemma-Szenarien des autonomen Fahrens in den letzten Jahren verstärkt mittels deskriptiver Ansätze aus dem Bereich der Moralpsychologie, insbesondere Methoden experimenteller Ethik, untersucht. Zweck dieser Vorgehensweise ist es, moralische Intuitionen, Urteile und Verhaltensweisen im Kontext von Unfallalgorithmen empirisch zu erforschen. Mithilfe der auf diese Weise ermittelten moralischen Präferenzen der Studienteilnehmer sollen Rückschlüsse über das öffentlich akzeptierte und erwünschte Verhalten autonomer Fahrzeuge gezogen werden, um deren Akzeptanz nachhaltig zu erhöhen.

Die Autoren einschlägiger Studien sind der Überzeugung, dass eine experimentelle Ethik relevante Erkenntnisse für die Programmierung von Unfallalgorithmen liefern kann. So betonen Awad et al. (2018, S. 59), dass datengestützte experimentelle Methoden geeignet sind, die moralischen Präferenzen der Öffentlichkeit im Sinne eines partizipatorischen Paradigmas zu erfassen. Ihnen kommt eine wichtige Rolle für die gesellschaftliche Akzeptanz von Innovationen zu, denn im Grunde werden alle Bürger von den Implikationen der Verkehrsautomatisierung betroffen sein, ob als Passagiere oder potenziell gefährdete Verkehrsteilnehmer (vgl. Krügel & Uhl, 2022, S. 2).

Zu den meistzitierten Forschungsartikeln des einschlägigen Diskurses zählen empirische Studien aus statistischen Laborexperimenten, die hypothetische Dilemma-Szenarien als Instanzen eines modifizierten, angewandten Trolley-Problems konstruieren. Ihr Ziel ist es, individuelle moralische Präferenzen bei Fahrentscheidungen in verschiedenen Szenarien mittels datengestützter Ansätze zu analy-

sieren.[118] Der Aufbau des klassischen Trolley-Problems wird dabei auf stochastische Szenarien übertragen, die dem Aspekt Rechnung tragen, dass es sich bei real-lebensweltlichen Unfallsituationen um Entscheidungen unter Risiko handelt. Die bis dato einflussreichsten Arbeiten auf diesem Gebiet wurden von einer Forschergruppe bestehend aus Psychologen und Verhaltensökonomen renommierter Universitäten initiiert.[119] Sehr prominent ist das sogenannte *Moral Machine Experiment* (vgl. Awad et al., 2018), eine Online-Experimentalplattform zur Erfassung umfangreicher Daten über moralische Präferenzen in spezifischen Dilemma-Szenarien, mittels derer bis zur Auswertung mehr als 40 Millionen Antworten registriert wurden. Im Stil philosophischer Narrative wurden den Teilnehmern nacheinander verschiedene modifizierte Trolley-Szenarien mit jeweils zwei möglichen Ausgängen präsentiert. Die untersuchten Szenarien thematisieren eines oder mehrere der folgenden neun Entscheidungsprobleme: präferierte Schonung von Menschen (gegenüber Tieren), Insassen (gegenüber Fußgängern), höherer Anzahl von Leben (gegenüber geringerer), Frauen (gegenüber Männern), Jüngeren (gegenüber Älteren), Gesünderen (gegenüber Ungesünderen), Menschen von höherem sozialen Status (gegenüber niedrigerem), Fußgängern mit regelkonformem Verhalten (gegenüber regelwidrigem Verhalten) sowie Spurhalten (gegenüber Ausweichen). Ziel des Experiments ist es einerseits, die relative Häufigkeit von moralischen Präferenzen hinsichtlich dieser Problemfelder zu bestimmen, und andererseits mögliche Zusammenhänge mit spezifischen persönlichen Charakteristika der Teilnehmenden zu untersuchen. Als wichtigstes Ergebnis sind drei Aspekte festzuhalten, zu denen die stärksten Präferenzen vorlagen: der Schutz von Menschen gegenüber Tieren, der höheren Anzahl von Leben gegenüber der geringeren und

118 Die einschlägigen empirischen Studien beziehen sich beinahe ausschließlich auf das Geschehen *vor* einer unabwendbaren Kollision. Dass auch das Verhalten autonomer Fahrzeuge unmittelbar *nach* einem Unfall eine ethische Dimension hat, erläutern z. B. Krügel et al. (2021), indem sie auf eine bestehende Regulierungslücke diesbezüglich verweisen.

119 Die Studien schließen sich an die Pionierarbeit von Joshua Greene auf dem Gebiet der Neuroethik an, der die Hirnaktivität von Entscheidungsträgern in moralischen Dilemma-Situationen erstmals um die Jahrtausendwende untersuchte (vgl. Greene, 2013).

junger Menschen (insbesondere von Kindern) gegenüber älteren.[120] Diese ermittelten Tendenzen sollten, so die Empfehlung der Autoren, zentrale Bausteine für Unfallalgorithmen darstellen (vgl. Awad et al., 2018, S. 60).

In einer anderen Studie weisen Frank et al. (2019) nach, dass normenkonformes Verhalten eine gewichtige Rolle bei der moralischen Bewertung von Dilemma-Szenarien spielt; so werden ›Regelbrecher‹ als deutlich weniger schutzwürdig angesehen als andere Gruppen. Altay et al. (2023) identifizieren den sozialen Status als wichtigste Determinante moralischer Entscheidungen, wohingegen dem Geschlecht potenzieller Betroffener die geringste Bedeutung zukommt. Der Fokus des Studiendesigns von Lucifora et al. (2021) liegt darauf, den Einfluss zeitlicher Restriktionen auf die Entscheidungsfindung in Dilemma-Szenarien zu analysieren. Ihre Auswertungen offenbaren, dass bei schnellen Entscheidungen vor allem risikoanalytische Methoden die besten Ergebnisse erzielen, wohingegen bei bewussten Entscheidungen moralische Beurteilungen Vorrang haben. Auch eine Studie von Sütfeld et al. (2017), die verschiedene Verhaltensmodelle aus psychologischer Perspektive testet, legt nahe, dass ethische Entscheidungen sowohl derselben Person als auch im interpersonellen Vergleich mit zunehmendem Zeitdruck inkonsistent werden.

Zu den Schwerpunkten weiterer relevanter experimenteller Studien zählt die Untersuchung der altruistischen Bereitschaft der Insassen autonomer Fahrzeuge, sich im Notfall selbst zu opfern. Frühere Studien der prominenten Forschergruppe um Jean-François Bonnefon, Azim Shariff und Iyad Rahwan (2015, 2016) widmen sich der moralischen Beurteilung von Akten der Selbstopferung sowie Erwartungen und Kaufbereitschaft im Hinblick auf eine mögliche gesetzliche Verankerung derselben. Um die Dilemma-Szenarien für

120 Zudem wurden mittels einer anhand demografischer, geografischer und kultureller Merkmale durchgeführten Clusteranalyse drei ›moralische Cluster‹ mit homogenen Vektoren moralischer Präferenzen gebildet: ein westlicher Cluster bestehend aus Nordamerika und christlich geprägten europäischen Ländern, ein östlicher Cluster mit Ländern des konfuzianistischen Kulturkreises wie Japan und Taiwan sowie islamischen Ländern wie Indonesien, Pakistan und Saudi-Arabien, und ein südlicher Cluster mit lateinamerikanischen Ländern Süd- und Mittelamerikas sowie ehemaligen französischen Herrschaftsgebieten. Ferner wurde untersucht, inwiefern moralische Präferenzen mit spezifischen kulturellen und ökonomischen Faktoren in den jeweiligen Clustern korrelieren (vgl. Awad et al., 2018, S. 61–63).

die Befragten realistisch erfahrbar zu machen, bedienen sich neuere einschlägige Studien immersiver Technologien. Frison et al. (2016) untersuchen anhand eines Studiendesigns mit physischem Fahrsimulator spezifische Einflussfaktoren auf die individuelle Bereitschaft zur Selbstopferung. Bedeutsamstes Ergebnis ihrer Studie ist die Erkenntnis, dass einerseits die Häufigkeit altruistischer Handlungen steigt, je höher die eigene Überlebenswahrscheinlichkeit ist, je mehr Betroffene es gibt und je jünger diese sind, während es andererseits unerheblich ist, ob es sich bei den Betroffenen um enge Freunde oder Fremde handelt.

Im Rahmen eines durch Virtual-Reality-Technologien realisierten Fahrsimulationsexperiments erforschen Bergmann et al. (2018) sowie Faulhaber et al. (2019), welchen Einfluss moralisch relevante Faktoren auf die Entscheidungsfindung der Befragten haben, z. B. das Alter potenzieller Opfer, die Bereitschaft zur Selbstopferung oder der Grad der Beteiligung am Unfallgeschehen. Auf Basis der ausgewerteten Daten folgern sie, dass die Befragten tendenziell auch dann altruistisch im Sinne eines quantitativen Gemeinwohls entscheiden, wenn dies der eigenen Selbsterhaltung entgegensteht (vgl. Bergmann et al., 2018, S. 6–7; Faulhaber et al., 2019, S. 407–413). Dies ist konsistent mit den Ergebnissen von Wintersberger et al. (2017), die zeigen, dass viele Teilnehmende grundsätzlich die Bereitschaft haben, sich selbst zugunsten unbeteiligter Fußgänger zu opfern. Eine Studie von Bruno et al. (2023a) belegt, dass eine Zurückweisung der Selbstopferung verbunden mit dem Erlangen eines persönlichen Nutzens selbst dann als unmoralisch empfunden wird, wenn dies im Sinne einer utilitaristischen Ausrichtung von Unfallalgorithmen das ›bestmögliche Ergebnis‹ darstellt.

4.3.2 Zur Relevanz deskriptiver Methoden: Eine Kritik

Trotz ihrer erheblichen Bedeutung für den Forschungsdiskurs stehen die verwendeten Ansätze experimenteller Ethik immer öfter in der Kritik – zum einen aus methodologischer Sicht, zum anderen

aufgrund ihrer zunehmend normativen Ausrichtung.[121] Insbesondere am *Moral Machine Experiment*, welches bis heute bereits mehr als eintausend Mal zitiert wurde, entzündet sich eine kontroverse Debatte, inwiefern eine solche Methodik tatsächlich einen Beitrag zur Gestaltung von Unfallalgorithmen leisten kann (vgl. z. B. Harris, 2020). Die von Kritikern aufgeführten Argumente beziehen sich auf zwei zentrale Schwächen, die experimentellen Ansätzen zugrunde liegen. Die erste betrifft das methodische Vorgehen, Entscheidungsstrategien für normative Probleme auf der Grundlage empirisch ermittelter moralischer Präferenzen zu begründen. Empirische Methoden sind im Kern deskriptiv; sie spiegeln falsifizierbare soziologische Tatsachen wider, die sich jedoch nicht ohne Weiteres in wahrheitsfähige moralische Urteile oder Normen überführen lassen. Letztere aus beobachteten Fakten zu folgern, käme einem naturalistischen Fehlschluss gleich: Aus dem Sein folgt kein Sollen. Die Frage, was die meisten von uns in einer bestimmten Situation intuitiv tun *würden*, ist unabhängig davon, was wir richtigerweise tun *sollten*; nur Letzteres ist Gegenstand der Ethik. Soziale Akzeptabilität ist kein genuin ethisches Kriterium und sollte nicht als (alleinige) Legitimationsgrundlage für die Programmierung von Unfallalgorithmen herangezogen werden (vgl. LaCroix, 2022, S. 4–5) – auch wenn experimentell ermittelte Präferenzen fälschlicherweise als Ausdruck einer öffentlichen Moral aufgefasst werden, wie es im *Moral Machine Experiment* der Fall ist:

> Majorities are not necessarily right; neither science nor ethics is produced by casting votes for particular ›answers‹; happy though such a possibility might seem to some! The Moral Machinists are proposing the moral equivalent of deciding whether the world is flat by finding out what people would prefer the answer to be. (Harris, 2020, S. 74)

Der Technikphilosoph Armin Grunwald, Mitglied der ehemaligen Ethik-Kommission und seit 2021 des Deutschen Ethikrats, schlussfolgert in ähnlicher Weise:

> Weder aus Spielen noch aus Umfragen kann etwas über die ethische Zulässigkeit von Normen gelernt werden. Ansonsten könnte nach jedem schweren Verbrechen eine Umfrage gemacht werden, die mit ziemlicher

121 Für einen Überblick über relevante Kritik am *Moral Machine Experiment* und an vergleichbaren Experimenten aus methodologischer und normativer Sicht siehe z. B. Paulo et al. (2023, S. 293–302).

> Sicherheit für die Einführung der Todesstrafe ausgehen würde. Ethik und Recht bedürfen anderer Quellen der Rechtfertigung, wie zum Beispiel einem gehaltvollen Menschenbild. (Science Media Center Germany, 2018)

Weiterhin neigen empirisch ermittelte moralische Intuitionen dazu, inkonsistent und weniger ethischen als vielmehr moralpsychologischen Ursprungs zu sein (vgl. Bruers & Braeckman, 2014, S. 266–267). Robinson et al. (2022, S. 444–445) verweisen auf eine Diskrepanz zwischen »what we say and what we do«, die die Verlässlichkeit von in hypothetischen Szenarien dokumentierten Präferenzen in Frage stellt. Diverse Forschungsarbeiten bestätigen, dass moralisch irrelevante Faktoren die Antworten von Studienteilnehmern verzerren können, beispielsweise spezifische Stimmungslagen oder affektive Einflüsse (vgl. Cao et al., 2017; Pastötter et al., 2013). Auch das spezifische Design der Experimente kann die Glaubwürdigkeit der Ergebnisse erschüttern: Spielt die Reihenfolge, in der die Szenarien präsentiert werden, eine (psychologische) Rolle? Oder die Perspektive, welche die Befragten einnehmen?[122] Sind moralische Präferenzen, etwa die Bevorzugung von Kindern aufgrund ihres jungen Alters, in jeder Situation unabhängig vom notwendigen Grad des Intervenierens?

Smith (2019, S. 120–122) stellt fest, dass Persönlichkeitsmerkmale bzw. Eigenschaften der Handelnden einerseits sowie deren ethische Einstellungen andererseits die individuellen Vorstellungen dahingehend maßgeblich beeinflussen, wie autonome Fahrzeuge in kritischen Situationen agieren sollten. Eine entscheidende Rolle spielt ebenfalls der Faktor ›Unsicherheit‹; Sensibilität gegenüber spezifischen Schadenswahrscheinlichkeiten und persönliche Risikoeinstellungen können moralische Präferenzen in nicht unerheblichem Maße beeinflussen (vgl. Schuessler, 2024). Ein Studiendesign, das dem Trolley-Paradigma folgt, zeugt ferner von naiven Annahmen über das Wesen der Moralität, die Aspekte wie den Charakter oder die

122 Die Zugänglichkeit verschiedener Perspektiven (*perspective-taking accessibility*) stellt einen wichtigen Faktor bei der Beeinflussung moralischer Urteile dar, der auch im Kontext von Unfallszenarien eine nicht-triviale Rolle spielt (vgl. Bruno et al., 2023b; Kallioinen et al., 2019; Mayer et al., 2021; Othman, 2023). Operationalisiert wurde dies vor allem anhand von experimentellen Studienkonzepten, die auf dem Rawls'schen *Schleier des Nichtwissens* basieren (siehe auch Kap. 7.3.3.3).

Intentionen der Handelnden außer Acht lassen: »Binary choice models are well-suited for experimentation since they enable the cut-and-dry variation of a variable. However, trolley-like dilemmas only permit deontic or utilitarian evaluations, failing to consider other important factors influencing moral judgment.« (Cecchini et al., 2023, S. 4)

Empirisch gewonnene Erkenntnisse lassen sich daher nur begrenzt auf real-lebensweltliche Kontexte übertragen (vgl. Siegel & Pappas, 2023, S. 224);[123] Cecchini et al. (2023) sprechen in diesem Zusammenhang von einer begrenzten ökologischen Validität:

> Another critical limitation of the trolley paradigm (and the MME particularly) is its lack of *ecological validity,* namely the extent to which some experimental results can be generalized to explain a wide range of real-life situations. In particular, [...] experiments based on trolley cases do not have sufficient *experimental, mundane,* and *psychological* realism. (Ebd., S. 4, Hervorh. i. Orig.)

Schlussendlich können normative Schlussfolgerungen nicht allein auf der Grundlage deskriptiver Argumente gezogenn werden, sondern diese nur unterstützen:

> Normative conclusions must be supplied by ethical theories. The empirical investigation only yields which of these theories is more aligned with society's practices and people's intuitions, or more specifically which factors are recognized by people in making moral decision. The empirical investigation may yield certain insights about which theory is preferable, but the normative significance is mainly derived from the theories themselves. (Bergmann et al., 2018, S. 4)

Einen weiteren hier einschlägigen Aspekt, der mit den strukturellen Unstimmigkeiten zwischen Trolley-Problem und Dilemma-Szenarien zusammenhängt, beschreibt Lundgren (2021, S. 407–409) als

123 Es ist umstritten, inwiefern Moral in virtuellen Umgebungen, wie sie im Rahmen der oben beschriebenen Experimente und Studien eingesetzt wurden, mit der Moral in der physischen Realität vergleichbar ist. Kenwright (2018, S. 21) merkt hier an, dass traditionelle moralische Verantwortungsmodelle sich nicht immer auf die digitale Welt übertragen lassen. Für weitere Untersuchungen zu diesem Themenfeld siehe z. B. Dunn (2012), McMillan und King (2017) sowie Ramirez und LaBarge (2018). Cecchini et al. (2023) schlagen ein alternatives experimentelles Studiendesign vor, das realistische Entscheidungssituationen durch Virtual-Reality-Umgebungen simuliert und das sogenannte *agent-deed-consequences (ADC)*-Modell als moralpsychologisches Framework integriert.

Inkongruenz zwischen Mensch und Maschine. Im Anschluss an maschinenethische Überlegungen (siehe Kap. 4.1.2) können Entscheidungsprozesse künstlicher Systeme nicht als analog zu menschlichen angesehen werden:

> Simply put, it is not evident that human preferences can be translated into rules for a machine. This is because choice-descriptions from a human and a machine perspective differs [sic] and may be incongruent. Indeed, the machines may both lack information humans have and vice versa, or the machine descriptions may be incompatible with human descriptions of reality, possibly making a translation impossible. Thus, it is not obvious that we can construct machines [sic] rules that satisfy the surveyed preferences, which potentially would provide a problem for policies based on such preferences. (Ebd., S. 407)

Zusammenfassend lässt sich festhalten, dass deskriptive Ansätze für die kollektive moralische Problemstellung des Designs von Unfallalgorithmen nur begrenzt hilfreich sind. Um direkte Implikationen für die Programmierung entsprechender Algorithmen abzuleiten, sind empirische Studien demnach nicht geeignet – insbesondere dann nicht, wenn sie dem Trolley-Problem nachempfunden sind:

> Philosophers use stylized tasks to analyse the complex and uncertain situations in which moral choices are actually made. Dilemmas have no meaning outside such discourse. Although survey responses might stimulate enquiry, taking them literally is an antithesis to philosophical practice. (Dewitt et al., 2019, S. 31)

Empirische Erkenntnisse über moralische Werturteile können lediglich als informierte Orientierungshilfe fungieren, um weiterführende Ansätze für die Gestaltung von Entscheidungsstrategien auf den Weg zu bringen. So können Methoden experimenteller Ethik für die Debatte über Unfallalgorithmen fruchtbar gemacht werden, indem sie in deskriptiver Weise die individuellen moralischen Präferenzen potenzieller Nutzer und damit auch die Erwartungen an autonome Fahrsysteme herausarbeiten. In dieser Hinsicht leisten sie einen wertvollen Beitrag zum Forschungsdiskurs, denn gänzlich ignoriert werden dürfen öffentliche Moralvorstellungen ebenfalls nicht (vgl. Savulescu et al., 2021, S. 655–656). Jedoch können sie aufgrund des bestehenden Wertepluralismus moderner Demokratien nicht unmittelbar in normative Richtlinien bzw. eine finale Legitimierung politi-

scher Entscheidungen übertragen werden (vgl. Brändle & Schmidt, 2021, S. 1491):

> As far as surveys of moral judgment are concerned, we admit that their outcomes can be also interesting, provided that we do not ask them for what they cannot deliver. They are not going to solve any moral dilemma by providing the correct answer (remember the is/ought fallacy); nor will they provide direct guidance for congresspersons. But they can provide useful information about the moral values and opinions of the members of the community that, together with further information concerning other social facts, may be considered by lawmakers when regulating AVs. […] The empirical knowledge that surveys can provide on these specific topics can be relevant not for the specification of the political ends but for designing the best means to achieve them. (Rodríguez-Alcázar et al., 2021, S. 829)

Insgesamt betrachtet legen die Ergebnisse der kritischen Reflexion deskriptiver Ansätze nahe, dass das Anwendungsproblem eine stärker normativ orientierte Herangehensweise erfordert. Mit der Verwendung des Trolley-Problems als dominantes Framework geht als zentrale Fragestellung des (normativen) Designproblems einher, welche ethischen Prinzipien der Entscheidung für eine der Handlungsoptionen zugrunde gelegt und als Teil des algorithmischen Entscheidungsprozederes implementiert werden sollen. Im Folgenden werden Begründungsversuche aus Sicht verschiedener ethischer Denkschulen zusammengetragen und im Hinblick auf ihre Eignung für das Anwendungsproblem kritisch geprüft.

4.4 Normative Ansätze: Begründungsversuche der philosophischen Ethik

Den wohl populärsten Ansatz zur Gestaltung von Unfallalgorithmen stellt das Prinzip der Schadensminimierung dar, welches als übergeordnete Zielvorgabe fordert, die Auswirkungen unvermeidbarer Unfälle so gering wie möglich zu halten bzw. im Zweifelsfall jeweils das ›geringere Übel‹ zu wählen. Als Problematik, die in einen praktischen Kontext der Lebenswelt eingebettet ist, müssen Unfallalgorithmen neben ihrer theoretischen Diskussion auch aus praktischer Perspektive vor dem Hintergrund des geltenden Rechtsrahmens bzw. dessen kritischer Betrachtung erörtert werden. In der

einschlägigen Forschungsliteratur werden diesbezüglich verschiedene Entscheidungsstrategien diskutiert, die sowohl ethisch-moralische als auch rechtsphilosophische Überlegungen, meist aus pragmatischer Perspektive, einbeziehen. Das Entscheidungskriterium der Schadensminimierung scheint auf den ersten Blick konsequentialistischen Denkstrukturen zu folgen; jedoch regen einige Autoren eine Debatte darüber an, inwiefern sich ein entsprechendes Design von Algorithmen auch aus der Sichtweise anderer ethischer Theorien normativ verteidigen lässt. Klassische Trolley-Szenarien sind in der philosophischen Tradition üblicherweise mit einer Gegenüberstellung von konsequentialistischen und deontologischen, insbesondere kantianischen, Prinzipien assoziiert worden; diese ethischen Denkschulen bilden auch die Basis des Diskurses moralischer Unfalldilemmata. Gleichzeitig wurden vereinzelt auch alternative ethische Theoriekonzepte vorgeschlagen, die sich an tugendethischen, kontraktualistischen oder rechtsphilosophischen Argumentationen orientieren.[124]

In den folgenden Unterkapiteln wird der umfangreiche normative Diskurs kursorisch rekonstruiert, zunächst ohne Argumente und Positionen im Einzelnen hinsichtlich ihrer Plausibilität zu bewerten. Zu den thematisierten Aspekten zählen die moralische Rechtfertigung spezifischer ethischer Werte bzw. Theorien, ihre Konformität mit geltenden Gesetzgebungen, ihre praktische Implementierbarkeit sowie ihre Vereinbarkeit mit gesellschaftlichen Erwartungen. Während einige der dargestellten Argumente im Hinblick auf eine Ethik für KI- Anwendungen im Allgemeinen gelten, beziehen sich andere auf die praktische Operationalisierbarkeit im spezifischen Kontext eines automatisierten Verkehrs. Ziel ist es deutlich zu machen, dass die bisher im Diskurs vorgeschlagenen ethischen Handlungsprinzipien in nicht-trivialer Weise an ihre Grenzen stoßen.

124 Einen aktuellen, strukturierten Überblick über den einschlägigen Diskurs präsentieren Poszler et al. (2023).

4.4.1 Klassische philosophische Ansätze zur moralischen Relevanz des Intervenierens

Wie bereits gezeigt, ist es fragwürdig, zu weitreichende Analogien zwischen Trolley-Fällen und Dilemma-Szenarien im Kontext des autonomen Fahrens zu ziehen. Dennoch ist zumindest die klassische philosophische Beschäftigung mit dem Trolley-Problem durchaus von Bedeutung für eine Auseinandersetzung mit Unfallalgorithmen, insbesondere in Bezug auf essenzielle deontologische Unterscheidungen.[125] Ausgelöst durch Foots Untersuchung (1978) wird die zeitgenössische Debatte über das Trolley-Problem vielfach mit der Kontrastierung von *doing* und *allowing* in Verbindung gebracht. Besteht ein moralisch relevanter Unterschied zwischen Handlungen, die anderen aktiv Schaden zufügen, und solchen, die entsprechende Schädigungen lediglich zulassen? Im Kern liefern die von Foot und Thomson vorgelegten Gedankengänge unterschiedliche Ansätze zur Begründung der moralischen Intuition, dass (aktives) Töten aus moralischer Sicht schlechter ist als (passives) Sterbenlassen. Foots Ansatz führt dabei über das ethische *Prinzip der Doppelwirkung* (*doctrine of the double effect*).[126] Dessen zentrale These ist es, dass ein moralisch relevanter Unterschied zwischen den Folgen einer Handlung, die beabsichtigt sind, und solchen, die lediglich als Nebenfolgen vorausgesehen werden, besteht.[127] Demnach kann eine Handlung als moralisch gerechtfertigt angesehen werden, wenn ihre negativen Folgen lediglich unbeabsichtigte Nebeneffekte darstellen:

> [...] sometimes it makes a difference to the permissibility of an action involving harm to others that this harm, although foreseen, is not part

125 Woollard (2022, S. 50) definiert deontologische Unterscheidungen wie folgt: »A deontological distinction is a distinction between how agents, victims, and harms are related, which appears to matter morally even though it does not affect the severity or type of harm suffered.«

126 Für eine Übersicht zur Verwendung des *Prinzips der Doppelwirkung* im ethischen Diskurs siehe z. B. Quinn (1989).

127 Foot (1978, S. 20) beschreibt den zugrundeliegenden Gedankengang wie folgt: »The doctrine of the double effect is based on a distinction between what a man foresees as a result of his voluntary action and what, in the strict sense, he intends. He intends in the strictest sense both those things that he aims at as ends and those that he aims at as means to his ends. The latter may be regretted in themselves but nevertheless desired for the sake of the end, as we may intend to keep dangerous lunatics confined for the sake of our safety.«

> of the agent's direct intention. An end such as earning one's living is clearly not such as to justify either the direct or oblique intention of the death of innocent people, but in certain cases one is justified in bringing about knowingly what one could not directly intend. (Ebd., S. 22)

Zu den wichtigsten Implikationen des *Prinzips der Doppelwirkung* zählt der Umstand, dass instrumentalisierende Handlungen als moralisch besonders fragwürdig bewertet werden: »The doctrine of double effect offers us a way out of the difficulty, insisting that it is one thing to steer towards someone foreseeing that you will kill him and another to aim at his death as part of your plan.« (Ebd., S. 23) Für den Fall des klassischen Trolley-Problems führt das Prinzip damit zu keiner Entscheidung, da es beide Optionen gleichermaßen verbietet. Foot legt daher einen alternativen Ansatz vor, der auf die Unterscheidung zwischen positiven und negativen Rechten bzw. Pflichten zurückgreift. Positive Rechte gehen mit positiven Pflichten einher, welche darin bestehen, dass eine bestimmte Handlung ausgeführt werden soll. Dem gegenüber stehen negative Rechte bzw. Pflichten, die ein Unterlassen einer bestimmten Handlung einfordern. Diese haben Vorrang vor positiven Pflichten für den Fall, dass verschiedene Pflichten miteinander in Konflikt geraten. So überwiegt gemäß Foot im klassischen Trolley-Problem das negative Recht des einzelnen Gleisarbeiters, nicht zum Zweck der Schadensminimierung und zugunsten der anderen fünf geopfert zu werden. Das Umleiten der Straßenbahn wäre unzulässig, weil so der Tod des einzelnen Gleisarbeiters nicht nur vorhersehbar, sondern als Teil des verfolgten Handlungsziels beabsichtigt wäre (vgl. ebd., S. 26–29).

Foots Ansatz erscheint intuitiv plausibel, jedoch in seiner konkreten Anwendung, besonders für komplexere Varianten wie das ›Fetter-Mann-Problem‹, beschränkt. Thomson kritisiert vor allem die implizierten Forderungen von Foots Konzeption: Negative Rechte müssten allen Beteiligten zugesprochen werden, sodass eine Hierarchiebildung unmöglich ist und alle Aktionen in Trolley-Dilemmas gleichermaßen verboten seien. Thomson rückt stattdessen die spezifischen, kontextabhängigen moralischen Ansprüche in den Vordergrund, welche Betroffene gegeneinander haben: Besitzt eine Partei einen legitimen höheren Anspruch gegenüber anderen, so sind ihre Interessen zu bevorzugen (vgl. Thomson, 1976, S. 208–211). Entsprechende Anspruchshierarchien können sich auf verschiedene Aspekte

beziehen wie Eigentumsrechte, mögliche Kompensationen für Risiken, Fahrlässigkeit, Versprechen oder besondere Verpflichtungen (vgl. Hübner & White, 2018, S. 693). In einer späteren, im weiteren Verlauf der Debatte kontrovers diskutierten Auseinandersetzung mit dem Trolley-Problem widmet Thomson (2008) sich der Rolle des Fahrers: Hat dieser die Möglichkeit, sich anstelle eines Unbeteiligten selbst zu opfern, so sollte er dies tun.[128]

Welche Relevanz hat all dies nun für den Kontext autonomer Fahrzeuge? Auch wenn in mehr als fünfzig Jahren philosophischer Auseinandersetzung mit dem Trolley-Problem keine eindeutige Entscheidungsstrategie vorgelegt werden konnte, können die entwickelten Gedankengänge für angewandte Probleme in gewisser Hinsicht fruchtbar gemacht werden. Mit der Situation ethischer Entscheidungsträger in Unfalldilemmata lässt sich am ehesten Thomsons *Bystander*-Variante vergleichen. Daraus ergeben sich Implikationen für eines der zentralen Postulate im Hinblick auf die Programmierung von Dilemma-Szenarien: die Unterscheidung zwischen Beteiligten und Unbeteiligten. Die Ethik-Kommission stellt zwar kurz und bündig fest, dass die »an der Erzeugung von Mobilitätsrisiken Beteiligten […] Unbeteiligte nicht opfern« (Di Fabio et al., 2017, S. 11, Regel 9) dürfen, liefert jedoch keine Argumentation, auf die sich die Legitimität der besonderen Schutzwürdigkeit Unbeteiligter zurückführen ließe. Um diese Begründungslücke zu schließen, können die Argumente und Impulse von Foot und Thomson in modifizierter Form hilfreich sein, wenngleich sie auch keine finale Rechtfertigung für eine spezifische Programmierung von Unfallalgorithmen darstellen.

Das Entscheidungsverhalten autonomer Fahrzeuge kennt aufgrund seines algorithmischen Charakters keine standardisierten Trajektorien, sodass nicht in unmittelbarem Sinne zwischen passivem Ansteuern und aktivem Ausweichen differenziert werden kann. Hübner und White (2018, S. 694–695) führen aus, dass die Begrifflichkeiten ›beteiligt‹ und ›unbeteiligt‹ auf zwei verschiedene Lesarten interpretiert werden können. Im Rahmen einer rechtebasierten

128 Thomson hat sich über einen Zeitraum von mehr als dreißig Jahren immer wieder neu mit dem Trolley-Problem auseinandergesetzt. In dieser Forschungsarbeit wird lediglich auf ihre früheren Werke Bezug genommen, um die Darstellung auf die für das Anwendungsproblem wesentlichen Aspekte zu beschränken.

Lesart, wie sie Foot entwirft, erhalten die Begriffe handlungstheoretische Bedeutung. Hier lässt sich mit negativen Pflichten plausibel argumentieren, dass autonomen Fahrzeugen das Ausweichen in Bereiche, die nicht unmittelbar mit der Verkehrssituation in Zusammenhang stehen, untersagt ist, z. B. in benachbarte Fahrspuren, auf Bürgersteige etc. Im Beispielszenario 5 ›Unbeteiligte auf Bürgersteig‹ kann die Fußgängerin auf dem Bürgersteig deshalb als unbeteiligt gelten, weil ihr nichts passieren würde, wenn das Fahrzeug nicht aktiv in ihre Richtung ausweichen würde. Mit Foot hätte sie folglich ein negatives Recht, nicht in die Situation involviert zu werden. Im Gegensatz dazu besitzen unmittelbar Beteiligte des Szenarios lediglich ein positives Recht, vor Schaden bewahrt zu werden. Da jedoch das negative Recht stärker wiegt, darf die Fußgängerin nicht geopfert werden. Auch in anderer Weise lässt sich die Unterscheidung zwischen *doing* und *allowing* auf den Kontext eines automatisierten Verkehrs übertragen: Sollte ein autonomes Fahrzeug, das ohne Passagiere unterwegs ist – beispielsweise wenn es als Taxi zu einem Abholort fährt – eingreifen, wenn es z. B. durch das Blockieren des Fahrtwegs eines anderen Fahrzeugs Schaden von Personen abwenden kann (vgl. Woollard, 2022, S. 58–60)?

Wenn wir im Gegensatz dazu davon ausgehen, dass Personen im Sinne Thomsons über bestimmte Ansprüche verfügen, dann verwenden wir die Begriffe in einer situativ-kontextuellen Bedeutung. Als Unbeteiligte wären all diejenigen Personen zu betrachten, die am Verkehrsgeschehen nicht teilnehmen und daher einen höheren Anspruch auf Sicherheit haben, sich also beispielsweise in verkehrsberuhigten Bereichen wie Straßencafés aufhalten. Gemäß Thomson ist entscheidend, dass sich die Fußgängerin aus dem Beispielszenario 5 ›Unbeteiligte auf Bürgersteig‹ im sicheren Fußgängerbereich aufhält und dadurch bewusst dem (motorisierten) Verkehrsgeschehen entzieht. Sie hat daher z. B. gegenüber jenen einen stärkeren Anspruch, die als Fahrzeugpassagiere in den Genuss der Vorteile des komfortablen autonomen Transports kommen (vgl. Hübner & White, 2018, S. 693–695). Der Fußgängerin dürfen nicht zugunsten der Insassen des autonomen Unfallfahrzeugs Nachteile auferlegt werden.

Ist in Anbetracht dessen eine mögliche Priorisierung des Insassenschutzes grundsätzlich noch zu rechtfertigen? Sowohl für Passagiere als auch für Fußgänger kann plausiblerweise angenommen werden, dass ein gewisser Grad an Eigenhaftung moralisch relevant ist; die-

ser steigt durch individuelles Fehlverhalten, etwa wenn regelwidrig die Straße überquert wird (vgl. Kamm, 2020, S. 94–98). Einen ähnlichen Gedankengang legt Lawlor (2022) zugrunde, der für die Priorisierung von bestimmten geschützten Räumen wie Bürgersteigen eintritt, in denen an Verkehrssituationen unbeteiligte Personen einen gewissen grundsätzlichen Schutz genießen:

> In any case in which driving onto the pavement would impose a risk of harm onto someone who is on the pavement, there should be a very weighty consideration against doing so, and this is the case even if those in the road are not responsible for being there. (Ebd., S. 198)

Für Hevelke und Nida-Rümelin (2015a, S. 222–224) hingegen ist weniger die Verantwortung als die Vorhersehbarkeit des Verhaltens im Straßenverkehr die ausschlaggebende Komponente, mittels derer sich die Priorisierung des Schutzes Unbeteiligter begründen lässt. Sie betonen, dass für ein funktionierendes Verkehrsgeschehen eine »starke[…] prima-facie-Pflicht zu regelkonformen [sic] oder […] zumindest absehbarem Verhalten« (ebd., S. 222) zugrunde gelegt werden muss, aus der sich wechselseitige moralische Ansprüche ableiten. Nicht-regelkonformes Verhalten anderer Personen, wie das regelwidrige Überqueren einer Straße wie in Beispielszenario 3 ›Rote Ampel‹, kann unerwartete Ausweichmanöver verursachen, die die Ansprüche der ›Verursacher‹ tangieren. Problematisch wird es auch dann, wenn regelkonformes Verhalten zu schlechteren Ergebnissen führt als ein Abweichen von dem, was erwartbar ist. Eine Priorisierung des Schutzes Unbeteiligter kann daher nicht als generelles Prinzip gelten, sondern muss im Einzelfall einer Abwägung unterzogen werden (vgl. Hevelke & Nida-Rümelin, 2017, S. 202–204).

4.4.2 Utilitaristische Ansätze

Eine Optimierung des Unfallverhaltens autonomer Fahrzeuge, die der Zielsetzung einer Minimierung resultierender Schäden folgt, legt eine Orientierung an konsequentialistischen Moralprinzipien nahe. Als prominenteste Form konsequentialistischer Ethik geht der Utilitarismus in seiner klassischen, systematisch entwickelten Variante auf Jeremy Bentham (1789; 1970) und John Stuart Mill (1861; 1963–91) zurück. Er bestimmt den ethischen Wert einer Handlung oder Norm allein auf der Grundlage ihrer (vorhersehbaren) Folgen

im Hinblick auf den erwarteten Gesamtnutzen (vgl. Bartneck et al., 2019, S. 26–27). In den Augen seiner Begründer macht das sogenannte *Greatest Happiness Principle* das Kernelement utilitaristischer Theorien aus, welches die Maximierung positiver Zielgrößen wie Glück, Wohlstand oder Nutzen forciert:

> The creed which accepts as the foundation of morals, Utility, or the Greatest Happiness Principle, holds that actions are right in proportion as they tend to promote happiness, wrong as they tend to produce the reverse of happiness. By happiness is intended pleasure, and the absence of pain; by unhappiness, pain, and the privation of pleasure. (Mill, 1861, S. 9–10)

Ausgehend von dieser klassischen Form hat sich der utilitaristische Grundgedanke im Laufe der Zeit ausdifferenziert und in verschiedenen Varianten ausgeprägt. Analog zum Literaturdiskurs finden diese im Folgenden jedoch nur in geringem Maße Berücksichtigung; sofern nicht explizit vermerkt, beziehen sich die beschriebenen Argumente auf den klassischen Utilitarismus.[129] Dieser befürwortet im Rahmen des obigen Handlungsprinzips nach Mill explizit das Opfern von Personen, um eine größere Anzahl zu retten. Bezogen auf das Beispielszenario 2 ›Einzelperson versus Gruppe‹ würde er diejenige Trajektorie bevorzugen, bei der nur eine Person verletzt wird, während die anderen drei verschont blieben. Die Nutzensumme der utilitaristischen Grundformel bemisst sich jedoch nicht nur an der Anzahl betroffener Personen, sondern auch an der Schwere entsprechender Schädigungen. So wäre im Beispielszenario 4 ›Motorradfahrer mit/ohne Helm‹ die Kollision mit demjenigen zu bevorzugen, für den aufgrund seiner Schutzausrüstung geringere Verletzungen zu erwarten sind. Konsequent zu Ende gedacht fordert eine utilitaristische Ausrichtung von Unfallalgorithmen auch die Selbstopferung Einzelner für das Gesamtwohl. Dies würde im Beispielszenario 7 ›Klippe‹ das Ausweichen des Fahrzeugs implizieren, wobei sich die Insassen zugunsten der Schulkinder opfern.

Gewissermaßen unklar ist, wie aus utilitaristischer Sicht entschieden werden soll, wenn Anzahl und Schadensausmaß der betroffenen Personen bei allen Handlungsoptionen entweder gleichermaßen

129 Mit konsequentialistischen Perspektiven auf ethische Fragen autonomer Fahrsysteme hat sich die Autorin bereits an anderer Stelle auseinandergesetzt (vgl. Schäffner 2018, 2020a).

oder kategorial unterschiedlich schlecht sind. In diesen Fällen gibt der Utilitarismus keine klare Präferenz vor. Haben die Optionen völlig gleiche Konsequenzen, wären beide erlaubt; eine Handlungsentscheidung kann beispielsweise durch einen Münzwurf erfolgen. Problematischer gestaltet es sich, wenn völlig unterschiedliche, miteinander unvergleichbare Optionen vorliegen. Kommensurabilität ist eine zwingende Voraussetzung für jegliche utilitaristische Erwägung. Nutzenwerte können zwar auch bei kategorial unterschiedlich schlechten Konsequenzen zugewiesen werden, jedoch müssten diese ethisch erst einmal begründet werden: Wiegt der erwartete Tod einer Person stärker oder die schweren Verletzungen einer ganzen Gruppe?

Ein utilitaristischer Ansatz setzt prinzipiell eine Quantifizierung des erwarteten Schadens im Sinne negativen Nutzens voraus, auf deren Basis dann diejenige Handlungsoption gewählt wird, die den geringsten Gesamtschaden verursacht bzw. das allgemeine Wohl maximiert. Den gängigsten Ansatz zur Integration des Aspekts unsicherer Handlungsfolgen in utilitaristische Frameworks stellt die entscheidungstheoretische Standardmethode der Risikoanalyse für die Entscheidungsfindung unter Risiko dar, welche auf der Maximierung des Erwartungsnutzens beruht (siehe Kap. 6.3). Als Optimierungsproblem mit relativ schematischer Grundformel ließe sich dies generell mittels mathematischer Kostenfunktionen in Algorithmen implementieren (vgl. Thornton et al., 2017, S. 1431–1437).

Doch obwohl der utilitaristische Ansatz zu den meistdiskutierten ethischen Prinzipien im Kontext von Unfallalgorithmen gehört, befürworten ihn nur wenige. Utilitaristische Positionen werden insbesondere im Kontext eines Verrechnungsverbots menschlicher Leben und sich daraus ergebender, moralisch fragwürdiger Ergebnisse kontrovers thematisiert. Der Kritik liegt dabei eine pragmatische Perspektive auf die Problemstellung zugrunde. Diese betrachtet mögliche Entscheidungsstrategien nicht nur theoretisch, sondern stets mit Blick auf situative Gegebenheiten und zur Verfügung stehende praktische Handlungsmöglichkeiten. Dies schließt vor allem den geltenden Rechtsrahmen sowie gesellschaftlich akzeptierte Grundwerte mit ein. Einschlägige Argumentationen stützen sich deshalb sowohl auf moralische als auch rechtsethische Aspekte und nehmen Bezug auf zentrale Kritikpunkte an einem ethischen Konsequentialismus (vgl. Nida-Rümelin et al., 2012, S. 130–133): Erstens berücksichtigen

utilitaristische Ansätze durch ihre Fokussierung auf eine intersubjektive, aggregierte Nutzensumme weder Interessen noch Motive der Einzelnen und lassen so Verteilungs- und Gerechtigkeitsaspekte unbeachtet. Da sich aus utilitaristischer Sicht jeder noch so große Nutzenverlust durch hinreichend viele kleine Nutzengewinne moralisch kompensieren lässt, ignoriert er insbesondere die im Rahmen von Rawls' Gerechtigkeitstheorie begründete Separatheit von Personen (*separateness of persons*):[130]

> This [utilitarian] view of social cooperation is the consequence of extending to society the principle of choice for one man, and then, to make this extension work, conflating all persons into one through the imaginative acts of the impartial sympathetic spectator. Utilitarianism does not take seriously the distinction between persons. (Rawls, 1971, S. 27)

Gerechtigkeitsnormen lassen sich nicht konsequentialistisch begründen; normativ relevante Wertefunktionen in konsequentialistischen Theorien sind stets außermoralisch zu bestimmen. Der aggregative Charakter der Erwartungsnutzensumme, die den Interessen der Einzelnen kein spezifisches Gewicht beimisst, verschärft die Missachtung von Verteilungsfragen gegenüber dem Standardutilitarismus, indem durch die Wahrscheinlichkeitskomponente eine komplexere Verteilung von Nachteilen zwischen Personen ermöglicht wird.[131]

Zweitens werden im Zuge der Abwägung von Einzelinteressen zum Zweck der Gesamtnutzenoptimierung die individuellen Rechte

---

130 An dieser Stelle sei angemerkt, dass die nachfolgende Analyse die in der Literatur vorherrschende kritische Einstellung gegenüber einer ausschließlich utilitaristischen Orientierung von Unfallalgorithmen widerspiegelt. Über eine theoretische Eignung utilitaristischer Entwürfe ist damit noch nichts gesagt. Es sei ausdrücklich auf den Variantenreichtum utilitaristischer Ansätze sowie deren kontroverse Beurteilung verwiesen; dabei wird eingeräumt, dass spezifische utilitaristische Entwürfe beispielsweise im Hinblick auf Verteilungsaspekte und individuelle Interessen aus theoretisch-analytischer Sicht weniger problematisch sind als andere. Jedoch erscheint die Grundausrichtung des utilitaristischen Kerns als (alleinige) Grundlage für eine pragmatisch orientierte Programmierung von Unfallalgorithmen zumindest fragwürdig.

131 Als Alternative zur Maximierung des Erwartungsnutzens erwähnt Hansson (2013, S. 24–26) noch den Ansatz des *Actual Consequence Utilitarianism* (»The utility of a mixture of potential outcomes is equal to the utility of the outcome that actually materializes.«), attestiert ihm aber ebenfalls eine mangelnde Eignung für praktische Fragen.

und die Autonomie derjenigen missachtet, die potenziell für das Allgemeinwohl geopfert würden. »In conglomerating the sufferings and enjoyments of all people, utilitarianism fails to recognize the importance of individual identity«, konstatiert Grau (2006, S. 54). Die Instrumentalisierung von Individuen zugunsten des Gesamtwohls ist ethisch unzulässig. So bemängelt u. a. Hansson (2013, S. 26–28) aus normativer Sicht, dass eine Orientierung am Erwartungsnutzen keinen Raum für risikoaverse Einstellungen lässt. Wahrscheinlichkeiten und moralische Dimension der Auswirkungen einer Handlung verhalten sich nicht unbedingt proportional zueinander – beispielsweise dann nicht, wenn Wirkungen katastrophalen Ausmaßes mit einer sehr geringen Wahrscheinlichkeit zu erwarten sind. Drittens kommt es aufgrund einer konsequentialistischen Nutzenorientierung zur Missachtung individueller Integrität, wenn Personen ihre eigenen Projekte um anderer willen aufgeben müssen.

Im Zusammenhang mit einer ethischen Konsequentialismuskritik werden grundsätzliche moralische Fragen nach dem Wert menschlichen Lebens aufgeworfen: Wie ließe sich der Wert eines Menschenlebens überhaupt objektiv bestimmen – und sind fünf Leben automatisch mehr wert als eines (vgl. Hevelke & Nida-Rümelin, 2017, S. 197–198; Santoni de Sio, 2017, S. 418)?[132] Der utilitaristische Ansatz bietet Anreize, Personen beispielsweise nach ihrem Nutzen (oder ihrer Belastung) für die Gesellschaft zu bewerten. Eine jüngere Person würde höher bewertet als eine ältere, ein prominenter Forscher höher als ein LKW-Fahrer (vgl. Liu, 2016, S. 168). In diesem Sinne wären unter utilitaristischen Gesichtspunkten auch Kollisionen mit schwereren bzw. leichteren Fahrzeugen im Hinblick auf den Gesamtschaden zu bevorzugen (vgl. Goodall, 2014b, S. 97; Lin, 2014a).[133]

132 In den Sozialwissenschaften ist beispielsweise der Indikator *disability-adjusted life years (DALY)* populär, der Schaden als die Anzahl der verlorenen Jahre eines gesunden Lebens beziffert (vgl. Murray, 1994).

133 Ein ähnliches Argument verwendet Bennett (2022, S. 198–203), um zu begründen, dass es aus utilitaristischer Sicht vertretbar wäre, wenn Unfallalgorithmen stets die Sicherheit der Fahrzeuginsassen priorisierten. Dabei zieht er statistische Daten heran, die zeigen, dass Insassen bei Unfällen häufiger tödlich verletzt werden als andere Gruppen von Verkehrsteilnehmern, welche hingegen eine höhere Zahl an Krankenhausaufenthalten infolge von Verkehrsunfällen aufweisen. Im Sinne eines utilitaristischen Kalküls ließe sich das Gesamtwohl erhöhen, wenn die Risiken für tödliche Unfälle reduziert, also die Fahrzeug-

Damit gehen verschiedene praktische Probleme einher. Unter Bezugnahme auf Harris' *Survival Lottery* (1975)[134] und Singers (1977) Antwort darauf erläutert Bennett (2022, S. 195–198), dass streng utilitaristische Algorithmen Anreize für egoistisches Verhalten schaffen: Individuen könnten sich die angestrebte Zielgröße der Maximierung des Gesamtwohls in manipulativer Weise zunutze machen, indem sie sich z. B. bewusst rücksichtslos verhalten und damit die Opferung besser geschützter Personengruppen, etwa Fahrzeuginsassen, forcieren.

Dennoch bedeutet jegliche Form der Aufrechnung eine Verletzung der moralischen Pflicht zur Achtung der individuellen Würde und ist damit auf grundrechtlicher Ebene mindestens fragwürdig. Aufgrund seiner Orientierung am Gesamterwartungsnutzen ist der utilitaristische Ansatz prinzipiell indifferent gegenüber moralisch relevanten Unterscheidungen zwischen Handlungen und deren Hintergründen. Wie in Kap. 4.4.1 dargestellt, ist die Differenzierung zwischen aktivem und passivem Schädigen einerseits sowie zwischen Beteiligten und Unbeteiligten andererseits von großer Bedeutung für den Anwendungskontext. Aus utilitaristischer Sicht können diese Aspekte nicht berücksichtigt werden. Problematisch ist das insbesondere aufgrund der Tatsache, dass autonome Fahrsysteme Indivi-

insassen besonders geschützt würden. Ein weiterer Effekt, der das Potenzial besitzt, die kollektive Wohlfahrt weiter zu steigern, besteht darin, dass andere Verkehrsteilnehmer bei derart programmierten Algorithmen über keinerlei Anreize für rücksichtsloses Verhalten verfügen; mehr Vorsicht und Rücksichtnahme würden die Unfallzahlen potenziell weiter senken. Nicht zuletzt würden bei einer Priorisierung des Insassenschutzes auch Kaufanreize geschaffen, die wiederum geeignet sind, die positiven Effekte der automatisierten Mobilität als solche zu befördern.

134 Die *Survival Lottery* ist ein von John Harris konzipiertes Gedankenexperiment. Es basiert auf der Idee, dass durch Organspenden mehr Leben zu retten sind als durch den Tod der Spender verloren gehen. Sobald mindestens zwei Mitglieder einer Gesellschaft ein Spenderorgan benötigen, wird per Los ein anderes, gesundes Individuum gezogen, von dem erwartet wird, sein Leben zugunsten der Kranken zu opfern. Die Argumentationsgrundlage dieses Experiments ist im Kern utilitaristisch, da sie stets das Wohlergehen der größten Zahl priorisiert. Harris' Experiment wurde vielfach (kritisch) rezipiert und auf verschiedene Anwendungsfragen übertragen, vor allem solche, die im Kontext von *Killing-versus-Letting-Die*-Problematiken stehen.

duen hohe Anreize für Missbrauch[135] und opportunistisches Verhalten bieten. Als direkte Folge unzureichender Regulierung entsteht Raum für bösartig manipulatives Verhalten immer da, wo ein bestimmter Umgang mit Technologie impliziert ist, wo Individuen sich an Technologien anpassen müssen und nicht umgekehrt:[136]

> […] whenever we expect human behavior to change to adapt to AVs, instead of vice-versa, we raise the possibility that unscrupulous, abusive humans will find new ways to change their behavior, with a resulting arms race. The Prisoner's Dilemmas raised by new technology will reward those who ingeniously defect to cause harm. (Abney, 2022, S. 261–262)

Als Beispiel für eine Situation, in der manipulatives Verhalten zu moralisch fragwürdigen Ergebnissen führt, gilt das sogenannte »Chicken Problem« (vgl. Abney, 2022, S. 259–260): Wenn autonome Fahrzeuge generell so programmiert würden, dass sie tendenziell einer Kollision ausweichen, um hohen Schaden zu vermeiden, wäre es anderen Verkehrsteilnehmern möglich, absichtlich eine Kollision zu provozieren, um das Fahrzeug zum Ausweichen zu zwingen. Neben der Vorhersehbarkeit des utilitaristischen Ansatzes an sich wird dabei zusätzlich die (generell gewünschte) Transparenz algorithmischer Entscheidungssysteme ausgenutzt: »[…] predictability opens the possibility to manipulation.« (Osório & Pinto, 2019, S. 41) Derartige Situationen sind weder unrealistisch noch einfach zu regulieren, ohne Nachteile hinsichtlich Kontrolle und Vorhersehbarkeit hinnehmen zu müssen:[137]

---

135 Abney (2022, S. 258) definiert Missbrauch in diesem Kontext folgendermaßen: »[…] to count as abuse, the use case must have a purpose that directly attacks or undermines the primary purpose, which […] for AVs I define as transporting people from one location to another by road in a safe, reliable, comfortable, and timely manner.«

136 Dies gilt nicht ausschließlich für utilitaristisch programmierte Fahrzeuge, ist aber in deren Kontext am offensichtlichsten.

137 Um die Anreize für manipulatives Verhalten zu verringern, müsste eine gewisse Unsicherheit hinsichtlich des Entscheidungsprozesses bestehen. Osório und Pinto (2019, S. 43) erklären, dass dies u. a. zulasten der Qualität des Entscheidungsprozesses geht, und stellen Ansätze interner und externer Unsicherheit gegenüber: »[…] in order to remove the incentives to manipulate and to solve the malicious pedestrian problem, individuals with bad intentions must hold some uncertainty about the decision and evaluation processes of the autonomous vehicle system. Noise or observation difficulties reduce the incentives to

> […] if we want to solve the chicken problem, it seems we need to forego perfect predictability and control of how an AV will react to an abusive human—we may need some amount of unpredictability, a lack of knowledge of what the AV will choose; that is, we need a certain lack of control. (Abney, 2022, S. 264)

Auch aus empirischer Sicht spricht wenig für eine rein utilitaristisch basierte Ausrichtung von Unfallalgorithmen. Zwar legen die Ergebnisse einschlägiger empirischer Studien nahe, dass die moralischen Präferenzen potenzieller Nutzer autonomer Fahrzeuge tendenziell utilitaristische Züge aufweisen bzw. dass Entscheidungsfaktoren als moralisch relevant erachtet werden, welche der Maximierung eines quantitativen Gemeinwohls entsprechen (vgl. Bergmann et al., 2018, S. 11; Faulhaber et al., 2019, S. 407–413). Allerdings ist umstritten, inwiefern diese ermittelten Resultate den spezifischen Umständen der Experimentumgebung geschuldet sind. So vertritt Kauppinen (2021, S. 632–633) die Auffassung, dass die vernunftgemäße Moral (*commonsense morality*) keineswegs utilitaristisch ist, insbesondere dann nicht, wenn Beeinträchtigungen individueller Rechte und Autonomie drohen. Anhand von Simulationsanalysen argumentieren Samuel et al. (2020, S. 3–5), dass in virtuellen Szenarien zwar häufig utilitaristisch entschieden wird, in realen Dilemma-Situationen mit zeitlichen Restriktionen hingegen kaum. Wie Lacroix (2018) und Edmonds (2018) darlegen, offenbart bereits das klassische Trolley-Problem, dass die meisten Personen keine strengen Utilitaristen sind. Vielmehr verfügen wir über ›kantianische Instinkte‹ – wir lehnen es intuitiv ab, Menschen für höhere Ziele zu instrumentalisieren und sind bereit, ab und an eine deontologische Regel zu akzeptieren, sofern diese das Gemeinwohl fördert.[138] Eine andere Perspektive auf die moralische Problematik utilitaristischer Folgenorientierung präsentieren Bodenschatz et al. (2021). Anhand dreier empirischer Studien demonstrieren sie, dass Studienteilnehmer es in Bezug auf Dilemmata durchaus als moralisch valide betrachten, über Handlungsoptionen zu randomisieren. Diese Präferenzen bestehen

misbehave. However, uncertainty may also reduce the quality of the decision process.«

138 Bruers und Braeckman (2014, S. 251–252) betonen, dass in radikaler Form weder utilitaristische noch deontologische Moralprinzipien unserer moralischen Intuition entsprechen.

vor allem dann, wenn keine eindeutige utilitaristische Alternative offensichtlich ist.[139]

Neben genuin moralischen Gesichtspunkten ergeben sich weitere Schwierigkeiten bei der konkreten Operationalisierung des utilitaristischen Moralprinzips. Hierbei sind zunächst Herausforderungen im Umfeld der Bestimmung von Nutzenwerten relevant. Es ist unklar, wie ein entstehender Personenschaden überhaupt quantifiziert werden soll, sowohl in kurz- als auch langfristiger Perspektive (vgl. Goodall, 2014b, S. 99). Metriken zum interpersonellen Nutzenvergleich sind zum gegenwärtigen Stand der Technik ebenso wenig verfügbar wie Verfahren zur Approximation von Wahrscheinlichkeiten in solcher Genauigkeit, wie sie das mathematische Optimierungsproblem erfordert.[140] Darüber hinaus bestehen Hürden, was eine mögliche Implementierung angeht. Selbst wenn Nutzenwerte und Wahrscheinlichkeiten hinreichend präzise bestimmt werden könnten, würde ein Algorithmus, der die Konsequenzen einer Handlung in Zeit und Raum umfassend einbezieht, sehr viel Rechenkapazität und -zeit benötigen, um alle relevanten Informationen zu verarbeiten.[141] Es wäre notwendig, dem System hinsichtlich der einzubeziehenden

139 An dieser Stelle sei darauf hingewiesen, dass Zufallsentscheidungen hinsichtlich ihrer ethischen Akzeptabilität stark umstritten sind, siehe z. B. Broome (1984) oder Misselhorn (2018b, S. 196–198).

140 Diese Probleme sind in der ethischen Tradition wohlbekannt. Jedoch sind bisher keine Lösungen entwickelt worden, die im Kontext von Algorithmen operationalisierbar wären. So weisen Geisslinger et al. (2021, S. 1045–1046) darauf hin, dass sich aus technischer Sicht die Schwere eines Unfalls grundsätzlich nur zu einem gewissen Grade präzise voraussagen lässt. Mit hinreichender Sicherheit bestimmt werden können nur drei Charakteristika: die Masse des potenziellen Kollisionsobjekts (durch Identifizierung des Typs des Verkehrsteilnehmers), die Differenzgeschwindigkeit der Unfallbeteiligten und der Aufprallwinkel. Aus diesen lässt sich die kinetische Energie berechnen, die bei einer Kollision freigesetzt wird. Die Schwere einer physischen Verletzung steigt dabei proportional zur aufgewendeten kinetischen Energie.

141 Dies gilt analog auch für die Implementierung deontologischer Systeme. Um die technischen Schwierigkeiten zu überwinden, schlägt Klincewicz (2017, S. 252–254) die Verwendung einer hybriden Systemarchitektur vor, welche auf dem Zusammenspiel deduktiver und induktiver Komponenten beruht. Symbolische Algorithmen, die spezifische ethische Prinzipien implementieren, generieren dabei eine Datenbasis paradigmatischer, moralisch eindeutiger Standardfälle. Diese werden sodann von einer separaten Systemkomponente bearbeitet und durch analoges Schlussfolgern auf Spezialfälle wie Dilemmata angewandt.

Parameter Grenzen zu setzen; doch auf welcher (moralischen) Basis sollen diese festgelegt werden (vgl. Allen et al., 2005, S. 151)? Ebenso stellt die Bestimmung einer konkreten utilitaristischen Nutzenfunktion eine anspruchsvolle Herausforderung dar:

> […] it is notoriously difficult to calculate a utility function for everyone involved and promote the outcome with the highest net utility, as classical utilitarianism would have it. There are just too many factors to take into account for a regular autonomous car in such a situation: how many persons are how likely to be how seriously injured with how much potential quality of life left, thus creating how much grief in how many relatives, just to name a few factors. (Loh & Loh, 2017, S. 44)

4.4.3 Deontologische Ansätze

Moralische Pflichten und grundlegende ethische Prinzipien wie die Achtung der Menschenwürde oder Gerechtigkeit stehen im Zentrum deontologischer Moralentwürfe. Im Gegensatz zu konsequentialistischen Theorien richten sie den Blick nicht nur auf die Folgen des Handelns, sondern auch auf dessen Merkmale und Voraussetzungen. Das kann die Absicht sein, mit der eine Handlung ausgeführt wird, oder die Kompatibilität mit einem formalen Prinzip bzw. einer Handlungsregel (vgl. Bartneck et al., 2019, S. 25). Letztere sind geeignet, bestimmte Typen von Handlungen als grundsätzlich moralisch unzulässig auszuweisen und auf diese Weise unverhandelbare Grenzen festzulegen. Zentraler Bestandteil einer deontologischen Ethik in der Tradition Kants ist der kategorische Imperativ als oberste Maxime – als Regel, die unser Handeln und Wollen bestimmt: »Handle nur nach derjenigen Maxime, durch die du zugleich wollen kannst, daß sie ein allgemeines Gesetz werde.« (Kant, 1900ff., GMS, AA 04: 421.07-08) Wie im Folgenden gezeigt wird, sind deontologische Ansätze für die Programmierung von Unfallalgorithmen ebenfalls nur beschränkt hilfreich. Zwar berücksichtigen sie die unter der utilitaristischen Perspektive vernachlässigten Postulate normativer Gleichheit und individueller Würde, stoßen aber hinsichtlich ihrer praktischen Operationalisierung sowohl an strukturelle als auch technisch-formale Grenzen.

Im relevanten Forschungsdiskurs finden deontologische Elemente in vielfältiger Weise Beachtung. Es wird beispielsweise ergründet, inwiefern spezielle moralische Pflichten gegenüber besonders schutz-

würdigen Gruppen, z. B. Kindern, Vorrang vor anderen ethischen Erwägungen haben. Da solche Ansätze jedoch eine qualifizierende Beurteilung der Betroffenen vornehmen, sind sie in direkter Form für die Praxis untauglich. Unter Berufung auf eine kantianisch geprägte ethische Tradition verbietet die Ethik-Kommission in ihren ethischen Leitlinien ausdrücklich, Menschenleben anhand persönlicher Merkmale gegeneinander aufzuwiegen (vgl. Di Fabio et al., 2017, S. 11). Personen zugunsten anderer zu opfern degradiert diese zu bloßen Objekten, mittels derer ein höheres Ziel erreicht werden soll. Handlungen, die Einzelne für einen derartigen Zweck instrumentalisieren, stellen einen Verstoß gegen die Selbstzweckhaftigkeit eines jeden Individuums dar, die in der Fähigkeit zum autonomen Handeln und dem Setzen eigener Ziele gegründet ist:

> Nun sage ich: der Mensch und überhaupt jedes vernünftige Wesen existirt als Zweck an sich selbst, nicht bloß als Mittel zum beliebigen Gebrauche für diesen oder jenen Willen, sondern muß in allen seinen sowohl auf sich selbst, als auch auf andere vernünftige Wesen gerichteten Handlungen jederzeit zugleich als Zweck betrachtet werden. (Kant, 1900ff., GMS, AA 04: 428.07-11)

Gemäß der prominenten Selbstzweckformel von Immanuel Kant besteht jedoch nicht nur eine Pflicht gegenüber anderen, sondern auch gegen sich selbst, die es verbietet, sich bzw. die eigene Selbstzerstörung für ein höheres Ziel als Mittel zu gebrauchen (vgl. ebd., S. 428–429). Auch Kauppinen (2021) erläutert, dass die aus der menschlichen Würde abgeleiteten Individualrechte eine gewissermaßen unverfügbare Grenze markieren, die eine Opferung Unschuldiger zugunsten anderer grundsätzlich untersagt:

> […] as long as people have rights, there are also possible situations in which it is not permissible to minimize harm. In the cases that are pertinent here, this is because someone has an intact right not to be harmed, and the only way to avoid violating it causes (or risks) greater or equal harm to someone who has lost their right not to be harmed that way. For example, if three robbers are trying to kill one innocent person to steal her wallet, it is morally permissible to kill all of them if necessary to save the one (even if they would afterwards become upright citizens), because they have forfeited their right not to be harmed, while the innocent person hasn't. (Ebd., S. 633)

Auf den ersten Blick erscheint eine Programmierung auf Schadensminimierung somit grundsätzlich unvereinbar mit deontologischen

Grundprinzipien. Jedoch wird diese Position im Kontext von Unfallalgorithmen durchaus kontrovers bewertet. Eines der häufigsten Argumente, die für die generelle Zulässigkeit einer Opferung Unbeteiligter auch aus deontologischer Perspektive sprechen, besagt, dass die Identität potenzieller Opfer zum Zeitpunkt der Programmierung, i. e. der vorgelagerten Entscheidungsfindung, noch nicht feststeht und deren Rechte deshalb nicht beeinträchtigt seien. Vor diesem Hintergrund kann die Empfehlung der Ethik-Kommission verstanden werden, die im Rahmen einer vorläufigen Bewertung eine Schadensminimierung für solche Fälle zulässig erklärt, in denen es darum geht, eine möglichst große Zahl an Unbeteiligten zu retten. Als Voraussetzung soll hier gelten, dass alle potenziell Betroffenen von dem implementierten Algorithmus in der Form profitieren, dass dieser das Risiko für alle in gleichem Maße reduziert:

> In der Konstellation einer vorweg programmierbaren Schadensminderung innerhalb der Klasse von Personenschäden liegt der Fall anders als der des Luftsicherheitsgesetzes oder der Weichensteller-Fälle. Hier ist nämlich eine Wahrscheinlichkeitsprognose aus der Situation zu treffen, bei der die Identität der Verletzten oder Getöteten (im Gegensatz zu den Trolley-Fällen) noch nicht feststeht. Eine Programmierung auf die Minimierung der Opfer (Sachschäden vor Personenschäden, Verletzung von Personen vor Tötung, geringstmögliche Zahl von Verletzten oder Getöteten) könnte insoweit jedenfalls ohne Verstoß gegen Art. 1 Abs. 1 GG gerechtfertigt werden, wenn die Programmierung das Risiko eines jeden einzelnen Verkehrsteilnehmers in gleichem Maße reduziert. Solange nämlich die vorherige Programmierung für alle die Risiken in gleicher Weise minimiert, war sie auch im Interesse der Geopferten, bevor sie situativ als solche identifizierbar waren. (Di Fabio et al., 2017, S. 18)

In eine ähnliche Richtung zielen auch Hevelke und Nida-Rümelin (2015c, S. 11–12): Wenn man davon ausgeht, dass eine Programmierung auf Schadensminimierung im Interesse jedes Einzelnen liegt, wird der scheinbare Widerspruch zum kantianischen Instrumentalisierungsverbot aufgelöst, denn wenn alle von der Regelung profitieren, ist jeder zugleich Zweck, nicht nur Mittel. Dies wird insbesondere deutlich, wenn man das autonome Fahren als Technologie betrachtet, die die Sicherheit jedes Einzelnen erhöht, und eine Programmierung, die die Tötung Unschuldiger beinhaltet, ihrerseits individuelle Freiheit und Selbstbestimmung befördert (vgl. Misselhorn, 2018b, S. 192).

Die Grundannahmen derartiger Argumente stoßen innerhalb des Forschungsdiskurses jedoch auf Kritik. So kann angesichts einer allgemeinen Skepsis gegenüber den allzu optimistischen Erwartungen an autonome Fahrzeuge zunächst bezweifelt werden, inwiefern diese tatsächlich förderlich für eine selbstbestimmte Lebensführung sind. Schränken sie nicht vielmehr die individuelle Freiheit ein, beispielsweise durch die Notwendigkeit, personenbezogene Daten für die vernetzte Kommunikationsinfrastruktur preisgeben zu müssen (vgl. ebd., S. 199–200)? Weiterhin ist die Annahme, dass die Unbestimmtheit der Identität potenzieller Opfer eine moralische Relevanz besitzt, zumindest fragwürdig:

> Moralisch und rechtlich gesehen bedeutet es nach den bisher akzeptierten Standards für die Unrechtsbewertung einer Tat […] keinen wesentlichen Unterschied, ob man das Opfer schon persönlich identifiziert hat oder ob die Identität des Opfers vom Zufall abhängt bzw. von Umständen, die z. Z. der Tötungshandlung noch nicht bekannt waren. Wer eine Drohne losschickt, die den nächstbesten Menschen tötet, handelt genauso unmoralisch und rechtswidrig wie derjenige, der die Drohne auf eine bestimmte, ihm bekannte Person zum Zweck ihrer Tötung hinsteuert. Auch im Dilemma-Fall ist bereits bei der Programmierung bekannt, dass ein Mensch geopfert werden wird; nur seine Identität steht noch nicht fest. Das Unrecht der Tat liegt in der Opferung des Menschen (als solchem), auf irgendwelche Identitätsmerkmale kommt es nicht an. (Hilgendorf, 2018a, S. 693)

Auch die Annahme, eine die Opferung Unschuldiger implizierende Programmierung sei im Interesse des Einzelnen, ist diskussionswürdig. Ein solches Interesse kann nur im Rahmen von Entwürfen angenommen werden, die auf einer unparteilichen Ausgangslage basieren, wie sie beispielsweise in John Rawls' (1971, S. 136–142) berühmtem *Schleier des Nichtwissens* (*veil of ignorance*) besteht.[142] Dieser ist zentraler Bestandteil der Beschreibung eines spezifischen Zustands der Menschen in einer fiktiven Entscheidungssituation; angewandt auf Unfalldilemmata würde er implizieren, dass nur dann von einem gleichen Interesse aller gesprochen werden kann, wenn alle Individuen gleichermaßen fürchten müssen, in Notsituationen geopfert zu werden. Tatsächlich erscheint es im gegebenen Kontext jedoch unplausibel davon auszugehen, dass entsprechende

142 Dies wird in Kap. 4.4.4.2 näher ausgeführt.

Risiken für alle gleichermaßen minimiert werden können. Relevante Verkehrsrisiken sind naturgemäß unterschiedlich; sie sind u. a. abhängig von der Fortbewegungsform bzw. der Häufigkeit ihrer Nutzung oder auch von der individuellen Bereitschaft bzw. vorhandenen Anreizen zu unvorsichtigem Verhalten. So können Fußgänger generell als gefährdeter gelten als durch Karosserie geschützte Personen in Fahrzeugen.[143] Jedoch gilt auch: Wer häufiger zu Fuß geht, dessen Risiko steigt gegenüber denjenigen, die meistens das Auto nehmen. Und wer zudem dazu neigt, ab und zu eine Regel zu missachten, dessen Risiko steigt weiter. Eine prinzipielle, systematische Opferung der Interessen bestimmter Personen zugunsten anderer setzt ferner Fehlanreize für unsoziales Verhalten (vgl. Hevelke & Nida-Rümelin, 2017, S. 200–201). Schließlich kann es auch andere, z. B. altruistische, Interessen als das egoistische Überlebensinteresse geben, die hier jedoch nicht berücksichtigt werden.

Jenseits dieser komplexen Schwierigkeiten bei der inhaltlichen Gestaltung eines Handelns aus Pflicht erweisen sich auch technisch-formale Aspekte deontologischer Ansätze als Herausforderung bei möglichen Implementierungen. Ein häufiger Kritikpunkt an deontologischen Entwürfen wie dem kantianischen lautet, dass die vollständige Konsistenz von Normensystemen ein unrealistisches Ziel darstellt. Aufgrund ihrer kategorischen Natur sind formale Prinzipien häufig zu unspezifisch, mehrdeutig und interpretationsbedürftig, als dass sie in komplexen lebensweltlichen Entscheidungssituationen Orientierung geben könnten (vgl. Misselhorn, 2019, S. 50). Dies ist im Kontext algorithmischer Entscheidungen, wo die menschliche Interpretationsfähigkeit sowie die Fähigkeit zu kontextsensitivem Handeln durch künstliche Akteure ersetzt werden, besonders problematisch.[144]

In der wissenschaftlichen Literatur existieren einige wenige konkrete Ansätze zur Implementierung moralischen Entscheidungsverhaltens für den Kontext autonomer Fahrzeuge. Der häufigste Implementierungsansatz besteht darin, deontologische Entwürfe für eine algorithmische Entscheidungsfindung als hierarchisch organisierte Regelsysteme zu konzipieren. Ein prominenter Bezugspunkt sind hier die von Isaac Asimov (1942) entwickelten Robotergesetze, die

143 Dieser Gedankengang wird in Kap. 7.3.3.2 präzisiert.

144 Siehe hierzu auch den Beitrag von Prakken (2017), der Herausforderungen hinsichtlich der Programmierung verkehrskonformen Verhaltens anhand einer Fallstudie zum niederländischen Verkehrsrecht erörtert.

in zahlreichen Anwendungskontexten adaptiert wurden, so auch für das autonome Fahren (vgl. Misselhorn, 2018b, S. 189–190). Die Gesetze lauten im Einzelnen:

> 1. A robot may not harm a human being, or, through inaction, allow a human being to come to harm. 2. A robot must obey the orders given to it by human beings except where such orders would conflict with the First Law. 3. A robot must protect its own existence, as long as such protection does not conflict with the First or Second Law. (Grau, 2006, S. 53)

Inwiefern derart hierarchische Gesetzessysteme für eine praktische Anwendung brauchbar sind, bleibt allerdings fragwürdig (vgl. Trappl, 2015, S. 6).

Konkretere Ansätze reichen über die Übersetzung ethischer Konstrukte in mathematische Äquivalente aus der Kontrolltheorie hin zur Bestimmung optimalen Verhaltens mittels Kosten- oder Wohlfahrtsfunktionen (vgl. Gerdes & Thornton, 2015; Kinjo & Ebina, 2017; Thornton et al., 2017) und die Identifizierung besonders schutzwürdiger Gruppen über externe Hardwarekomponenten (vgl. Liu, 2018, S. 162–168). Für den (allgemeineren) Kontext künstlicher Systeme sind zudem verschiedene Ansätze auf deontologischer Basis entwickelt worden, die eher auf die Programmierung von Algorithmen zugeschnitten sind als die traditionelle Deontologie, z. B. indem sie Abstufungen von Pflichten hinsichtlich ihrer normativen Verbindlichkeit vornehmen. So legt Powers (2006) den Entwurf einer ›kantianischen Maschine‹ vor, die normative Ansprüche gemäß der Kategorien ›verboten – erlaubt – geboten' clustert. Zu erwähnen sind in diesem Zusammenhang auch die auf William David Ross (1930) zurückgehenden *Prima-Facie*-Pflichten[145], die zwar gültig sind, aber in bestimmten Fällen durch andere überschrieben werden können. Mithilfe von Rawls' Konstrukt des *Überlegungsgleichgewichts* (*reflective equilibrium*) sind diese prinzipiell in ethische Algorithmen implementierbar (vgl. Anderson et al., 2005, S. 2–4). So sind sie auf der Grundlage der Prinzipienethik von Beauchamp

145 Ross schlägt insgesamt sieben *Prima-Facie*-Pflichten vor: Treue (*fidelity*), Wiedergutmachung (*reparation*), Dankbarkeit (*gratitude*), Gerechtigkeit (*justice*), Fürsorge (*beneficence*), Nichtschädigung (*nonmaleficence*), Selbstvervollkommnung (*self-improvement*).

und Childress (1994)[146] bereits für den Kontext medizinethischer Anwendungen in Form des Prototypen *MedEthEx* realisiert worden (vgl. Anderson et al., 2006). In jüngeren Publikationen werden zunehmend *MPC* (*model-predictive-control*)-Frameworks verwendet, um moralische Werte und Normen in Softwarelösungen zu integrieren (vgl. Németh, 2023; Pan et al., 2016).

Als Grundlage für Steuerungsalgorithmen bergen sogenannte *commandment models* diverse Probleme. So benötigen autonome Systeme aufgrund ihrer Softwarearchitektur als deterministische Automaten für alle Fälle klare Handlungsvorgaben, um von einem Zustand in den nachfolgenden gelangen zu können. Jedes Regelsystem stößt angesichts der Komplexität real-lebensweltlicher Situationen irgendwann an seine Grenzen (vgl. Goodall, 2014a, S. 62), sodass Unklarheiten auf verschiedene Weisen entstehen können. Einerseits können formale Prinzipien zu spezifisch sein, sodass sie in bestimmten Fällen nicht anwendbar sind. Je expliziter die Kriterien sind, desto schwieriger ist es, sie in einer Weise zu formalisieren, die von Maschinen verstanden wird (vgl. Goodall, 2014b, S. 98). Andererseits können Regeln aber auch zu allgemein sein, sodass sie der Komplexität der Entscheidungssituationen nicht gerecht, unter gewissen Voraussetzungen außer Kraft gesetzt oder gar unerfüllbar werden (vgl. Misselhorn, 2018b, S. 190, 2019, S. 50). So würden sowohl der Kategorische Imperativ als auch das erste Asimov'sche Gesetz ein autonomes Fahrsystem praktisch handlungsunfähig machen, indem sie grundsätzlich untersagen, dass einer Person durch eine Aktion eines Fahrroboters Schaden zugefügt wird. Jegliche Form eines probabilistischen Absolutismus, der alle Handlungen mit positiver Schadenswahrscheinlichkeit verbietet, erscheint in praktischer Hinsicht unplausibel. Wie Hansson (2013, S. 28–34) erläutert, trifft dies ebenfalls auf Vorgehensweisen zu, die auf einer Gewichtung von Pflichten proportional zur Eintrittswahrscheinlichkeit der korrespondierenden Ereignisse beruhen. Demnach sollen absolute Regeln nur in Fällen gelten, die oberhalb einer definierten Wahrscheinlichkeitsgrenze liegen. Im Grunde stellen sich dabei dieselben

146 Die Arbeiten von Beauchamp und Childress (1994) beziehen sich ursprünglich auf den Kontext der Medizinethik, lassen sich aber teilweise verallgemeinern. Sie benennen vier grundlegende Prinzipien, die *prima facie* gelten und bei Konflikten abgewogen werden müssen: Autonomie, Nichtschädigung, Fürsorge, Gerechtigkeit.

Probleme, die auch jedes utilitaristische Kalkül aufwirft: Was ist als hohe Wahrscheinlichkeit zu bewerten und was nicht? Wie können Vorteile, die eventuell durch eingegangene Risiken entstehen, einbezogen werden?

Zudem kann auch bei hierarchischen Ansätzen nicht ausgeschlossen werden, dass Konflikte zwischen einzelnen Regeln auf hoher Komplexitätsebene auftreten (vgl. Allen et al., 2005, S. 150). Entscheidungsdilemmata zeichnen sich im Speziellen dadurch aus, dass nicht verschiedene, sondern ein und dieselbe Regel zu widersprüchlichen Handlungsempfehlungen führt:

> From the mathematical perspective, dilemma situations represent cases that are mathematically infeasible. In other words, there is no choice of control inputs that can satisfy all of the constraints placed on the vehicle motion. The more constraints that are layered on the vehicle motion, the greater the possibility of encountering a dilemma situation where some constraint must be violated. Clearly, the vehicle must be programmed to do something in these situations beyond merely determining that no ideal action exists. (Gerdes & Thornton, 2015, S. 94)

In derartigen Situationen sind deontologische Gewichtungen wenig hilfreich. Vielmehr müsste das System über die festgeschriebenen Regeln hinausgehen und spezifische Fälle individuell evaluieren. Abwägungen, die sich an deontologischen Pflichten orientieren, sind höchst situativ, interpretationsbedürftig und kaum generalisierbar. Über entsprechende Fähigkeiten verfügen jedoch nur Menschen; für Maschinen erscheint es höchst problematisch, deontologische Implementierungen korrekt zu interpretieren und in wünschenswerte Aktionen umzusetzen: »These undesirable outcomes result from the inherent literalness of computers and from the inability of humans to articulate their own morals.« (Goodall, 2014a, S. 62)

4.4.4 Alternative Ansätze und pluralistische Frameworks

4.4.4.1 Tugendethische Ansätze

Tugendethische Konzepte befinden sich im digitalen Zeitalter wie nie auf dem Vormarsch (vgl. Ess, 2009; Spiekermann, 2015; Vallor, 2016). Im ethischen Kontext meist assoziiert mit den Schriften von Aristoteles, legen Tugenden keine konkreten Handlungsregeln fest,

sondern stellen den Charakter der handelnden Person in den Mittelpunkt. Dessen tugendhafte Konstitution strebt die Kultivierung eines guten Lebens an, das durch eine innere moralische Orientierung motiviert wird und in der Herausbildung moralischer Weisheit (*phronesis*) mündet. Tugenden können als ein tugendhaftes Leben befördernde, qualitative Merkmale des Charakters und dessen Fähigkeiten beschrieben werden, welche nicht gelehrt, sondern sich nur durch (Lebens-)Erfahrung angeeignet werden können (vgl. Hursthouse & Pettigrove, 2023). Individuelle Entscheidungen in spezifischen Situationen werden nicht durch normative Vorgaben geleitet, sondern dem Urteilsvermögen tugendhafter Personen überlassen: »A virtuous act is thus a rational act based on a wise, purposeful assessment of the factual situation, chosen for a pure motive and consistent with a steady disposition of the actor's character.« (Whetstone, 2001, S. 104)

Für den Kontext von Unfallalgorithmen erscheinen tugendethische Ansätze durchaus ansprechend, indem sie die dominanten konsequentialistischen und deontologischen Theorien komplementieren. Sie ermöglichen konsistente und kontextsensitive Entscheidungen, welche die zugrundeliegende Motivation der Handelnden hinterfragen und praktische Weisheit fördern (vgl. Gerdes, 2020, S. 110–111; Kumfer & Burgess, 2015, S. 133). Jedoch sind die Implikationen tugendethischer Entwürfe für konkrete Implementierungen (noch) unklar. Wie genau lassen sich ethische Tugenden in einem Maschinencode abbilden? Welche Rolle kann die Tugendethik in Design- und Gestaltungprozessen von Softwaresystemen spielen? In der Forschung existieren zwar einige grundsätzliche Ideen, die aber bisher nicht ausreichend konkretisiert wurden. Prinzipiell sind zwei Optionen vielversprechend: Erstens könnten tugendethische Entwürfe im weiteren Kontext von Systemen wirken, indem sie die Handlungen involvierter Personen während der Entwicklung und Nutzung der Systeme in einer Weise prägen, die resultierenden Schaden möglichst gering hält. Interpretiert als Paradigma der Technikethik würde die Tugendethik vorgeben, über welche Eigenschaften bzw. Tugenden ein guter Ingenieur verfügen sollte, z. B. Verantwortungsbewusstsein oder Aufrichtigkeit (vgl. Weber & Zoglauer, 2019, S. 149). Zusätzlich könnten Anwender durch ein spezifisches ethisches Design dazu motiviert werden, gewisse Tugenden auszubilden,

z. B. indem sie sich bei der Nutzung selbstfahrender Fahrzeuge vorsichtig und verantwortlich zeigen (vgl. Nyholm, 2018b, S. 6–7).

Zweitens ist auch denkbar, dass tugendethische Elemente direkt in den Softwarecode einfließen. Als mögliche ›Eigenschaften‹ eines tugendhaften Fahrsystems sind in der Literatur u. a. Fairness, Respekt gegenüber Autoritäten im Sinne von Gesetzen bzw. Regeln, Verantwortungsbewusstsein, Rücksichtnahme und Sorge um andere sowie Mut genannt worden (vgl. Gerdes, 2020, S. 111–113; Nyholm, 2018b, S. 6–7; Pan et al., 2016, S. 3–5). Tugenden lassen sich dabei als Resultat maschinellen Lernens generieren, indem das System aus einer Datenbasis über tugendhaftes Verhalten in Form einer ›Belohnungsfunktion‹ entsprechende Handlungsprinzipien erlernt. Vielversprechend ist auch die Idee, utilitaristisch oder deontologisch basierte Systeme um tugendethische Elemente zu ergänzen. So könnten Tugenden in Form relativer Gewichtungen für implementierte Kostenfunktionen oder Einschränkungen realisiert werden, die die Zielfunktionen autonomer Fahrsysteme dahingehend verändern, dass diese tugendhaftes Verhalten widerspiegeln (vgl. Gerdes & Thornton, 2015, S. 92). Auf diese Weise ließen sich unterschiedliche Instanzen einer Rollenmoral implementieren, die verschiedenen Typen von Fahrzeugen (private Personenbeförderung, Taxis, Krankenwagen, etc.) in spezifischen Verkehrssituationen besondere Zugeständnisse macht (vgl. Wang et al., 2022, S. 11). Fahrzeugen würde im Hinblick auf ihre spezifische soziale Rolle im übertragenen Sinne ein mit entsprechenden Tugenden ausgestatteter Charakter verliehen (vgl. Thornton et al., 2017, S. 1436–1437). Allerdings ist die Anwendung von Methoden maschinellen Lernens im Kontext ethischer Entscheidungsprobleme generell fragwürdig (siehe Kap. 4.1.2). Eine der größten Herausforderungen für tugendethisch basierte Ansätze besteht in der mangelnden Erklärbarkeit der zugrundeliegenden Entscheidungslogik, woraus sich insbesondere Schwierigkeiten bei der Zuschreibung von Verantwortung ergeben können (vgl. Geisslinger et al., 2021, S. 1040).

4.4.4.2 Vertragstheoretische Ansätze

Kontraktualistischen Ethikentwürfen liegt die Idee zugrunde, die Vertretbarkeit moralischer Grundsätze durch einen hypothetischen,

zwischen freien und gleichen Individuen geschlossenen Vertrag zu begründen. Normativ gültig ist das, was allgemein zustimmungsfähig ist. Bis dato versuchen einige wenige Beiträge, dem Prinzip der Schadensminimierung im Kontext von Unfallalgorithmen eine vertragstheoretische Grundlage zu geben, indem sie sich entweder auf John Rawls' (1971) politische Moral- und Gerechtigkeitstheorie im Sinne eines Gesellschaftsvertrags (*social contract*) oder die ethische Theorie von Thomas Scanlon (1998) beziehen. Auch wenn es nicht immer explizit deutlich gemacht wird, greifen entsprechende Argumentationen im einschlägigen Forschungsdiskurs zumeist auf eine heuristische Variante von Rawls' *Schleier des Nichtwissens* zurück. Dieser setzt Individuen im Rahmen der Entscheidungsfindung über einen Wertekonsens in Unkenntnis ihrer Rolle im gesellschaftlichen Gefüge und damit auch des Ausmaßes, in dem diese von den Folgen der getroffenen Entscheidung betroffen sein werden. Unter dieser Voraussetzung kann davon ausgegangen werden, dass sie als rationale Individuen einer Regelung zustimmen würden, welche die Interessen aller gleichermaßen mitbedenkt:

> […] I assume that the parties are situated behind a veil of ignorance. They do not know how the various alternatives will affect their own particular case and they are obliged to evaluate principles solely on the basis of general considerations. […] It is assumed, then, that the parties do not know certain kinds of particular facts. First of all, no one knows his place in society, his class position or social status; nor does he know his fortune in the distribution of natural assets and abilities, his intelligence and strength, and the like. Nor, again, does anyone know his conception of the good, the particulars of his rational plan of life, or even the special features of his psychology such as his aversion to risk or liability to optimism or pessimism. (Rawls, 1971, S. 136–137)

Im Rahmen einer spieltheoretischen Analyse modellieren Gogoll und Müller (2017) Entscheidungsdilemmata als Problem strategischer Interaktion. Sie zeigen, dass die Minimierung des Schadens nicht nur für die Gesamtgesellschaft, sondern auch für am Eigeninteresse orientierte Individuen optimal ist. Den bisher einzigen explizit kontraktualistisch argumentierenden Ansatz legt Leben (2017) vor. Unter Verwendung entscheidungstheoretischer Konzepte entwirft er einen formalen Algorithmus nach dem Vorbild von Rawls' Gerechtigkeitstheorie. Dessen Grundidee ist es, die vom Fahrzeug geschätzte Überlebenswahrscheinlichkeit für jede Partei bei jeder

möglichen Aktion zu erfassen und zu berechnen, welcher dieser Aktionen jeder Einzelne zustimmen würde, wenn er sich in einer ursprünglichen Verhandlungsposition der Fairness befände. Unter der Annahme eigennütziger Akteure würde dies derjenigen Alternative entsprechen, die im schlechtesten denkbaren Fall den geringsten Schaden für das am schlechtesten gestellte bzw. schwächste Individuum bedeutet.[147] Leben versteht seinen Ansatz als Antwort auf die Probleme, denen sich utilitaristische und deontologische Prinzipien gegenübersehen:

> The chief advantage of a Rawlsian algorithm is its respect for persons as equals, and its unwillingness to sacrifice the interests of one person for the interests of others. Certainly, this can produce surprising results, but ones that any Rawlsian believes the foundations of morality must inevitably lead one towards. (Ebd., S. 114)

Auch traditionelle vertragstheoretische Ansätze sind jedoch nur bedingt geeignet, um Dilemma-Szenarien zu entscheiden. Eine von Hansson (2013, S. 37–42) vorgelegte Kritik richtet sich im Allgemeinen gegen die Schwierigkeiten, die sich für kontraktualistische Ansätze insbesondere hinsichtlich Risiko und Unsicherheit ergeben. Die Zulässigkeit von schädigenden Handlungen könnte demnach lediglich über (hypothetische) Zustimmung gerechtfertigt werden, für welche allerdings die nötigen Voraussetzungen im lebensweltlichen Kontext nicht gegeben sind.[148] Eine spezifischere, kritische Auseinandersetzung mit Lebens Entwurf liefert Keeling (2018a). Er bemängelt, dass zwischen der ursprünglichen Rawls'schen Argumentation und ihrer Verwendung in Lebens Ansatz Unstimmigkeiten bestehen, beispielsweise in Bezug auf Lebens Annahmen zur Anwendung der *Maximin*-Regel. Der von Leben entworfene Algorithmus beinhaltet Implikationen, die in gewissen Szenarien-Konstellationen problematisch sein können; diese sollten durch ein unabhängiges Argument begründet werden.[149]

Hübner und White (2018) wenden sich ebenfalls gegen eine kontraktualistische Grundausrichtung bei der Programmierung von Un-

147 Neben Leben (2017) folgen auch Dogan et al. (2020) einer Argumentation zugunsten des am schlechtesten gestellten Individuums.

148 Eine intensive Auseinandersetzung mit dem risikoethischen Kriterium der Zustimmung erfolgt in Kap. 7.2.3.2 und Kap. 7.2.3.3.

149 Siehe auch Kap. 7.3.3.3 für eine detaillierte Darstellung relevanter Kritikpunkte.

fallalgorithmen. Ausgehend von Foot und unter Bezugnahme auf Harris' (1975) Gedankenexperiment der *Survival Lottery*, welche die Verteilung von Spenderorganen an utilitaristischen Zielgrößen bemisst, beziehen sie Stellung gegen die Eignung eines kontraktualistischen Entwurfs, vor allem wenn dieser auf Schadensminimierung abzielt. Dabei substanziieren sie ihre Kritik mit der mangelnden Unterscheidung zwischen positiven und negativen Rechten, die eine zentrale Herausforderung für Unfallalgorithmen darstellt, jedoch im Rahmen einer auf rationalem Eigeninteresse beruhenden Optimierungsentscheidung nicht berücksichtigt wird.

4.4.4.3 Rechtsphilosophische Ansätze

Neben dezidiert ethischen Untersuchungen sind im Diskurs auch rechtsphilosophische Ansätze wiederzufinden, die normative Prinzipien für Notsituationen in anderen Anwendungskontexten erörtern. Vor dem Hintergrund der deutschen Rechtsprechung stellt Hilgendorf (2018a, S. 683–690) einen kontroversen Ansatz vor, um Schadensminimierung und deontologische Prinzipien in Einklang zu bringen. Dabei greift er auf Wertungen tradierter Dilemma-Fälle zurück, die in der deutschen Jurisprudenz bereits existieren und sich grundsätzlich auf das *Prinzip des geringsten Übels* stützen bzw. dieses zur Anwendung bringen. Steht jedoch Leben gegen Leben, sind die Prinzipien der Menschenwürde, der Menschenrechte und der Rechtsstaatlichkeit höher zu priorisieren als das reine Abwägen von Übeln. Auch Gasser (2015, S. 558) schreibt: »Eine Abwägung mit dem gleichwertigen und im Fall des Lebens als verfassungsrechtlicher ›Höchstwert‹ geschütztes […] Grundrecht anderer Grundrechtsträger hat zu unterbleiben und ist unzulässig.«

Ein prominentes einschlägiges Beispiel aus neuerer Zeit ist das Urteil des Bundesverfassungsgerichts zum Luftsicherheitsgesetz, dem zufolge die Opferung unbeteiligter Menschen durch den Abschuss einer entführten Passagiermaschine grundsätzlich als rechtswidrig einzustufen ist. Dies gilt auch dann, wenn auf diese Weise eine größere Zahl von Menschenleben gerettet werden könnte (vgl. BVerfG, 2006). Dieses Urteil lässt sich aufgrund seiner situativen Merkmale prinzipiell auf Entscheidungsdilemmata im autonomen Fahren übertragen. Jedoch weist Hilgendorf (2019, S. 368–370) da-

rauf hin, dass im Fall von Unfallalgorithmen eine juristisch relevante Unterscheidung zwischen symmetrischer und asymmetrischer Gefahrenlage bedeutsam würde, die folgende Implikationen hat: Das zuvor begründete Quantifizierungs- und Abwägungsverbot ist nicht absolut gültig, sondern lediglich in Fällen, in denen eine Aufopferungspflicht Unbeteiligter ausgeschlossen ist. Während eine quantitative Abwägung im Sinne des *Prinzips des geringsten Übels* nur zulässig ist, sofern alle Beteiligten von Anfang an gleichermaßen gefährdet sind, bleibt das Opfern zunächst Unbeteiligter zugunsten anderer prinzipiell unzulässig. Begründen ließe sich dies sowohl rechtsethisch als auch verfassungsrechtlich; eine Verpflichtung zur Selbstopferung existiert für Insassen auch aus juristischer Sicht nicht:[150]

> Auch im Notstand dürfen Menschenleben daher nicht gegeneinander ›aufgerechnet‹ werden. Nach dieser Position ist das Individuum als ›sakrosankt‹ anzusehen; dem Einzelnen dürfen keine Solidarpflichten auferlegt werden, sich für andere aufzuopfern, auch dann nicht, wenn nur so andere Menschen gerettet werden können. (Di Fabio et al., 2017, S. 18)

Auch wenn diese Aspekte grundsätzlich plausibel erscheinen, lässt sich damit eine Programmierung auf Schadensminimierung letztlich nicht rechtfertigen. Hilgendorf (2018a, S. 692) hält fest: »Aus der Tatsache, dass ein Umsteuern des Wagens als rechtswidrig einzustufen ist, folgt nicht, dass ein Weiterfahren rechtmäßig wäre. Es handelt sich vielmehr um eine tragische Situation, in der jede mögliche Geschehensvariante Unrecht verwirklicht.« Wie soll ein Fahrzeug nun agieren, wenn es aus rechtlicher Sicht keine ›richtige‹ Option gibt? Sind nicht das Existieren und zugleich die Möglichkeit einer straffreien, rechtstreuen Alternative Voraussetzungen dafür, wie eine Handlung juristisch zu bewerten ist? Aus der dilemmatischen Struktur der betrachteten Entscheidungssituationen folgt analog zur ethischen Perspektive, dass ausgehend von der heutigen Grundrechtsdogmatik keine der Alternativen juristisch ›korrekt‹ ist. Hinzu kommt, dass Fahrroboter nicht den juristischen – und mora-

150 Ob eine Handlung aus juristischer Sicht als rechtswidrig einzustufen ist, ist zunächst unabhängig von der Frage, wie diese zu bestrafen ist. Daher würde eine primär strafrechtliche Begründung des Verbots der Opferung Unbeteiligter an dieser Stelle zu kurz greifen.

lischen – Status menschlicher Akteure besitzen. Somit ist die Problematik strafrechtlicher Bewertungen in ihrem Fall besonders komplex und eng verknüpft mit Fragen der zivil- und strafrechtlichen Produkthaftung.

Ungeachtet der juristischen Bewertung besteht für Hilgendorf dennoch eine moralische Pflicht, so viele unschuldige Leben wie möglich zu retten. Was bedeutet dies in Hinsicht auf das Aufrechnungsverbot? Eine differenzierte kritische Antwort auf Hilgendorf findet sich bei Misselhorn (2018b, S. 193–195), die sich der Problematik über eine Auseinandersetzung mit den Grundideen der kantianischen Ethik nähert. Sie stellt die Frage nach der moralischen Zulässigkeit einer Schadensminimierung in den Zusammenhang der Universalisierbarkeit von Normen: *Dürfen* Personen geopfert werden, oder *sollten* sie es? Misselhorn veranschaulicht ihre Argumentation, indem sie auf Kants Unterscheidung zwischen verschiedenen Graden der Verbindlichkeit moralischer Pflichten zurückgreift, welche dieser in seiner *Grundlegung zur Metaphysik der Sitten* (1900ff., GMS, AA 04) anhand verschiedener deontischer Kategorien vornimmt. Während vollkommene Pflichten rechtlich bindend sind, bleiben unvollkommene Pflichten eher vage und unbestimmt in der Art und Weise, wie sie zu erfüllen sind. An Hilgendorfs Argumentation kritisiert Misselhorn nun, dass dieser die Zulässigkeit der Opferung Unschuldiger unter gewissen Umständen im Sinne einer vollkommenen Pflicht interpretiert, was die Tötung Unschuldiger zum moralischen Gebot erhebt. Im Hinblick auf Kants vollkommene Pflicht, unschuldige Personen unter keinen Umständen zu töten, kann dies jedoch unmöglich geboten sein – unabhängig davon, ob es sich bei Schadensminimierung um eine vollkommene oder unvollkommene Pflicht handelt. Auch Schlussfolgerungen über ein grundsätzliches Erlaubt-Sein derartiger Handlungen, wie sie Hilgendorf im Rahmen seiner Unterscheidung zwischen »zunächst Unbeteiligten« und »von Anfang an Beteiligten« zieht, sind an dieser Stelle inadäquat und können die Tötung Unschuldiger nicht rechtfertigen.

Weitere rechtsphilosophische Auseinandersetzungen beziehen sich auf spezifische Elemente der Rechtsprechung. Eine kritische Rekonstruktion der Prinzipien und Rechtsnormen, die der vor allem in der anglo-amerikanischen Jurisprudenz verankerten Notwendigkeitslehre (*doctrine of necessity*) zugrunde liegen, präsentiert Santoni de Sio (2017). Er analysiert, inwiefern juristische Konzepte wie

Notwehr auf den Kontext von Unfallalgorithmen anwendbar sind und welche Implikationen sich daraus für eine Programmierung auf Schadensminimierung nach utilitaristischer Lesart ergeben. Dabei geht er insbesondere auf das Problem der Inkommensurabilität ein, die aus rechtlicher Sicht das zentrale Hindernis für ein rein utilitaristisches Kalkül darstellt. Er diskutiert im weiteren Sinne auch anwendungsorientierte vertragstheoretische Komponenten wie beispielsweise vertragliche Übereinkünfte durch spezifische Versicherungspolicen, die er aber aus praktischen Gründen für ungeeignet befindet. Als kritische Antwort auf Santoni de Sio formuliert Keeling (2018b) eine alternative Interpretation der Notwendigkeitslehre. Zentrales Element ist ein eingeschränktes Pareto-Prinzip, das sich – im Gegensatz zu Santoni de Sios Enwurf – aus utilitaristischer, deontologischer und kontraktualistischer Sicht gleichermaßen verteidigen lässt. Auf dieser Grundlage erarbeitet Keeling einen Vorschlag, der sich auf bestimmte, ausgewählte Dilemma-Szenarien anwenden lässt.

Coca-Vila (2018) wiederum greift die Problematik von Dilemma-Situationen aus der Perspektive des Strafrechts auf. Er erklärt, dass die Maximierung des gesellschaftlichen Nutzens keine schädlichen Eingriffe in die Rechtssphäre einer Person rechtfertigt und sieht das Prinzip der Schadensminimierung daher prinzipiell im Widerspruch zu den Grundsätzen eines liberalen Rechtssystems. Als mögliche Strategie schlägt er vor, die Programmierung von Unfallalgorithmen auf die Basis eines deontologischen Verständnisses von strafrechtlichen Doktrinen der Rechtfertigung zu stellen. Als Voraussetzung für eine Entscheidung von Dilemmata fordert er eine vorherige Analyse der Rechtspositionen aller Beteiligten, die im Hinblick auf Prinzipien der Autonomie und Solidarität zu erfolgen hat.

4.4.4.4 Meta-normative Ansätze

Der überwiegende Teil des Diskurses um Unfalldilemmata geht von einem moralischen Universalismus aus, demzufolge moralische Prinzipien objektiv und allgemeingültig sind. Einige wenige Ansätze jedoch beurteilen die Problemstellung aus dem Blickwinkel eines moralischen Relativismus, der die Gültigkeit moralischer Urteile stets an die kulturellen und sozialen Gegebenheiten innerhalb von

Gesellschaften oder Gruppen knüpft. In diesem Sinne erläutern Bhargava und Kim (2017), dass moralische Unsicherheit – die Unsicherheit dahingehend, dass nicht eindeutig erkennbar ist, was in einer bestimmten Situation moralisch richtig ist – der Ausgangspunkt der ethischen Debatte über Unfallalgorithmen sein sollte. Um zwischen den normativen Vorschriften konkurrierender Moraltheorien vermitteln zu können, sei ein meta-normatives Framework notwendig. Sie schlagen vor, dieses an einem erwarteten moralischen Wert (*expected moral value*) auszurichten, in dessen Rahmen relevanten ethischen Aspekten einer Entscheidungssituation quantitative Werte zugewiesen werden, um (mathematisch) eindeutige Lösungen zu ermitteln:

> As such, an adequate solution to the problem of moral uncertainty must take into account the moral values associated with the particular normative proposition, weighted by their respective probabilities, not merely the probability that the normative proposition in question is true. (Ebd., S. 9)

Die Idee eines relativistischen Fahrzeugs im Sinne eines allgemeinen Begriffsverständnisses stellt Pölzler (2021) zur Diskussion. Anstelle von Prinzipien, die konkreten ethischen Theorien entstammen, kommen dabei spezifische Verfahren zur Entscheidungsfindung (*decision-making procedures*) zur Anwendung. Diese legen vernünftige moralische Forderungen zugrunde, welche unabhängig von spezifischen ethischen Prinzipien akzeptabel sein können. Bei diesem Ansatz geht es also nicht darum, die moralisch akzeptabelste Antwort zu finden, sondern akzeptable algorithmische Verfahrensweisen zu bestimmen, die dann ihrerseits Antworten generieren. Eine konkrete Entscheidungsstrategie, die im Kontext von Unfallalgorithmen aufgegriffen wurde, ist die sogenannte *Ethical Valence Theory*, die den Entscheidungsprozess in unvermeidbaren Unfallsituationen als algorithmische Vermittlung ethischer Forderungen auffasst:

> […] different road users hold different moral claims on the vehicle's behavior, and the vehicle must mitigate these claims as it makes decisions about its environment. Specifically, it must find an optimal response to these claims in cases of unavoidable collision, or in ›dilemma scenarios‹; one which captures most efficiently the moral claims and relations which exist within the vehicle's decision context, and aligns best with user expectations. (Evans et al., 2020, S. 3286)

Dieser Ansatz hat die Form eines Optimierungsproblems: Für konkrete Situationen werden zunächst relevante Forderungen identifiziert und sodann mit relativen Gewichtungen bewertet, welche schließlich in einer Weise verrechnet werden, die das Ergebnis optimiert. Zwei Faktoren sind bei der Bewertung dabei jeweils wichtig: zum einen das Ausmaß des Schadens, der im Fall einer Nichtberücksichtigung der Forderung entsteht, und zum anderen deren ethische Valenz, d. h. der Grad ihrer sozialen Akzeptanz (vgl. Evans et al., 2020).

4.4.4.5 Pluralistische Frameworks

Wie zuvor dargestellt, stoßen traditionelle ethische Prinzipien und Theorien angesichts der spezifischen Herausforderungen von Unfallalgorithmen an ihre Grenzen. Einige Forscher und Philosophen reagieren auf diese Problematik mit der Forderung eines pluralistischen Frameworks (vgl. z. B. Brändle & Schmidt, 2021; Goodall, 2014b, 2020; Hübner & White, 2018; Nyholm, 2018b; Wang et al., 2020). Dabei wirken Entscheidungsprinzipien aus verschiedenen ethischen Denktraditionen zusammen, um deren jeweilige soziale, moralische, rechtliche und funktionale Vor- und Nachteile zu integrieren (vgl. Poszler et al., 2023, S. 5–15). Wallach und Allen (2008, S. 78) führen dazu aus: »Given the range of perspectives regarding the morality of specific values, behaviors, and lifestyles, perhaps there is no single answer to the question of whose morality or what morality should be implemented in AI.«

Pluralistische Konzeptionen lassen sich in softwaretechnischer Hinsicht über hybride Softwarearchitekturen realisieren.[151] Einen konkreten Vorschlag diesbezüglich legen Gerdes und Thornton (2015) vor, indem sie ethische Konzepte in mathematische übersetzen. Dabei werden Unfallalgorithmen grundsätzlich als (utilitaristisches) Optimierungsproblem interpretiert, das deontologische Prinzipien in Form von Nebenbedingungen berücksichtigt.[152] Um zu gewährleisten, dass das System auch im (andernfalls mathematisch

151 Poszler et al. (2023, S. 15) stellen verschiedene Möglichkeiten für hybride Kombinationen überblicksartig vor.

152 Sütfeld et al. (2019) schlagen eine ähnliche Konzeption vor: »A possible solution to unite robustness of the decision making logic and reasonableness of the

unlösbaren) Dilemma-Fall zu einer Entscheidung gelangen kann, werden diese als sogenannte weiche Bedingungen (*soft constraints*) implementiert, deren Verletzung mit Kosten unterschiedlicher Höhe sanktioniert wird. Diese Bedingungen können bei Bedarf überschritten werden, nachdem eine utilitaristisch basierte Gewichtung der Kosten möglicher Handlungsoptionen stattgefunden hat:

> From the mathematical perspective, dilemma situations represent cases that are mathematically infeasible. In other words, there is no choice of control inputs that can satisfy all of the constraints placed on the vehicle motion. The more constraints that are layered on the vehicle motion, the greater the possibility of encountering a dilemma situation where some constraint must be violated. Clearly, the vehicle must be programmed to do something in these situations beyond merely determining that no ideal action exists. A common approach in solving optimization problems with constraints is to implement the constraint as a ›soft constraint‹ or slack variable […]. The constraint normally holds but, when the problem becomes infeasible, the solver replaces it with a very high cost. In this way, the system can be guaranteed to find some solution to the problem and will make its best effort to reduce constraint violation. A hierarchy of constraints can be enforced by placing higher weights on the costs of violating certain constraints relative to others. The vehicle then operates according to deontological rules or constraints until it reaches a dilemma situation; in such situations, the weight or hierarchy placed on different constraints resolves the dilemma, again drawing on a consequentialist approach. (Ebd., S. 94–95)

Obwohl Verkehrsregeln von Natur aus deontologisch sind, werden sie in der Praxis oft konsequentialistisch behandelt: Nicht selten übertreten wir bewusst Regeln im Interesse anderer Ziele, z. B. bei medizinischen Notfällen oder um einen konstanten, risikoärmeren Verkehrsfluss zu gewährleisten. Wir wägen also implizit die Einhaltung der jeweiligen Regel in Bezug auf ihren Nutzen und ihre Kosten

---

resulting decision would be to conceptualize ethically relevant properties on a continuous scale, and treat moral rules as soft constraints to the car's behavior. This would allow for a compromise between deontological and utilitarian considerations. The system would principally base its decisions on a comparison of the stakes involved for different parties in a situation, but could additionally disincentivise against the violation of important moral rules, as well as traffic violations.«

in spezifischen Situationen ab.[153] Diesen Umstand greifen Thornton et al. (2017) auf, indem sie den ursprünglichen Ansatz erweitern und eine Gewichtung der Kosten bzw. der Stärke der deontologischen Bedingungen anhand einer auf tugendethischen Überlegungen basierenden Rollenmoral in ihre Überlegungen einbeziehen, die eine Ausdifferenzierung spezifischen Verhaltens für verschiedene Typen von Fahrzeugen erlaubt.

Befürworter pluralistischer Frameworks betonen häufig, dass durch eine Integration von Komponenten verschiedener ethischer Theorien nicht nur die jeweiligen Schwächen der einzelnen Konzepte aufgefangen werden können, sondern auch dem Pluralismus ethischer Wertvorstellungen entsprochen wird, der liberale Gesellschaften in lokaler und globaler Perspektive prägt. Allerdings sind auch pluralistische Ansätze nicht frei von Schwachstellen. Es ist z. B. unklar, inwiefern sich die spezifischen Schwächen der jeweiligen Prinzipien bzw. Theorien überzeugend kompensieren lassen. Das sich aus seiner mangelnden Berücksichtigung individueller Rechte ergebende Konfliktpotenzial des utilitaristischen Ansatzes ließe sich beispielsweise nur mithilfe eines sehr komplexen Systems deontologischer Einschränkungen entschärfen, welches seinerseits die spezifischen Problematiken deontologischer Regelsysteme aufwirft.

Andere Ansätze wie tugendethische Überlegungen sind dagegen hinsichtlich ihrer Implementierbarkeit noch nicht ausgereift genug, um für angewandte Probleme im Kontext von Systemen Künstlicher Intelligenz kurzfristig in Frage zu kommen. Auch ist nicht trivial ersichtlich, wie pluralistischen moralischen Werturteilen durch ein solches Framework adäquat entsprochen werden kann, wenn favorisierte Werte bzw. Prinzipien stets durch andere in ihrer Geltung eingeschränkt sind. Zumindest angesichts des gegenwärtigen Forschungsstands erscheint es wenig nachvollziehbar, pluralistische Frameworks aus Komponenten traditioneller ethischer Theorien als *die* Strategie für die Programmierung von Unfallalgorithmen zu betrachten. Auch ihnen gelingt es nicht, die strukturellen Probleme, die tradierten ethischen Prinzipien im Hinblick auf ihre Operationalisierbarkeit und Rechtfertigungsgrundlage anhaften, vollständig aufzulösen.

153 Sütfeld et al. (2019, S. 13–14) merken an, dass Regelübertretungen u. U. nicht nur akzeptabel, sondern sogar ethisch geboten sein können.

4.5 Zwischenergebnis: Ungeklärte Fragen des Diskurses

Der relevante Forschungsdiskurs wird von einem methodischen Zugang dominiert, der Unfallalgorithmen als moralisches Designproblem begreift. Dilemma-Szenarien werden dabei mehrheitlich als spezifische Instanzen eines modifizierten Trolley-Problems gedeutet. Eine eingehende Analyse bisheriger Forschungsbeiträge enthüllt jedoch einige Unstimmigkeiten dieser Vorgehensweise. Sie belegt die erste zentrale These dieser Arbeit, welche besagt, dass bis dato vorherrschende Herangehensweisen an die Gestaltung von Unfallalgorithmen viele für die Entwicklung und den Einsatz autonomer Fahrsysteme essenzielle Fragen ungeklärt lassen.

Wie in diesem Kapitel dargelegt, lässt sich diese Schlussfolgerung anhand verschiedener Argumente begründen. Aus der strukturierten Analyse einschlägiger Forschungsliteratur wird deutlich, dass sich Entscheidungssituationen im Kontext autonomer Fahrsysteme aufgrund struktureller und epistemischer Unterschiede nicht adäquat mithilfe des Trolley-Frameworks darstellen lassen. Aus dem Blickwinkel des gesellschaftlichen Kontextes, in den die Thematik der Gestaltung von Unfallalgorithmen eingebettet ist, treten methodische Schwierigkeiten dominanter Forschungszugänge hervor. Der spezifischen Komplexität ethischer, sozialer und rechtlicher Verflechtungen, welche lebensweltliche Dilemma-Szenarien charakterisieren, kann eine Reduzierung auf das Framework des Trolley-Problems nicht gerecht werden. Entscheidungen über Unfallalgorithmen orientieren sich nicht primär an individuellen moralischen Urteilen, sondern werden durch allgemein akzeptierte gesellschaftliche Moralvorstellungen geleitet. Strategien für die Programmierung selbstfahrender Fahrzeuge als sozio-technische Systeme müssen mit einem allgemein akzeptierten Wertekodex vereinbar sein, dessen ethische Begründbarkeit Gegenstand einer kontinuierlichen kritischen Prüfung ist. Die gesellschaftlich-soziale Dimension von Unfallalgorithmen wird bei einer Fokussierung auf das Trolley-Problem ausgeblendet, welches sich lediglich auf der Ebene individueller Moralpräferenzen bewegt. Letztere geht an der eigentlichen praktischen Problemstellung vorbei, die nach den normativen Grundlagen sozialen Zusammenlebens fragt und regulierende Richtlinien erfordert. Mögliche alternative Ansätze der politischen Philosophie stellen

zum gegenwärtigen Zeitpunkt lediglich impulsartige Heuristiken dar, woraus sich Desiderate für weiterführende Forschung ableiten.

Weiterhin versäumen es bisherige Ansätze, zentrale entscheidungstheoretische Charakteristika von Dilemma-Situationen miteinzubeziehen. So sind Entscheidungen über die Programmierung von Unfallalgorithmen stets mit moralisch relevanten Unsicherheiten bezüglich der Handlungsfolgen bzw. des Eintretens verursachter Umweltzustände behaftet. Unter dem Framework des Trolley-Problems werden diese jedoch fälschlicherweise als Entscheidungen unter Sicherheit verstanden. Daraus ergeben sich Schwierigkeiten für die normative Bewertung. Die mangelnde Berücksichtigung von Risiken und Unsicherheiten ist wesentlich dafür verantwortlich, dass bisherige normative Begründungsansätze aus dem Bereich der philosophischen Ethik als mögliche Entscheidungsstrategien an ihre Grenzen stoßen. Aufgrund inhärenter Schwächen vor allem hinsichtlich ihrer Operationalisierbarkeit im real-lebensweltlichen Problemkontext können weder utilitaristische noch deontologische, tugendethische, kontraktualistische oder pluralistisch orientierte Ansätze rechtfertigbare Entscheidungsstrategien final begründen.

Auch Ansätze aus der Moralpsychologie, die auf Methoden experimenteller Ethik zurückgreifen, weisen in Bezug auf ihre Eignung für die Problematik erhebliche Nachteile auf. Zum einen sind sie als deskriptive Ansätze für normative Fragestellungen grundsätzlich methodisch fragwürdig. Zum anderen beruht das Design der im Kontext von Dilemma-Situationen durchgeführten empirischen Studien ebenfalls auf Instanzen modifizierter, bezogen auf den praktischen Problemkontext inadäquater Trolley-Szenarien; deren eigentlicher Zweck besteht nicht darin, Probleme zu lösen, sondern sie aufzuwerfen. Die mit dem Trolley-Problem assoziierte sogenannte *Trolleyology*-Methodik will nicht als Modell für einzelne angewandte Situationen fungieren, sondern zielt darauf ab, moralische Werturteile zwischen verschiedenen Fällen zu vergleichen und Erklärungen für mögliche Unterschiede in deren intuitiver Bewertung zu erforschen. Innerhalb des Experiments gewonnene Erkenntnisse lassen sich nicht unmittelbar in Form von Prinzipien oder Regeln auf lebensweltliche Probleme übertragen.

Schließlich begründet die in diesem Kapitel vorgelegte Analyse eine grundlegende Skepsis gegenüber bis dato dominanten Forschungszugängen. Um geeignete Entscheidungsstrategien zu erarbei-

ten, muss der ethische Diskurs über eine angewandte Trolley-Perspektive hinausgehen. Im nachfolgenden Kapitel wird der Horizont des hier entwickelten Arguments durch eine metaethische Analyse moralischer Dilemma-Strukturen erweitert. Wie sich zeigen wird, ergeben sich auch aus diesem Blickwinkel sowohl praktische als auch theoretisch-formale Implikationen, die im Rahmen bisheriger Forschungszugänge unberücksichtigt geblieben sind.

5. Die Komplexität moralischer Dilemma-Strukturen: Rekonstruktion aus metaethischer Sicht

Dilemmatische Problemstrukturen, wie sie sich im Kontext unvermeidbarer Unfallsituationen manifestieren, sind vielschichtig und erfordern eine gründliche ethische Auseinandersetzung. Um eine solche zu gewährleisten, sind nicht nur anwendungsbezogene Überlegungen im Speziellen notwendig, sondern auch ein Anknüpfen an bestehende Analysen der zugrundeliegenden metaethischen Problematik von Dilemmata im Allgemeinen. Daher wird in diesem fünften Kapitel das formale Problem, welches moralische Dilemma-Strukturen auszeichnet, zunächst metaethisch rekonstruiert und schließlich im Hinblick auf praktische Dilemma-Situationen im Kontext des autonomen Fahrens erörtert.[154] Die nachfolgende metaethische Diskussion ist weitgehend abstrakt gehalten, was sie aber auch sein muss, um das Verhältnis zwischen metaphysischen, logischen und pragmatischen Zusammenhängen zu klären, die für den Anwendungskontext relevant sind.

5.1 Einführung: Dilemmata als Grenzsituationen moralischen Handelns

5.1.1 Beispiele und Narrative aus Philosophie, Literatur und lebenspraktischen Kontexten

Moralische Dilemmata gehören zu den anspruchsvollsten, zeitlos relevanten Problemstrukturen der Ethik und Moralphilosophie. Ob die (utilitaristische) Aufrechnung von Menschenleben, die absolute

154 Eine metaethische Analyse der spezifischen Dilemmastrukturen, die im Kontext des autonomen Fahrens auftreten können, hat die Autorin bereits an anderer Stelle publiziert (vgl. Schäffner, 2024).

Gültigkeit vollkommener Pflichten oder die Vermittlung zwischen Egoismus und Altruismus – die Beantwortung der von Immanuel Kant formulierten Grundfrage der Ethik ›Was soll ich tun?‹ ist in Entscheidungssituationen mit dilemmatischen Strukturen besonders herausfordernd. Angesichts der immer komplexer werdenden gesellschaftlichen Wirklichkeit treten moralische Dilemmata zunehmend in vielfältigen praktischen Anwendungszusammenhängen auf. Sowohl in metaethischen als auch anwendungsbezogenen Diskursen der Gegenwart wird dabei häufig auf tradierte Beispiele Bezug genommen, von denen im Folgenden zunächst einige exemplarisch vorgestellt werden.

Bereits antike Philosophen und Dichter griffen zur nachdrücklichen Veranschaulichung ihrer Lehren auf Narrative mit dilemmatischen Strukturen zurück. So schildert Sokrates im Gespräch mit Kephalos über das Wesen der Gerechtigkeit eine beispielhafte Situation, in der jemand von einem Freund eine Waffe zur Verwahrung empfangen hat. Als dieser sie nun mit offensichtlich fragwürdiger Gesinnung zurückfordert, sieht sich der Akteur mit einem Konflikt zweier moralischer Normen konfrontiert: Einerseits soll er die Waffe ihrem Besitzer zurückgeben, andererseits kann es jedoch auch als seine Pflicht angesehen werden, Dritte vor möglichen Übergriffen seitens des Freundes zu bewahren. Gemäß Sokrates' Ansicht genießt Letzteres Priorität vor der moralischen Verpflichtung zur Aushändigung der Waffe (vgl. Platon, ca. 375 v. Chr., 331b-c). Ein weiteres antikes Beispiel ist das Dilemma Agamemnons aus der gleichnamigen Tragödie des griechischen Dichters Aischylos (ca. 458 v. Chr.). Mitten im Trojanischen Krieg wird Agamemnons Kriegsflotte von der Göttin Artemis aufgehalten; diese lässt sich nur durch die Opferung von Agamemnons Tochter Iphigenie besänftigen. Das Dilemma besteht hier zwischen Agamemnons konfligierenden Verpflichtungen als Befehlshaber seiner Truppen einerseits und Vater seiner Tochter andererseits.

Auch in Philosophie und Literatur der neueren Zeit sind dilemmatische Entscheidungsprobleme allgegenwärtig. Der französische Philosoph Jean-Paul Sartre (1946) beschreibt den inneren Konflikt eines Schülers, der Rachegefühle wegen seines gefallenen Bruders hegt und sich deshalb den französischen Streitkräften anschließen möchte. Zugleich fühlt er sich jedoch auch der Fürsorge seiner alleinlebenden Mutter verpflichtet:

> Er fand sich also zwei sehr verschiedenen Typen von Handlungen gegenüber: einer konkreten, unmittelbaren, die allerdings nur einem Individuum galt; oder einer anderen, die sich auf eine unendlich größere Gesamtheit, eine nationale Kollektivität richtete, die aber eben dadurch zweideutig war und auf ihrem Weg unterbrochen werden konnte. Zugleich schwankte er zwischen zwei Typen von Moral. Einerseits eine Moral der Sympathie, der individuellen Hingabe; andererseits eine weiter gespannte Moral, jedoch von fragwürdigerer Wirksamkeit. Er musste zwischen beiden wählen. (Ebd., S. 156–157)

An Tragik kaum zu überbieten ist das Dilemma der Protagonistin des Romans *Sophie's Choice*, welcher aus der Feder des amerikanischen Autors William Styron (1980) stammt. Er erzählt die Geschichte der Polin Sophie, die während des Zweiten Weltkriegs zusammen mit ihren beiden Kindern in ein Konzentrationslager deportiert wird. Direkt nach ihrer Ankunft wird sie von einem Aufseher vor die Wahl gestellt, welches ihrer beiden Kinder getötet werden soll; trifft sie keine Entscheidung, so werden beide Kinder getötet.

Außer in klassischen philosophischen und literarischen Beispielen treten moralische Konflikte mit dilemmatischen Strukturen auch in verschiedenen modernen Anwendungskontexten der realen Lebenswelt auf. Weithin bekannt ist das sogenannte Terroristen-Dilemma, das in der Frage besteht, ob Folter im Zuge der Kriminalitäts- und Terrorbekämpfung vertretbar ist, um von einem Verdächtigen Informationen zu erhalten, die notwendig sind, um Unbeteiligte zu schützen. Es stehen sich dabei das Grundrecht der Unantastbarkeit der Menschenwürde und die moralische Verpflichtung zum Schutz der Bevölkerung gegenüber. Eine weitere, häufig referenzierte beispielhafte Entscheidungssituation entwirft Foot (1978): Das Leben einer werdenden Mutter kann nur durch einen komplizierten operativen Eingriff gerettet werden, bei dem allerdings das ungeborene Kind mit hoher Wahrscheinlichkeit sterben würde. Wessen Lebensrecht kann bzw. soll hier Vorrang haben?

Eines der berühmtesten Beispiele ist das sogenannte Gefangenendilemma (*prisoner's dilemma*). Heutzutage wird es vorwiegend in der mathematischen Spieltheorie verwendet, um strategische Entscheidungssituationen mit mehreren Interagierenden zu modellieren: Zwei Untersuchungshäftlinge werden zu einem gemeinsam verübten Verbrechen unabhängig voneinander verhört und stehen

vor der Wahl zwischen zwei Handlungsalternativen: gestehen oder schweigen. Wenn beide Häftlinge schweigen, werden sie aufgrund kleinerer Delikte zu einer geringen Haftstrafe verurteilt. Gestehen beide, bekommen sie jeweils eine moderate Strafe. Für den Fall, dass nur einer der beiden gesteht, kommt dieser mit einer geringfügigen Strafe davon, während der jeweils andere Komplize die Höchststrafe verbüßen muss. Aufgrund der Unsicherheit über die Entscheidung des anderen erscheint es aus kollektiver Sichtweise für beide optimal zu schweigen. Jedoch liegen für jeden der beiden Anreize vor, von dieser Strategie abzuweichen und dadurch die persönliche Strafe durch ein Geständnis zu reduzieren. Dies gilt allerdings nur, solange der jeweils andere weiterhin schweigt. Nur durch ein beiderseitiges Geständnis entsteht eine Situation, in der keiner sich durch Schweigen besserstellen kann und daher keinen Anreiz mehr hat abzuweichen (vgl. Lütge, 2011, S. 17–18). Dies wird in der Spieltheorie als Nash-Gleichgewicht bezeichnet.[155] Es erscheint jedoch paradox, dass die beiden Komplizen eine Strafe in Kauf nehmen, die sie hätten vermeiden können, wenn sie beide geschwiegen hätten. Das Dilemma liegt hier also darin, dass die Gleichgewichtsstrategie ›Gestehen – Gestehen‹ zu einem sowohl individuell als auch kollektiv betrachtet schlechteren Ergebnis führt.

Mit seiner Darstellung des Konflikts zwischen individueller Optimierung und kollektiver Vernunft, der durch die wechselseitige Interdependenz der individuellen Ergebnisse für beide Akteure verschärft wird, hat das Gefangenendilemma weitreichende Berühmtheit erlangt und wurde in einer Vielzahl von Disziplinen rezipiert. Ein stetiger Prozess der Neuinterpretation seiner Dilemma-Strukturen macht es zu einem beliebten Modell für diverse fachspezifische Fragestellungen, z. B. in Ökonomie, Soziologie, Biologie, Psychologie und Rechtswissenschaften. Auch in der Philosophiegeschichte finden sich die wesentlichen Strukturen des Gefangenendilem-

155 Die Bezeichnung ›Nash-Gleichgewicht‹ geht zurück auf den amerikanischen Mathematiker John Forbes Nash. Ein solches Gleichgewicht liegt genau dann vor, wenn kein Spieler unter den gegebenen Strategien der anderen einen Anreiz hat, als Einziger von seiner gewählten Strategie abzuweichen (vgl. Diekmann & Voss, 2004, S. 23). Weitere Annahmen, die dem Gefangenendilemma zugrunde liegen, sind: Die Häftlinge sind räumlich getrennt voneinander untergebracht, was eine Abstimmung ihrer Aussagen unmöglich macht. Beide Akteure sind zudem streng rational.

mas an verschiedenen Stellen wieder, beispielsweise bei Aristoteles, Locke oder Hume (vgl. Lütge, 2011). Der Philosoph Julian Nida-Rümelin (vgl. 1993, Kap. 14) betrachtet das Gefangenendilemma als ein Paradigma, dem jede Form kooperativen Verhaltens folgt. Nicht zuletzt hat das Gefangenendilemma auch eine Relevanz für die Ethik. Indem es zeigt, dass individuell egoistisches Verhalten ohne das Berücksichtigen der Interessen anderer zu einem suboptimalen Ergebnis führt, wird es oft als Indikator interpretiert, der einem rationalen Egoismus seine Grenzen aufzeigt.

Nicht zuletzt bringt die technologische Durchdringung moderner Gesellschaften neue Herausforderungen für ethische Entscheidungsprobleme mit sich. Dies erleben wir auch angesichts von Unfallszenarien des autonomen Fahrens: Wie soll sich eine Maschine in Situationen verhalten, mit denen sich bisher nur Menschen konfrontiert sahen? Dilemmata sind Grenzsituationen moralischen Handelns, die tradierte moralische Werte und Normen – unseren ›moralischen Kompass‹ – kritisch zur Diskussion stellen. Um diese Problematik wissenschaftlich untersuchen zu können, wird im folgenden Unterkapitel zunächst eine präzise Definition der relevanten dilemmatischen Problemstrukturen entwickelt. Diese erlaubt es nicht nur, Situationen zweifelsfrei als Dilemmata zu identifizieren, sondern umreißt auch bereits grob die zentralen Schwierigkeiten, die im Rahmen möglicher Entscheidungsstrategien zu bewältigen sind.

5.1.2 Kriterien und Definition moralischer Dilemma-Strukturen

Aus theoretisch-formaler Sicht stellen moralische Dilemmata einen spezifischen Typ moralischer Entscheidungsprobleme dar, bei dem sich miteinander inkompatible Handlungsalternativen gegenüberstehen, die alle aus moralischen Gründen geboten bzw. verboten sind. Aus der intensiven philosophischen Auseinandersetzung mit moralischen Dilemmata im zurückliegenden Jahrhundert ist eine Vielzahl unterschiedlich plausibler Definitionen hervorgegangen.[156] Die gemeinsame Basis aller Ansätze bilden drei Merkmale, die den Kern

156 Eine ausführliche Aufarbeitung unterschiedlicher Perspektiven auf den Dilemmabegriff legt Raters (2013) vor.

des zugrundeliegenden moralischen Konflikts beschreiben. Diese haben sich jedoch im Verlauf des Diskurses als unpräzise erwiesen, denn sie umfassen zu viele Fälle, die bei genauerer Betrachtung keine echten Dilemmata sind. Überdies lassen sie auch hinsichtlich der verwendeten Begrifflichkeiten viele Fragen offen. Im gegenwärtigen moralphilosophischen Diskurs haben Dilemmata zwei wesentliche Kennzeichen: das Vorliegen einer Konfliktsituation einerseits und einen Akteur andererseits, der nicht in der Lage ist zu entscheiden, welche Handlung er wählen soll (vgl. Statman, 1995, S. 6). Fraglich ist bereits an dieser Stelle, ob eine dilemmatische Problemstruktur überhaupt eine ›Entscheidung‹ im engeren Sinne erlaubt – wenn alle Alternativen schlecht sind, kann dann überhaupt eine bewusste, abwägungsbasierte Wahl stattfinden in dem Sinne, dass eine der Alternativen als willentlich ›gewählt‹ gelten kann? Im Folgenden soll der Begriff der ›Entscheidung‹ verstanden werden als Handlung, die eine der möglichen Optionen als Ergebnis eines ethischen Reflexionsprozesses verwirklicht,[157] der diese im spezifischen Fall als bestmögliche Antwort ausweist. Dilemmata stellen somit Situationen dar, in denen der »Protagonist weiß, dass er entscheiden muss, obwohl eine wirkliche Entscheidung nicht möglich zu sein scheint.« (Raters, 2013, S. 57)

Die nachfolgend genannten Merkmale (1)-(3), die in allen ernstzunehmenden wissenschaftlichen Definitionen implizit oder explizit enthalten sind, stellen damit zwar notwendige, aber keine hinreichenden Bedingungen für das Vorliegen *echter* moralischer Dilemmata dar.

(1) *Der Handelnde muss zwischen zwei Handlungsalternativen wählen, die beide moralisch geboten sind.*
(2) *Die Alternativen schließen sich gegenseitig aus.*
(3) *Jede der Alternativen hat moralisch negative Konsequenzen in dem Sinne, dass in jedem Fall ein moralisches Gebot verletzt wird.*

Eine anspruchsvolle, systematische Auseinandersetzung mit den Merkmalen moralischer Dilemmata legt der amerikanische Philo-

157 Dies gilt für menschliche Akteure; in Bezug auf Maschinen ist der Begriff der ›Entscheidung‹ bzw. ›Handlung‹ grundsätzlich metaphorisch zu verstehen, siehe die entsprechende Anmerkung im Vorwort dieses Buches.

soph Walter Sinnott-Armstrong vor. Seine Abhandlung *Moral Dilemmas* (1988) wird im einschlägigen Diskurs oft referenziert, teilweise auch kritisch. Darin entwickelt er eine präzise Definition auf der Basis von Argumenten sowohl der Befürworter als auch der Gegner moralischer Dilemmata. Ins Zentrum seines Definitionsansatzes rückt Sinnott-Armstrong zwei Aspekte: Zum einen wird intensiv erörtert, was es bedeutet, dass etwas *aus moralischen Gründen geboten* ist bzw. dass jemand etwas tun *soll*. Sinnott-Armstrong greift die häufig geäußerte Kritik an der fehlenden Präzision und der Ambivalenz des einschlägigen Sollen-Begriffs (*ought*) auf. Er trägt Ansätze zusammen, die zur Klärung des Terminus existieren, und reflektiert diese im Hinblick auf ihre Glaubhaftigkeit: Was bedeutet es, dass eine Alternative moralisch geboten ist? Welche Voraussetzungen müssen erfüllt sein, damit es Handelnden überhaupt möglich ist, eine der Optionen zu wählen?[158] Zum anderen wird eingehend untersucht, was es heißt, dass ein Argument das andere aufhebt. Es ist offensichtlich, dass eine Abwägung bzw. Hierarchisierung von moralischen Argumenten sich nicht in jedem Fall als möglich erweist. Entweder *können* miteinander in Konflikt stehende Gebote nicht gegeneinander abgewogen werden, weil sie z. B. unterschiedlicher Art sind und so keinem gemeinsamen Bewertungsmaßstab unterliegen. Oder sie *dürfen* nicht gegeneinander abgewogen werden, weil sie beispielsweise die Interessen verschiedener Personen tangieren, deren moralische Ansprüche nicht aufgerechnet werden dürfen.

Es liegt also nahe, dass der oben dargestellten, bisherigen Definition moralischer Dilemmata noch ein entscheidender Punkt fehlt: Es darf keine triviale Abstufung bzw. Abwägung der zugrundeliegenden moralischen Gründe erkennbar sein. Anhand von Sinnott-Armstrongs Gedankengang zur Entwicklung einer differenzierten Definition werden im Folgenden zentrale Merkmale moralischer Dilemmata skizziert. Die Abhandlung hat zum Ziel, ein intuitives Verständnis der spezifischen Problematik zu entwickeln, die morali-

158 Eine Alternative zur sollensbasierten definitorischen Konzeption moralischer Dilemmata, wie sie Sinnott-Armstrong entwickelt, bilden Definitionen auf der Basis konfligierender Handlungsgründe (vgl. z. B. Nagel, 1979a, 1979b). Diese sind allerdings nicht präzise genug (vgl. z. B. Raters, 2013) und daher für die in dieser Arbeit verwendete Definition moralischer Dilemmata von nachrangiger Bedeutung.

sche Dilemmata auszeichnet. Daher erfolgt die Darstellung an dieser Stelle kursorisch, auf einzelne Aspekte sowie kritische Einwände wird in Kap. 5.2 und Kap. 5.3 näher eingegangen.

Im Allgemeinen lässt sich sagen, dass ein Akteur eine Handlung aus moralischer Sicht ausführen soll (*ought*), wenn ein moralischer Grund dafür vorliegt. Ein solcher ist immer dann gegeben, wenn die betreffende Handlung bestimmte Eigenschaften oder Konsequenzen aufweist, die moralisch relevant sind (vgl. Sinnott-Armstrong, 1988, S. 8). Doch was meint moralisch relevant, und wie lassen sich moralische und nicht-moralische Gründe unterscheiden? Für die Beantwortung dieser Fragen existieren verschiedene Antworten, die sich auf unterschiedliche Definitionen von Moralität zurückführen lassen. Trotz zahlreicher Meinungsverschiedenheiten stimmen die meisten Ethiker darin überein, dass jede Handlung, die direkt und absichtlich zum Tod eines Unschuldigen führen oder ihm Schmerz zufügen würde, moralisch relevant ist und ein moralischer Grund dagegen vorliegt. Obwohl es wiederum unterschiedliche Auffassungen darüber gibt, *worin* der moralische Grund genau besteht, so sind sich doch alle einig, *dass* es (mindestens) einen gibt.

Dennoch sind nicht alle Konflikte zwischen moralischen Gründen auch moralische Dilemmata. Es geht auch darum zu klären, ob für eine Handlung tatsächlich oder nur scheinbar moralische Gründe vorliegen bzw. inwiefern diese als gültige Gründe anerkannt werden können. Sinnott-Armstrong (ebd., S. 11) nimmt an dieser Stelle eine heuristische Kategorisierung moralischer Gründe vor, die die kantianische Unterscheidung zwischen vollkommenen und unvollkommenen Pflichten aufgreift, aber konkreter in ihrer Ausführung ist: Er differenziert zwischen Geboten (*requirements*), welche Pflichten, Verpflichtungen und Rechte umfassen, und Idealen (*ideals*). Ein moralisches Ideal ist z. B. das Spenden an eine wohltätige Organisation. Dabei geht es um den Akt des Spendens im Allgemeinen, der im moralischen Sinne idealisiert wird. Ein moralisches Gebot hingegen liegt gemäß Sinnott-Armstrongs Definition vor, wenn es für eine Alternative keine moralische Legitimation (*justification*) für ihr Unterlassen gibt, es also moralisch falsch wäre, sie nicht zu wählen: »A moral reason to adopt an alternative is a moral requirement if and only if it would be morally wrong not to adopt that alternative if there were no moral justification for not adopting it.« (Ebd., S. 12) Ein Beispiel für ein moralisches Gebot wäre das Einhalten eines

Versprechens: Solange keine überzeugende moralische Legitimation vorliegt, ist es moralisch falsch, ein Versprechen zu brechen.[159]

Auf Basis dieser Kategorisierung sind nun drei Arten von moralischen Konflikten denkbar: 1. Konflikte zwischen moralischen Idealen, 2. Konflikte zwischen moralischen Idealen und moralischen Geboten, 3. Konflikte zwischen moralischen Geboten. Der Fokus dieser Arbeit richtet sich auf letzteren Fall, da dieser von Gegnern und Befürwortern moralischer Dilemmata am häufigsten diskutiert wird und auch für die angewandte Fragestellung der Arbeit einschlägig ist. Als revidiertes Kriterium lässt sich damit formulieren:

(1a) Der Handelnde muss zwischen zwei Handlungsalternativen wählen, für die jeweils ein moralisches Gebot vorliegt.

Aus rein struktureller Perspektive muss für das Vorliegen eines Dilemmas als Bedingung gelten, dass sich die gegebenen Alternativen gegenseitig ausschließen, sodass es nicht möglich ist, beide (nacheinander oder gleichzeitig) zu wählen. D. h. es ist eine *echte* Entscheidung im Sinne eines exklusiven Oders erforderlich. Neben diesem relativ offensichtlichen Charakteristikum beinhalten die meisten soliden Definitionen moralischer Dilemmata noch einen weiteren Aspekt, der sich auf die grundsätzliche Erfüllbarkeit der Optionen bezieht. So liegt nur dann ein gültiges moralisches Gebot vor, wenn der Handelnde auch tatsächlich in der Lage ist, es zu erfüllen (siehe Kap. 5.2.3). Dies ist eine notwendige Voraussetzung dafür, dass er jede der beiden Alternativen wählen kann. Auch hier existieren, ähnlich wie bei *ought*, verschiedene Interpretationsansätze zu den Voraussetzungen für dieses ›Können‹ (*can*). Weitgehende Einigkeit herrscht in Bezug auf die Auffassung, dass eine Person genau dann etwas im Sinne von *can* tun kann, wenn sie körperlich oder moralisch dazu in der Lage ist.[160]

---

159 Hier stellt sich die berechtigte Frage, was eine angemessene Legitimation sein kann. Lässt sich eine solche nur aus subjektiver Intuition bestimmen? Hierbei handelt es sich zweifelsohne um ein berechtigtes Problem, das an dieser Stelle nicht gelöst werden kann, im dritten Teil des Buches aber aufgegriffen wird. Im Hinblick auf den Zweck dieses Unterkapitels, der darin besteht, moralische Dilemmata von (verhältnismäßig trivialen) Konflikten abzugrenzen, ist es zunächst von vernachlässigbarer Bedeutung.

160 Ein Beispiel wäre folgende Situation: Angenommen, die Mutter von Person A möchte, dass diese seinem Bruder bei den Hausaufgaben hilft. Zeitgleich

Ein anderes Beispiel zeigt, dass *can* genau dann in dem Sinne erfüllt ist, sodass sich daraus ein moralisches Gebot ableitet, wenn keine anderen moralischen Argumente eines der beiden ursprünglich in Konflikt stehenden Gebote Vorrang haben. Angenommen, Person B hat versprochen, jeweils bei C und D Rasen zu mähen. Demnach hat B ein moralisches Gebot, jeden der beiden Rasen zu mähen. Problematisch wird dies erst, als ein suizidgefährdeter Freund B bittet, zu ihm zu kommen. Folgt B dieser Bitte, schafft er es zuvor nur, den Rasen von C zu mähen. B steht also vor dem Problem, dass er entweder auch noch den Rasen von D mähen oder zu seinem Freund gehen kann. Da Letzteres moralisch gesehen offenbar ein stärkeres moralisches Gebot darstellt, hat es Vorrang gegenüber demjenigen, den Rasen von D zu mähen. Aufgrund dieser Vorrangbeziehung ist B moralisch gesehen nicht mehr in der Lage, den Rasen von D zu mähen. Somit befindet sich B auch nicht in einem echten moralischen Dilemma (vgl. Sinnott-Armstrong, 1988, S. 26–27).[161] Auf der Basis dieses Gedankengangs kann Kriterium 2 präzisiert werden:

(2a) Der Handelnde ist in der Lage, jede der beiden Alternativen zu wählen, aber nicht beide gleichzeitig.

Stehen inkompatible moralische Gebote in Konflikt, liegt die Schlussfolgerung nahe, dass durch eine Entscheidung zwangsläufig eines davon verletzt wird; im Dilemma-Fall wäre dies als unausweichliche, moralisch ungünstige (*unfavorable*), unangenehme (*dis-*

möchte ein Freund, dass A ihn beim Umzug unterstützt. Da A ein gebrochenes Bein hat, ist er körperlich gar nicht in der Lage, beim Umzug zu helfen; deshalb liegt für diese Alternative kein gültiges moralisches Gebot und damit auch kein moralisches Dilemma vor. Andere Aspekte wie verfügbare Zeit usw. werden in diesem Beispiel aus Vereinfachungsgründen nicht berücksichtigt.

161 Umstritten ist hingegen, wie Fälle zu behandeln sind, in denen die körperlichen und moralischen Voraussetzungen erfüllt sind, aber der Handelnde nicht über genug faktisches Wissen verfügt. In diesem Fall spricht man von einem epistemischen Konflikt, wobei es umstritten ist, ob diese Fälle zu Dilemmata zählen (siehe Kap. 5.3.1). Unsicherheit kann z. B. im Hinblick darauf bestehen, ob nicht vielleicht doch eine Kompromisslösung existieren oder die Inkompatibilität der Alternativen anderweitig aufgehoben werden könnte. Zeitdruck und fehlender Einfallsreichtum der Handelnden werden mitunter als weitere Möglichkeiten genannt, die es verhindern, dass jede der beiden Alternativen tatsächlich gewählt werden kann (vgl. Sinnott-Armstrong, 1988, S. 26–28).

agreeable) oder unerwünschte (*undesirable, unwanted*) Konsequenz zu werten. Doch ist dies immer der Fall? Nun sind Konflikte zwischen moralischen Geboten (Typ 3) offensichtlich nicht immer auch moralische Dilemmata, nämlich z. B. dann nicht, wenn ein Gebot sehr viel stärker als das andere ist. In diesen Fällen ließe sich eine Hierarchie bilden, der zufolge dem stärkeren Argument in der konkreten Situation Vorrang vor dem schwächeren gewährt wird. Das bedeutet, dass aufgrund einer Ungleichheit in der Stärke der moralischen Argumente entschieden werden könnte. Derartige Fälle können also keine *echten* moralischen Dilemmata repräsentieren, sondern lediglich moralische Konflikte, die nur scheinbar als Dilemmata auftreten.

Von entscheidender Bedeutung für eine brauchbare Definition ist folglich die Unterscheidung zwischen moralischen Konflikten und Dilemmata. Zentrales Kriterium für die Abgrenzung ist das Verhältnis, in dem die jeweiligen moralischen Gebote der beiden Optionen zueinander stehen. Echte moralische Dilemmata sind dadurch charakterisiert, dass sie sich nicht trivial und eindeutig entscheiden lassen. Welche Anforderungen an das Verhältnis der konfligierenden moralischen Gebote allerdings konkret gestellt werden, wird im Diskurs kontrovers beurteilt. Insbesondere ist umstritten, inwiefern moralische Dilemmata grundsätzlich lösbar sein können und sich dennoch von moralischen Konflikten unterscheiden. So konstituiert Unlösbarkeit im Rahmen von Argumentationsansätzen, welche sich auf moralische Erfahrungen und Empfindungen berufen (siehe Kap. 5.2.2.1), keine notwendige Bedingung für das Vorliegen echter Dilemmata. Hingegen hängen Argumentationen, die die Inkommensurabilität von Werten in den Fokus rücken, entscheidend von der Ansicht ab, dass nur unlösbare Fälle echte Dilemmata darstellen (vgl. Statman, 1990, S. 191–193). Als gemeinsame Basis verschiedener Positionen kann folgende Voraussetzung betrachtet werden: Ein echtes Dilemma liegt dann vor, wenn kein Gebot das andere überwindet (*override*[162]) oder gar aufhebt. Dies ist zunächst einmal dann

162 Im angelsächsischen Diskurs werden hauptsächlich die Begriffe *override* und *defeat* gebraucht. Eine intuitive Übersetzung lautet ›außer Kraft setzen‹, was im Zusammenhang mit moralischen Geboten und Dilemma-Situationen aber gerade *nicht* gemeint ist. Daher wird in diesem Buch primär die Übersetzung ›überwinden‹ im Sinne von ›Vorrang haben‹ verwendet.

erfüllt, wenn die jeweiligen Gebote keine (relevanten) Unterschiede hinsichtlich Stärke und Geltungskraft aufweisen. Allerdings muss hier die Forderung gelten, dass es sich um ein Aufheben im moralisch relevanten Sinne handelt, während es unerheblich ist, ob ein Gebot im außer-moralischen Sinne stärker ist als das andere.[163] Ein moralisches Dilemma kann also keine Gebote einschließen, die von einem anderen in moralisch relevanter Weise überwunden werden. Ist dies erfüllt, so bleiben beide Gebote in ihrer Geltungskraft bestehen, womit unweigerlich eine Verletzung des jeweils nicht befolgten moralischen Gebots einhergeht (siehe Kap. 5.3).

Nach Sinnott-Armstrong (1988, S. 17–18) *dürfen* moralische Dilemmata keine überwundenen und *können* aus logischen Gründen keine überwindenden Gebote enthalten.[164] Damit verbleibt noch als letzte Möglichkeit, dass keines der Gebote Vorrang hat: »If moral requirements conflict, but neither moral requirement overrides the other, then neither is overriding, but also neither is overridden. […] To capture these situations, moral dilemmas can be defined as conflicts between […] non-overridden moral requirements.« (Ebd., S. 18) Demnach sind echte moralische Dilemmata dadurch gekennzeichnet, dass sie nicht-überwundene Gebote (*non-overridden requirements*) in sich bergen.

Eine alternative Auffassung vertritt David Brink (1994, S. 247). Als Kritiker moralischer Dilemmata erachtet er Sinnott-Armstrongs Entwurf als zu schwach, um Konfliktsituationen tatsächlich als Dilemmata ausweisen zu können. Er kritisiert, dass nicht-überwundene Gebote lediglich als nicht-überwundene *Prima-Facie*-Verpflichtungen anzusehen seien. Damit würden sie der von zahlreichen prominenten Befürwortern vertretenen Sichtweise nicht gerecht, dass es sich bei echten Dilemmata tatsächlich um Konflikte zwischen allumfassenden Verpflichtungen (*all-things-considered obligations*) handle.

163 So ließe sich im Beispiel von Styrons Romanfigur Sophie zwar argumentieren, dass der Grund, sich für das jüngere Kind zu entscheiden, schwächer sei, denn es habe ohnehin schlechtere Chancen, die Bedingungen im Konzentrationslager zu überleben. Für Sinnott-Armstrong (1988, S. 54) handelt es sich dabei jedoch um einen moralisch irrelevanten Unterschied.

164 Kritiker moralischer Dilemmata sehen sich durch diese Argumentation darin bestätigt, dass moralische Dilemmata unmöglich sind. Diese Schlussfolgerung basiert allerdings auf einer begrifflichen Ungenauigkeit, die alternative Möglichkeiten der Charakterisierung von Geboten außer Acht lässt.

Jenseits dieser beiden Ansätze dreht sich eine rege Debatte um die Interpretation des Begriffes ›*prima facie*‹ (siehe Kap. 5.3.1), deren Positionen sich in einem Spannungsfeld zwischen begrifflicher Präzision und Irritation bewegen. Für ein intuitives Verständnis der spezifischen Problematik, die moralische Dilemmata auszeichnet, genügt es an dieser Stelle festzuhalten, dass keines der in Konflikt stehenden moralischen Gebote sich durch ein anderes in (zunächst nicht näher spezifizierter) moralisch relevanter Weise überwinden lässt. Kriterium 3 lautet in revidierter Form daher folgendermaßen:

(3a) Keines der in Konflikt stehenden moralischen Gebote lässt sich durch ein anderes in moralisch relevanter Weise aufheben.

Zusammenfassend lässt sich Folgendes konstatieren: Der kontroverse Charakter, welcher den philosophischen Diskurs um moralische Dilemmata kennzeichnet, offenbart, dass moralische Dilemmata vielschichtig sind und einer akkuraten Definition bedürfen. Häufig fehlt es den verwendeten Begriffen an konzeptioneller Tiefe, was sie für Kritik leicht angreifbar macht. Die bislang dargestellten Ausführungen entspringen dem Versuch, häufig genannte Merkmale moralischer Dilemmata kritisch zu hinterfragen und zu analysieren, inwiefern sie als hinreichende Bedingungen taugen. Vor diesem Hintergrund ergibt sich in Summe folgende präzisierte Definition moralischer Dilemmata, die im weiteren Verlauf dieser Arbeit zugrunde gelegt wird:

> *Ein moralisches Dilemma ist eine Situation, in der (1a) ein Akteur zwischen zwei Alternativen wählen muss, für die jeweils ein moralisches Gebot vorliegt, (3a) von denen sich keines in moralisch relevanter Weise durch das andere aufheben lässt und (2a) er zwar jede der beiden Alternativen wählen kann, aber nicht beide gleichzeitig.*[165]

Situationen, die diese Anforderungen erfüllen, sind offenbar komplex und keineswegs alltäglich. Existieren sie in der real-lebenswelt-

165 In der englischen Originalversion lautet die Definition: »A moral dilemma is any situation where at the same time: (1) there is a moral requirement for an agent to adopt each of two alternatives, (2) neither moral requirement is overridden in any morally relevant way, (3) the agent cannot adopt both alternatives together, and (4) the agent can adopt each alternative separately.« (Sinnott-Armstrong, 1988, S. 29).

lichen Praxis überhaupt? Oder sind sie gar theoretisch unmöglich? Diese Fragen sind Gegenstand eines intensiven metaethischen Diskurses, der die Möglichkeit und Realität moralischer Dilemmata seit Mitte des letzten Jahrhunderts facettenreich thematisiert. Prominente Thesen und Positionen dieses Diskurses werden im folgenden Unterkapitel überblicksartig dargestellt, um weitere Besonderheiten moralischer Dilemma-Strukturen zu eruieren.

5.2 Von der (Un-)Möglichkeit und (Nicht-)Existenz moralischer Dilemmata

5.2.1 Überblick und Einführung in den Diskurs

Der Frage nach der Möglichkeit und Existenz moralischer Dilemmata kommt innerhalb der Metaethik eine nicht unwesentliche Bedeutung zu. Zum einen zeigt sie auf, dass moralische Urteile mehr bieten als nur eine einmalige Handlungsorientierung; zum anderen macht sie deutlich, dass von Moraltheorien nicht legitimerweise verlangt werden kann, komplette Entscheidungsprozeduren bereitzustellen. In diesem Sinne fordern uns moralische Dilemmata auf, sowohl Zwecke als auch Grenzen von Moraltheorien und -urteilen neu zu überdenken (vgl. Sinnott-Armstrong, 1988, S. 189). Gelegentlich wird versucht, die Existenz moralischer Dilemmata über das Vorliegen praktischer Beispiele zu beweisen. Ein solches Vorgehen, das sich rein auf die statistische Häufigkeit des Auftretens gründet, ist jedoch wenig überzeugend (vgl. Holbo, 2002, S. 259). Nur weil jemand in einer konkreten Situation nicht *weiß*, was er tun soll, folgt daraus noch nicht, dass es grundsätzlich keine richtige Entscheidung *gibt* (vgl. McConnell, 2022).

Vielmehr hängen Möglichkeit und Existenz moralischer Dilemmata entscheidend davon ab, welche Annahmen über die Natur der Moral getroffen werden. Dies trifft den Kern der Metaethik, die das Fundament, die Grundprinzipien und die Kohärenz ethischer Theorien und Urteile untersucht sowie über formale und semantische Analysen zu Aussagen und Hypothesen über das Wesen moralischer Urteile und Normen gelangt. In der Philosophiegeschichte sind Möglichkeit bzw. Existenz moralischer Dilemmata seit jeher Gegenstand kontroverser Dispute. Angesichts der Problematik kolli-

dierender moralischer Gebote wird eine Grundsatzdebatte um die (universelle) Gültigkeit von Moraltheorien als solche entfacht. Lange war in der Moralphilosophie die Ansicht vorherrschend, dass eine gültige Moraltheorie keine echten moralischen Dilemmata zulassen, sondern in jeder denkbaren Situation eine moralisch eindeutige und konkrete Handlungsorientierung bieten sollte. Damit geht das Postulat einher, dass echte moralische Dilemmata nicht auftreten können. Diese Position prägte eine Tradition der Moralphilosophie, die – vertreten von prominenten Ethikern wie Thomas von Aquin (1916), Kant (1900ff., insb. GMS, AA 04 & MS, AA 06), Mill (1861) und Ross (1930, 1939) – bis heute fortbesteht.

Seit Mitte des 20. Jahrhunderts äußerten moderne Philosophen jedoch vermehrte Zweifel bezüglich der hohen Ansprüche an universelle Moraltheorien: »We might not know which situations are moral dilemmas, until we know which substantive moral theory is true, but we can still have strong reasons to believe that some situations are moral dilemmas.« (Sinnott-Armstrong, 1988, S. 35) Zu den ersten, die eine grundlegende Skepsis gegenüber der Zurückweisung der Möglichkeit moralischer Dilemmata äußerten, gehört Jean-Paul Sartre;[166] vor allem deontische Logiker sind gefolgt. Seitdem hat besonders die auf verschiedenen Argumentationsebenen geführte angelsächsische Debatte Fahrt aufgenommen, die ihren Höhepunkt in den 1960er- bis 1980er-Jahren erreichte und bis heute fortgeführt wird. Die vorgebrachten Argumente sind sowohl konzeptioneller als auch phänomenologischer Art; sie lassen sich in drei Argumentationslinien untergliedern, die von unterschiedlichen moraltheoretischen Positionen und deren Anwendung auf (scheinbare) Dilemmata ausgehen.[167] Im Rahmen dieser Forschungsarbeit wird der

166 Sartres Beispiel eines Mannes, der sich zwischen der Aufnahme des Militärdienstes und der Pflege seiner Mutter entscheiden muss (siehe Kap. 5.1.1), zählt zu den bekanntesten Beispielen moralischer Dilemmata in der neueren Literatur (1946, S. 156–157).

167 Die Darstellung in dieser Arbeit folgt der Kategorisierung relevanter Argumente nach Gowans (1987). Einen alternativen Versuch, die wesentlichen Positionen des Diskurses zusammenzufassen, macht Statman (1990, S. 191–193). Er sagt, dass alle prominenten Argumente einer von zwei Argumentationslinien folgen: Sowohl das *Argument der Pluralität von (echten moralischen) Werten* als auch das *Argument der Einzelwert-Konflikte* zeigen die Realität moralischer Dilemmata nur indirekt über das Konzept der Unlösbarkeit. Einen direkten

Diskurs nur überblicksartig skizziert. Die Darstellung in den nachfolgenden beiden Unterkapiteln erhebt daher keinen Anspruch auf Vollständigkeit und fokussiert sich lediglich auf diejenigen Aspekte, die für den Anwendungskontext des autonomen Fahrens und die Zielsetzung dieses fünften Kapitels relevant sind.[168]

5.2.2 Phänomenologische und konzeptionelle Perspektiven

5.2.2.1 Phänomenologischer Ansatz: Das Argument des moralischen Empfindens

Theoretiker, die die Realität moralischer Dilemmata negieren, gehen davon aus, dass man die Problematik allein auf der Grundlage konzeptioneller Überlegungen aus dem Bereich deontischer Logik oder rationaler Abwägungsprozesse entscheiden kann. Jegliche Bezugnahme auf moralische Emotionen halten sie für irrational. Befürworter moralischer Dilemmata betonen dagegen häufig die Bedeutung moralischer Erfahrungen, woraus sich das *Argument des moralischen Empfindens* (*argument from moral sentiment*) entwickelt hat. Die Auseinandersetzung mit der Möglichkeit moralischer Dilemmata wird daher auch als Diskurs zwischen Rationalisten und Experientialisten bezeichnet. Das Kernargument der Letzteren begründet die Möglichkeit und Existenz moralischer Dilemmata über die Erklärung und Bewertung von Rolle und Ursprung moralischer Gefühle wie Bedauern, Reue oder Schuld. Dilemmata können demnach bei der praktischen Entscheidung moralischer Konflikte entstehen, wenn wir uns für eine Option entscheiden, zugleich aber die Nicht-

Beweis hingegen versucht das *Argument des moralischen Empfindens* vorzulegen. Dabei wird die Frage nach der Realität moralischer Dilemmata als identisch angesehen mit derjenigen nach der Realität überwundener Pflichten. Raters (2013) präsentiert ebenfalls eine alternative Darstellung des Diskurses, indem sie sich der vielschichtigen Dilemmaproblematik über verschiedene Zugänge mit spezifischen Schwerpunkten aus Sicht prominenter metaethischer Positionen anzunähern versucht.

168 Ausführliche und präzise Darstellungen des Diskurses finden sich in Gowans' Einleitung zu dem von ihm herausgegebenen Sammelband (1987, S. 5–14) sowie in der umfangreichen Monografie von Raters (2013).

berücksichtigung der anderen bedauern oder uns gar schuldig fühlen.[169]

Systematisch entwickelt wird das *Argument des moralischen Empfindens* zum ersten Mal bei Bernard Williams (1987). Er versucht zu beweisen, dass moralische Konflikte zwischen Wünschen (*desires*) und nicht zwischen faktischen Überzeugungen (*beliefs*) bestehen. Diese Sichtweise hat eine bedeutende Implikation: Wenn zwei moralische Gebote – verstanden als Wünsche – in Konflikt stehen und wir nach einem davon handeln, wird der andere nicht zwangsläufig eliminiert, sondern bleibt weiterhin existent:

> [...] it is surely falsifying of moral thought to represent its logic as demanding that in a conflict situation one of the conflicting ought's must be totally rejected. One must, certainly, be rejected in the sense that not both can be acted upon; [...] But this does not mean they do not both (actually) apply to the situation; or that I was in some way mistaken in thinking that these were both things that I ought to do. (Ebd., S. 134)

Seine These sieht Williams durch die Tatsache untermauert, dass wir die Nichterfüllung bedauern, auch wenn wir glauben, das Bestmögliche getan zu haben; auf diese Weise entsteht ein ›moralischer Restwert‹ (*moral remainder*) bzw. ›moralischer Rückstand‹ (*moral residue*). Terrance McConnell (1996, S. 37–38) formuliert vier Annahmen, die für das Vorliegen eines solchen moralischen Restwerts erfüllt sein müssen: (1) der Akteur empfindet Bedauern, Gewissensbisse oder Schuld, wenn er handelt; (2) das Empfinden dieser Emotionen ist angemessen und geboten; (3) der Akteur hätte ebenfalls so gefühlt, wenn er nach der anderen der konfligierenden Anforderungen gehandelt hätte; (4) diese Emotionen wären ebenso angemessen und geboten gewesen.

In Antizipation möglicher Einwände präzisiert Williams (1987, S. 131), dass Gefühle des Bedauerns im Allgemeinen zwar als entweder irrational oder unmoralisch angesehen werden, es aber durchaus Fälle geben kann, in denen Handelnde vernünftigerweise ihre Entscheidung bedauern. So bedauert Agamemnon die Opferung seiner Tochter, weil er weiß, dass er sie nicht hätte opfern dürfen, auch wenn dies in seiner Situation im moralischen Sinne das Beste war,

169 In der Folge werden die prominentesten Argumente der traditionellen Debatte beschrieben. Für einen neueren Beitrag siehe Tessman (2015, Kap. 2).

was er tun konnte. Agamemnons Bedauern ist berechtigt und keinesfalls irrational, denn er glaubt, dass das, was er tun musste, an sich etwas Schlechtes war, unabhängig von möglichen Alternativen. Williams weist in diesem Zuge die Behauptung zurück, Bedauern sei Ausdruck dessen, dass der Handelnde erkennt, dass er falsch gehandelt habe. Vielmehr sei ein Gefühl des Bedauerns durchaus konsistent mit dem inneren Glauben, man hätte die Option wählen sollen, die man nicht gewählt hat, während man zugleich aber glaubt, dass man das Bestmögliche unter den gegebenen Umständen getan hat. Roger Trigg (1971) führt aus, dass ein Gefühl des Bedauerns an spezifische Umstände der dilemmatischen Situation geknüpft ist:

> Regretting having to do something involves wishing I did not have to. I would not be doing it if I had a completely free choice. I am reluctant to do it. When I have done it, I would not be pleased with what I have done even if I recognised that I had done the best thing in the circumstances. My whole attitude to what I had to do both before and after would be different from that of the man who could see no wrong in it. He would do it readily. I would be more reluctant and would make quite sure that this was indeed the lesser of the two evils facing me. He would be willing to do a similar action in the future. I would not, unless I was again presented with the same kind of dilemma. (Ebd., S. 49)

Ein weiteres bekanntes Argument in diesem Kontext wird von Ruth Marcus (1980) vorgetragen, die moralisch motivierte Schuldgefühle in den Fokus rückt. Sie erläutert, dass es eine kausale Verbindung zwischen dem Gefühl des Bedauerns und der Tatsache gibt, dass die nicht gewählte Handlung mit einer moralischen Pflicht belegt ist. Dilemmata als Situationen, in denen moralische Pflichten kollidieren, können also Bedauern rechtfertigen. Nach Marcus würde eine Verneinung moralischer Dilemmata implizieren, dass Schuldgefühle hinsichtlich der nicht gewählten Option unangemessen und irrational seien. Bas van Fraassen (1973, S. 13) fügt hinzu, dass es nur dann angemessen ist, sich schlecht zu fühlen, wenn man sich tatsächlich schuldig gemacht hat. Ein in diesem Sinne berechtigtes Schuldgefühl entsteht nur dann, wenn man eine moralische Pflicht verletzt hat. Beide Ansätze sehen angesichts derart in Konflikt stehender Pflichten die Existenz moralischer Dilemmata als bewiesen an.

Ansätze, die die Existenz moralischer Dilemmata über das *Argument des moralischen Empfindens* zu begründen versuchen, sind im Allgemeinen sehr umstritten. Zentrale Gegenargumente stützen

sich darauf, dass es andere Erklärungen von Bedauern gibt, die unabhängig von der Forderung existierender Dilemmata sind. Zwei Haupteinwände prägen den Diskurs.[170] Der erste stellt das gegebene Argument in seiner Plausibilität in Frage. So argumentiert Earl Conee (1982, S. 91–97), dass moralische Gefühle nicht unbedingt mit der Verletzung einer faktisch vorliegenden moralischen Pflicht zusammenhängen, sondern lediglich mit dem subjektiven Moralverständnis des Handelnden, dem gemäß er eine bestimmte Handlung hätte ausführen sollen:[171]

> Feeling guilty is subjectively appropriate when the belief that one has failed which prompts the feeling fits one's moral principles. […] When someone does what is morally best while neglecting something his morality requires, his feeling guilty is therefore appropriate only because it is called for by morality as he sees it. […] This sort of appropriate guilt does not imply that a moral mistake has been made. So an opponent of moral dilemmas can consistently hold that feeling guilty about a morally superior act is clearly appropriate at times–but in this subjective way, not in light of any omitted actual moral obligation. (Ebd., S. 91–92)

Grundsätzlich ist ein Gefühl des Bedauerns auch möglich, obwohl die eigene Handlung moralisch richtig ist, z. B. wenn man ein Kind bestraft. Der Unterschied zwischen Bedauern und Reue bzw. Schuld besteht darin, dass zwar in allen Fällen ein negatives Gefühl vorhanden ist, aber nur in Letzterem eine kognitive Komponente hinzukommt – die Überzeugung, etwas falsch gemacht zu haben. Diese kann jedoch nur im Fall eines Dilemmas vorhanden sein. Das Vorliegen eines Dilemmas ist demnach lediglich eine Annahme und keine Schlussfolgerung, wie sie im Zuge des *Arguments des moralischen Empfindens* gezogen wird. Welches moralische Gefühl in einem konkreten Fall vorliegt, muss in anderer Weise spezifiziert werden, als es das gegebene Argument leistet; die Berufung auf moralische Rückstände allein kann unter dieser Perspektive noch nicht die Realität moralischer Dilemmata begründen.

Der zweite Einwand fordert die Annahme heraus, dass Reue und Schuldgefühle nur angemessen sind, wenn der Handelnde fühlt, dass er etwas falsch gemacht hat. So kann man sagen, dass es in

170 Weitere, hier nicht im Einzelnen dargestellte einschlägige Argumente finden sich u. a. bei Richard M. Hare (1987) und Philippa Foot (1987).

171 Ein verwandtes Argument präsentiert McConnell (1978, S. 277–282).

jedem Fall Teil unserer moralischen Verantwortung ist, sich für negative Konsequenzen zu entschuldigen, nicht nur wenn man meint, etwas falsch gemacht zu haben. Auch wenn die Konsequenzen einer Handlung so gravierend sind, dass man sich unabhängig von moralischen Aspekten auf jeden Fall schuldig fühlt, etwa wenn man jemanden unabsichtlich überfährt, kann dies zutreffen (vgl. McConnell, 1996, S. 42–44).

Als Resümee lässt sich festhalten: Das Empfinden negativer moralischer Emotionen angesichts eines Konflikts zwischen moralischen Geboten können sowohl Befürworter als auch Gegner moralischer Dilemmata erklären. Der Zusammenhang zwischen ethischen Konflikten und moralischen Emotionen ist sehr komplex; nur tiefergehende Analysen können diesem gerecht werden.[172]

5.2.2.2 Konzeptionelle Ansätze I: Das Argument der Pluralität von (echten moralischen) Werten

Eine aus konzeptioneller Sicht kritische Haltung gegenüber der Möglichkeit moralischer Dilemmata gründet sich im Wesentlichen auf Ansprüche an die universelle Gültigkeit von Moraltheorien. Befürworter moralischer Dilemmata sind hingegen der Ansicht, dass kein plausibler Grund für die Forderung vorliegt, dass eine gültige Moraltheorie Dilemmata kategorisch ausschließen muss. Zentrale einschlägige Argumentationslinien verstehen sich als Antworten auf die Argumente, welche von Seiten der Kritiker als vermeintliche Belege für die Unmöglichkeit moralischer Dilemmata vorgebracht werden.

Einige wenige Kritiker verfolgen einen Argumentationsansatz, demzufolge nicht nur moralische Dilemmata zu verneinen sind, sondern gar jegliche Form moralischer Konflikte an sich (vgl. Sinnott-Armstrong, 1988, S. 31–32). Ihr Kernargument bezieht sich dabei auf die metaphysische Struktur von Moralität: Auf der Grundlage einer monistischen Position wird erläutert, dass moralische Gebote immer nur vermeintlich in Konflikt stehen, weil niemals alle zugleich den höchsten Wert maximieren können, auf den sich alle gemeinsam

172 Entsprechende Untersuchungen legen z. B. Greenspan (1995) und Tessman (2015) vor.

reduzieren lassen. Diese Auffassung wird vor allem von Utilitaristen vertreten, die die Gültigkeit des Nutzenprinzips als höchsten Wert jeglicher moralischen Bewertung zugrunde legen.[173] In der zeitgenössischen Debatte wird die utilitaristische Sichtweise vor allem von Hare (1987) repräsentiert. Hingegen führen kantianisch geprägte Kritiker wie Conee (1982) und Donagan (1984) auf der Grundlage eines pluralistischen Standpunkts relevante Gegenargumente an, die sich auf Merkmale und Gültigkeit moralischer Gebote beziehen.

Die überwiegende Mehrheit der Dilemma-Kritiker erkennt an, dass moralische Konflikte prinzipiell möglich sind.[174] Moralische Pflichten sind divers und unsere Netzwerke von Beziehungen und sozialen Verpflichtungen komplex, sodass es naheliegt, dass diese von Zeit zu Zeit in Konflikt geraten; daraus folgt jedoch noch nicht zwangsläufig, dass Dilemmata existieren (vgl. McConnell, 1978, S. 279). Das *Argument der Pluralität von (echten moralischen) Werten* (*argument from a plurality of (genuine moral) values*) knüpft an die Vorstellung der universellen Gültigkeit von Moraltheorien an, interpretiert diese jedoch nicht ganz so streng. Kritische Positionen nehmen an, dass eine Moraltheorie, die echte Dilemmata zulässt, inkompatible Handlungsempfehlungen enthält, da sie in Dilemma-Situationen keine eindeutige Handlungsorientierung zu geben vermag. Deshalb müsse ihr die Fähigkeit abgesprochen werden, in allen denkbaren Fällen handlungsleitend zu sein, was wiederum ihren universellen Geltungsanspruch untergraben würde (vgl. McConnell, 2022). In der Konsequenz seien moralische Dilemmata auszuschließen.[175]

---

173 Für eine kritische Auseinandersetzung mit dem utilitaristischen Argument siehe Sinnott-Armstrong (1988, S. 74–81).

174 Eine Zwischenposition geht davon aus, dass moralische Dilemmata zwar möglich sind, aber nicht wirklich auftreten. Da es zahlreiche Evidenz für dilemmatische Situationen in real-lebensweltlichen Zusammenhängen gibt (vgl. Sinnott-Armstrong, 1988, S. 31), erscheint diese Position tendenziell implausibel und wird daher an dieser Stelle nicht weiter ausgeführt.

175 McConnell (1978) liefert ein weiteres Argument, indem er zwei verschiedene Arten von Inkonsistenz vorschlägt, die bei echten moralischen Dilemmata auftreten können: einerseits als Inkonsistenz in unserem ethischen Denken, andererseits als Forderung, dass eine angemessene Theorie solche Fälle ausschließt, um Inkohärenzen zu vermeiden. Ihm zufolge gibt es gute Gründe anzunehmen, dass eine gültige Moraltheorie zwangsläufig moralische Dilemmata ausschließen muss.

Einen alternativen Ansatz zur traditionellen Perspektive auf moralische Dilemmata, welche diese als Entscheidungsproblem zwischen Alternativen auffasst, legt Conee (1982) vor. Er betrachtet moralische Dilemmata nicht als Entscheidung zwischen zwei (oder mehr) Werten, sondern als Gelegenheit, kreative Ansätze zu entwickeln, die eine Koexistenz beider Werte erlaubt, sodass Dilemmata konsequenterweise zu verneinen sind. Solche Entscheidungsstrategien sind dabei immer individuell und tragen einem Wertepluralismus Rechnung, wobei sie über traditionelle ethische Frameworks hinausgehen.

Positionen, die die Existenz moralischer Dilemmata hingegen befürworten, werden u. a. von deontischen Logikern vertreten (siehe Kap. 5.2.3). In einer der ersten systematischen metaethischen Auseinandersetzungen präzisiert Edward J. Lemmon (1962) sein Verständnis eines moralischen Sollen-Begriffs im Sinne der Begründbarkeit eines moralischen Gebots aus (mindestens) einer von drei Quellen: Pflichten (*duties*), eingegangene Verpflichtungen (*obligations*)[176] und moralische Prinzipien (*moral principles*). In moralischen Dilemmata stehen sich schließlich diese verschiedenen Quellen gegenüber.[177] Als kritische Antwort auf Lemmons Konzept argumentiert Conee (1982, S. 88–89), dass Pflichten oder Verpflichtungen nicht notwendigerweise eine moralische Dimension besitzen, z. B. dann nicht, wenn die entsprechenden Handlungen an sich unmoralisch sind, wie z. B. bei beruflichem Töten. Auf ihrer Basis allein lasse sich also kaum die Existenz moralischer Dilemmata rechtfertigen.

Eine andere Version des pluralistischen Arguments entwickelt Thomas Nagel (1979b). Er erörtert Probleme praktischer Konflikte, die aus der Diskrepanz zwischen einer Zersplitterung der Werte einerseits und der Einheitlichkeit moralischer Entscheidungen andererseits resultieren. Dabei geht er von einer Unterscheidung von

176 Pflichten beziehen sich auf eine bestimmte Position, einen Status oder eine Rolle innerhalb der Gesellschaft, während sich Verpflichtungen aus einer vorherigen Handlung ergeben, z. B. aufgrund eines gegebenen Versprechens, eines unterschriebenen Vertrags etc. (vgl. Lemmon, 1962, S. 140–142).

177 Mögliche Entscheidungsstrategien lägen dann beispielsweise in der Bildung von Hierarchien hinsichtlich dieser Quellen, z. B. von moralischen Prinzipien über Pflichten über Verpflichtungen, oder in einer utilitaristischen Orientierung an den jeweiligen Konsequenzen der konfligierenden Optionen (vgl. Lemmon, 1962, S. 148–153).

Werten in fünf verschiedene Typen aus, die mit- und untereinander in Konflikt stehen können: spezifische Verpflichtungen gegenüber Personen oder Institutionen, allgemeine Rechte, Nützlichkeitserwägungen (»the consideration that takes into account the effects of what one does on everyone's welfare«), perfektionistische Ziele im Sinne des intrinsischen Werts von Errungenschaften, z. B. Kunstwerken (»the intrinsic value of certain achievements or creations, apart from their value to individuals who experience or use them«) und persönliches Engagement für eigene Projekte (vgl. ebd., S. 175–177). Vor diesem Hintergrund muss nach Nagel der Anspruch der Moralphilosophie auf universelle Gültigkeit und Konsistenz aufgegeben werden, nicht aber die Moralphilosophie als systematische Disziplin selbst. Im Gegenteil: Diese wird »gerade dadurch unverzichtbar, dass individuelle moralische Akteure mit der Entscheidung moralischer Dilemmata […] überfordert sind.« (Raters, 2013, S. 343). Das Scheitern moralphilosophischer Strategien hat seine Ursache darin, dass Werte unterschiedlichen Ursprungs sind. Nagel unterscheidet dabei zwischen personalen und impersonalen Handlungsgründen. Diese beiden Kategorien sind so heterogen und formal unterschiedlich, dass es schwierig ist, verschiedene Erwägungen in einem einzigen bewertenden Urteil zusammenzubringen; ergo gibt es kein einheitliches Vorgehen zur Entscheidung echter moralischer Dilemmata. Der wesentliche Kern von Nagels Argument liegt darin, dass personale und impersonale Handlungsgründe unüberbrückbar sind, weil sie unterschiedliche Sichtweisen auf die Welt konstituieren:[178]

> My general point is that the formal differences among types of reason reflect differences of a fundamental nature in their sources, and that this rules out a certain kind of solution to conflicts among these types. Human beings are subject to moral and other motivational claims of very different kinds. This is because they are complex creatures who can view the world from many perspectives […] But when conflict occurs between them, the problem is still more difficult. […] The capacity to view the world simultaneously from the point of view of one's relations to others, from the point of view of one's life extended through time, from the point of view of everyone at once, and finally from the

178 Raters (2013, S. 319–332) entwickelt im Rahmen einer kritischen Auseinandersetzung mit Nagels Perspektive einige weiterführende Fragen, die es erlauben, die Grundsatzfrage nach der Möglichkeit moralischer Dilemmata in einen größeren Zusammenhang zu stellen.

> detached viewpoint often described as the view of sub specie aeternitatis is one of the marks of humanity. This complex capacity is an obstacle to simplification. (Nagel, 1979b, S. 180)

Raters (2013, S. 330) spricht in diesem Zusammenhang von einem Abgrund, »der sich als Wesensmerkmal unserer Vernunft durch keine philosophische Strategie ›beseitigen‹ […] lässt.«

5.2.2.3 Konzeptionelle Ansätze II: Das Argument der Einzelwert-Konflikte

Eine weitere, konzeptionelle Argumentationslinie knüpft an die Anspruchshaltung bezüglich der universalen Gültigkeit von Moraltheorien an. Mögliche Entscheidungsstrategien für Situationen, in denen konkrete Prinzipien einer Moraltheorie miteinander in Konflikt geraten, sind zwar im Einzelfall denkbar, lassen sich aber nicht verallgemeinern. Insbesondere bleibt offen, wie Situationen zu behandeln sind, in denen der dilemmatische Konflikt durch ein und denselben moralischen Wert begründet ist. Hier kommt das sogenannte *Argument der Einzelwert-Konflikte* (*argument from single-value conflicts*) zum Tragen. In ihrem Aufsatz entfaltet Marcus (1980) den Gedankengang, dass ein einziges moralisches Prinzip unter bestimmten Bedingungen mit sich selbst in Konflikt stehen kann. Ihr Ziel ist es nachzuweisen, dass es keinen Grund für die Annahme gibt, dass sich Konflikte völlig vermeiden lassen; keiner moralischen Theorie sei es möglich, echte moralische Dilemmata grundsätzlich auszuschließen. Vielmehr können moralische Konflikte selbst dann auftreten, wenn pluralistische Positionen falsch sind und es nur ein einziges (höchstes) Prinzip gibt, beispielsweise in dem Sinne, wie es Utilitaristen behaupten.

Sinnott-Armstrong (1988, S. 54–58) hingegen beurteilt symmetrische Fälle als unausweichlich dilemmatisch. Zur Illustration seines Gedankengangs zieht er ein häufig zitiertes Beispiel einer symmetrischen Konstellation moralischer Gründe heran, wie sie in *Sophie's Choice* (1980) thematisiert wird. In diesem Beispielfall kann kein

moralisches Gebot das andere überwinden, weil es keinen moralisch relevanten Unterschied zwischen den betroffenen Kindern gibt.[179]

Die von Dilemma-Kritikern angeführten Einwände gegen die Möglichkeit symmetrischer moralischer Konflikte überschneiden sich teilweise mit jenen, die gegen pluralistische Wertekonflikte vorgebracht wurden. So ist hier das Argument von Hare (1987) einschlägig, dass auch zwei symmetrische Gründe niemals beide einen höchsten Wert maximieren können. Eine differenzierte, prominente Antwort auf Marcus' Begründung symmetrischer Konflikte stammt von Alan Donagan (1984). Er kritisiert, dass praktische Konflikte fälschlicherweise allzu häufig für moralische gehalten werden. Die Frage, was man in Konfliktsituationen tun soll, ist nur dann eine moralische Frage, wenn man annimmt, dass moralische Restriktionen hinsichtlich dessen gelten, was man tun soll. Die Symmetrie einer Situation impliziert jedoch, dass es für keine der Optionen Gründe gibt, weder moralische noch außermoralische. Das Fehlen moralischer Gründe wiederum bedeutet, dass z. B. die Frage, welchen von zwei sich in einer Notlage befindenden Zwillingen man retten soll, keine moralische ist, und die einzig mögliche rationale Antwort lautet, dass es unerheblich ist, für *welche* der Alternativen man sich entscheidet, solange man *eine* davon wählt. Die moralischen Verpflichtungen sind in diesem Fall disjunkt, d. h. es soll eine der Möglichkeiten gewählt werden – welche, bleibt dem Akteur überlassen:

> Where the lives of identical twins are in jeopardy and I can save one but only one, every serious rationalist moral system lays down that, whatever I do, I must save one of them. By postulating that the situation is symmetrical, Marcus herself implies that there are no grounds, moral or nonmoral, for saving either as opposed to the other. Why, then, does she not see that, as a practical question, Which am I to save? has no rational answer except ›It does not matter,' and as a moral question none except ›There is no moral question‹? Certainly there is no moral conflict: from the fact that I have a duty to save either a or b, it does not follow that I have a duty to save a and a duty to save b. Can it be

179 Tatsächlich geht Styron selbst in seinem Roman davon aus, dass das Alter der Kinder und damit deren Überlebenschancen im Konzentrationslager durchaus einen moralisch relevanten Unterschied begründen. Sinnott-Armstrong allerdings modifiziert das Beispiel in einer Weise, sodass kein Unterschied angenommen werden kann.

> seriously held that a fireman, who has rescued as many as he possibly could of a group trapped in a burning building, should blame himself for the deaths of those left behind, whose lives could have been saved only if he had not rescued some of those he did? (Ebd., S. 307)

Donagan versucht, Marcus' These, dass auch ein moralisches System mit nur einem einzigen Prinzip zu Konflikten führen könne, unter Bezugnahme auf Kants Unterscheidung zwischen vollkommenen und unvollkommenen Pflichten zu widerlegen. Dabei stützt er sich auf dessen These, dass sich vollkommene Pflichten grundsätzlich nicht widersprechen können und Gründe der Verbindlichkeit im Konfliktfall keine bindenden Verpflichtungsgründe mehr darstellen.[180] Donagan führt weiterhin ins Feld, dass Prinzipien immer implizite Bedingungen enthalten, mittels derer sich mögliche Konflikte entscheiden lassen. Insbesondere im Hinblick auf symmetrische Konflikte erläutert er, dass ein Versprechen nur dann moralische Gültigkeit besitzt, wenn der Akteur keinen Konflikt antizipieren konnte. Wird dennoch ein zweites Versprechen gegeben, das dann mit dem ersten in Konflikt gerät, so ist dies nicht auf ein Versagen des moralischen Systems zurückzuführen, sondern auf persönliches moralisches Fehlverhalten im Sinne eines Abweichens von einer gültigen moralischen Norm, durch das wider die Moraltheorie gehandelt wird. Dieser Gedankengang ist auch als *Argument der vor-*

180 Im Kant'schen Original lautet die entsprechende Textstelle: »Ein Widerstreit der Pflichten (*collisio officiorum s. obligationum*) würde das Verhältnis derselben sein, durch welches eine derselben die andere (ganz oder zum Theil) aufhöbe. – Da aber Pflicht und Verbindlichkeit überhaupt Begriffe sind, welche die objective praktische Nothwendigkeit gewisser Handlungen ausdrücken, und zwei einander entgegengesetzte Regeln nicht zugleich nothwendig sein können, sondern wenn nach einer derselben zu handeln es Pflicht ist, so ist nach der entgegengesetzten zu handeln nicht allein keine Pflicht, sondern sogar pflichtwidrig: so ist eine Collision von Pflichten und Verbindlichkeiten gar nicht denkbar (*obligationes non colliduntur*). Es können aber gar wohl zwei Gründe der Verbindlichkeit (*rationes obligandi*), deren einer aber oder der andere zur Verpflichtung nicht zureichend ist (*rationes obligandi non obligantes*), in einem Subject und der Regel, die es sich vorschreibt, verbunden sein, da dann der eine nicht Pflicht ist. – Wenn zwei solcher Gründe einander widerstreiten, so sagt die praktische Philosophie nicht: daß die stärkere Verbindlichkeit die Oberhand behalte (*fortior obligatio vincit*), sondern der stärkere Verpflichtungsgrund behält den Platz (*fortior obligandi ratio vincit*).« (Kant, 1900ff., MS, AA 06: 224.09-26, Hervorh. i. Orig.). Kants *Metaphysik der Sitten* wurde erstmals 1797 veröffentlicht.

angegangenen moralischen Fehlleistung bekannt und zeigt auf, dass aus der Existenz moralischer Dilemmata noch nicht die Inkonsistenz der Moralphilosophie zu folgern ist (vgl. ebd., S. 303).

Nachdem Donagan zu dem Schluss gekommen ist, dass moralische Dilemmata möglich sind, entwickelt er ein kasuistisches Verfahren, mit dessen Hilfe sich diese zuverlässig entscheiden lassen. Dabei soll nach dem ›Geist‹ konkreter Prinzipien im jeweiligen Fall entschieden werden, nicht nach deren Wortlaut, was sich in der Praxis allerdings häufig als unklares und daher unzuverlässiges Vorgehen herausstellt (vgl. Raters, 2013, S. 361–362). Im Rahmen von Marcus' Auseinandersetzung und verwandten Positionen sei die rationalistische Position, dass moralische Verpflichtungen niemals kollidieren, nur vermeintlich falsifiziert worden:

> [...] the rationalist position that moral obligations never collide has not been shown to be false and [...] the prevalent impression that it has been springs from three sources: confusion of practical conflict generally with moral conflict; overlooking the distinction between moral conflict *simpliciter* and moral conflict *secundum quid*; and neglect of the casuistical resources of the various rationalist ethical traditions. (Donagan, 1984, S. 309, Hervorh. i. Orig.)

5.2.3 (Vermeintliche) Inkonsistenzen in Theoriesystemen: Argumente der deontischen Logik und Thesen logischer Widersprüchlichkeit

Eine weitere, von moraltheoretischen Positionen unabhängige Klasse von Argumenten bewegt sich auf der Ebene von Annahmen über Wahrheit und Konsistenz von Prinzipien der deontischen Logik. Entsprechende kritische Positionen diskutieren auf konzeptioneller Ebene, dass die Existenz moralischer Dilemmata als Indikator für das Vorliegen von Inkonsistenzen innerhalb von Moraltheorien oder anderen ihr zugrundeliegenden Prämissen sowie logischen Prinzipien zu werten sei. Daher sei die (universelle) Gültigkeit zweier Standardaxiome der deontischen Logik grundlegend unvereinbar mit der These, dass moralische Dilemmata möglich sind. Kann ein Akteur nicht beide Optionen A und B zugleich wählen, so folgt daraus, dass das Negieren von A zugleich Bedingung von B ist. Da es jedoch eine notwendige Bedingung moralischer Dilemmata ist, dass beide Optionen moralisch geboten sind, hieße das, dass Option A zugleich geboten und verboten (als Bedingung von B) sein

müsse, was dem Grundsatz der deontischen Konsistenz (*principle of deontic consistency*) widerspricht. Dieses Prinzip wird auch von Dilemma-Befürwortern als grundlegend und notwendig angesehen und daher kaum angezweifelt (vgl. Lemmon, 1965, S. 51).

Dagegen sind Diskussionen um das Prinzip der deontischen Logik (*principle of deontic logic*) eher gängig, welches fordert, dass, wenn B aus A folgt und A moralisch geboten, dann auch B moralisch geboten ist. In Bezug auf beide Prinzipien sind jedoch rein konzeptionelle Widerlegungsversuche wenig erfolgversprechend. Komplexere Antworten stellen dagegen den Geltungsbereich der Prinzipien in Frage. So argumentiert Holbo (2002, S. 267–270), dass die wesentlichen deontischen Prinzipien nur in einer idealen Welt uneingeschränkt gelten. In der realen Welt sind sie lediglich als heuristische Bedingungen zu verstehen, wobei sie im Fall von Konflikten die Akteure im pragmatischen Sinne anleiten, nach zulässigen Optionen zu suchen, auch wenn es schlussendlich keine gibt:[181]

> These heuristics will function as instructions for blueprinting deontically ideal possible worlds. If a contradiction emerges in the construction process, that means the (tacit) assumption that the actual situation is not hopeless must be false. The unavoidability of obligation violations is proved. (Ebd., S. 268–269)

Das bekannteste Argument auf Seiten der Dilemma-Kritiker ist das sogenannte ›Sollen-impliziert-Können-Argument‹, welches auf zwei Axiomen deontischer Logik basiert. Das ›*ought-implies-can*‹-Prinzip (*OIC*) fordert, dass ›Sollen‹ (*ought*) immer zugleich auch ›Können‹ (*can*) impliziert. Wann immer ein Akteur eine Handlung tun *soll*, muss ihm diese auch *möglich* sein. Das Zusammenfallen von ›Sollen‹ und ›Können‹ wird – die Realität moralischer Dilemmata vorausgesetzt – durch ein anderes Prinzip der deontischen Logik, das sogenannte Agglomerationsprinzip (*agglomeration principle/AGG*), in Frage gestellt. Letzteres besagt, dass, wenn ein Akteur A und B aufgrund vorliegender moralischer Gebote tun soll, er auch tatsäch-

181 In ähnlicher Weise formuliert Marcus (1980, S. 128 bzw. 129, Hervorh. i. Orig.): »[…] we can define a set of rules as consistent if there is some possible world in which they are all obeyable in all circumstances in *that* world. « Und: »[…] rules are consistent if there are possible circumstances in which no conflict will emerge. […] a set of rules is inconsistent if there are *no* circumstances, no possible world, in which all the rules are satisfiable«.

lich beide (A und B) tun soll:[182] „›I ought to do *a*‹ and ›I ought to do *b*‹ together imply ›I ought to do *a* and *b*‹ (which I shall call the *agglomeration* principle).« (Williams, 1987, S. 130, Hervorh. i. Orig.) Kann der Akteur nicht A und B gleichzeitig wählen, so kann es nicht wahr sein, dass beide Optionen im Sinne von *ought* geboten sind. Denn aus dem Agglomerationsprinzip folgt, dass der Akteur beide Optionen wählen soll, dies aber in einer dilemmatischen Situation definitionsgemäß ausgeschlossen ist. Das entsprechende moralische Gebot ist nicht bindend und es besteht kein Dilemma; niemand kann verpflichtet sein, das Unmögliche zu tun (vgl. Sinnott-Armstrong, 1988, S. 109–113; van Fraassen, 1973, S. 12).

Wird hingegen die Möglichkeit moralischer Dilemmata angenommen, so muss entweder *AGG* oder *OIC* verworfen werden, wodurch nach Ansicht der Dilemma-Kritiker die Konsistenz unserer Moralsysteme in Frage gestellt würde (vgl. McConnell, 1978, S. 270–272). Befürworter betonen, dass dies jedoch nicht notwendigerweise der Fall ist. Sowohl auf Seiten der Kritiker als auch der Befürworter stellt die Plausibilität von *OIC* nicht nur die bekannteste, sondern auch die kontroverseste Debatte dar.[183] Ein Großteil der konzeptionellen Forschung zu moralischen Dilemmata in den letzten sechzig Jahren dreht sich darum, wie sich die von den Kritikern angezeigten logischen Widersprüche vermeiden lassen. Im Zentrum steht die kritische Analyse der Axiome (*OIC* und *AGG*), auf die sich das Argument gründet. Insbesondere die Interpretation des Begriffs *ought* erweist sich als problematisch und begründet eine der zentralen Fragestellungen der frühen Debatte.

So fassen einige metaethische Positionen die Existenz moralischer Dilemmata als Widerlegung des *OIC*-Prinzips auf. Lemmon (1962, S. 148) argumentiert beispielsweise, dass Pflichten, Verpflichtungen und moralische Prinzipien als mögliche Quellen moralischen Sollens miteinander in Konflikt stehen können. Einen alternativen, häufig referenzierten Ansatz zur Interpretation von *OIC* legt Trigg (1971) vor. Im Anschluss an Lyons' (1965, S. 21) Unterscheidung zwischen ›starken‹ und ›schwachen‹ moralischen Prinzipien vertritt er die Auffassung, dass sich Konflikte zwischen diesen beiden trivial

182 Die formale Notation des *AGG* lautet: (OA & OB → O(A & B)).

183 Ein ausführlicher Überblick über die Debatte und einschlägige Argumente finden sich bei Sinnott-Armstrong (1988, Kap. 4 & 5).

entscheiden lassen, wohingegen starke Prinzipien nicht außer Kraft gesetzt werden können und im Konfliktfall moralische Inkompatibilität erzeugen. Wenn zwei Regeln ein Dilemma hervorrufen, ist das zunächst einmal nicht ungewöhnlich. Es liegt vielmehr in der Natur moralischen Lebens, ohne dass dadurch die Gültigkeit von Prinzipien in Zweifel zu ziehen sei:[184] »The fact that a couple of moral rules happen to clash once does not affect either their general validity or their application in that situation. There is no logical incoherence here. It is just that life is sometimes tough.« (Trigg, 1971, S. 55) Moralische Dilemmata zeichnen sich in diesem Sinne dadurch aus, dass die konfligierenden Regeln eigentlich befolgt werden sollen, aber in einer konkreten Situation unter bestimmten Umständen – z. B. bei Zielkonflikten im Krieg – miteinander in Konflikt stehen. In diesen besonderen Fällen müssen wir zeitweise eine der beiden Regeln aufgeben, obwohl wir normalerweise beide anerkennen. Dilemmata entstehen, wenn wir aus moralischen Erwägungen zwei Dinge verbinden wollen, die sich praktisch nicht verbinden lassen; dabei handelt es sich jedoch keineswegs um eine logische Unmöglichkeit (ebd., S. 43–45).

Marcus (1980) bietet eine weitere Revision des *OIC* an. Sie präsentiert eine überarbeitete Version von Kants Prinzip, in das sie moralische Pflichten einbezieht und gleichzeitig Konfliktsituationen zwischen ihnen möglich macht. Die zugrundeliegende Idee von Marcus' Ansatz entstammt der Modelltheorie: Ein Set an Regeln ist genau dann konsistent, wenn es eine mögliche Welt gibt, in der alle diese Regeln unter allen Umständen einhaltbar sind. Wenn ein System in dieser Weise konsistent ist, dann muss *OIC* als Argument umformuliert werden in zwei Prinzipien: Das Prinzip erster Ordnung ist, dass jedes nicht-agglomerierte *ought* seinerseits *can* impliziert. Das Prinzip zweiter Ordnung ist, dass jedes agglomerierte *ought*, wenn es nicht unter allen Umständen eingehalten werden kann, so beschaffen sein muss, dass es möglich ist, eine Welt herzustellen, in der es eingehalten werden kann. Dieses zweite Prinzip ist aber lediglich regulativ, es impliziert nicht *can*. Es gibt also keinen glaubhaften Grund anzunehmen, dass wir Konflikte völlig vermeiden können, denn die Welt besteht schlichtweg aus Zufälligkeiten.

184 Lediglich wenn zwei Regeln dauerhaft in Konflikt stehen, liegt vermutlich eine grundlegende Inkohärenz vor (vgl. Trigg, 1971, S. 43).

Andere Kritiker begegnen ausgehend von der These der Existenz moralischer Dilemmata der Gültigkeit des *AGG* skeptisch. In einem einflussreichen Essay, der den Diskurs nachhaltig geprägt hat, demonstriert Williams (1987, S. 130–134) eine zentrale Schwäche des *AGG*, indem er zeigt, dass dieses nicht zwangsläufig aus *OIC* folgt,[185] sondern vielmehr das Gegenteil zutrifft: Wenn ein Akteur nicht beides (A und B) tun *kann*, so kann es nicht der Fall sein, dass er beides (A und B) tun *soll*.[186]

Auch Van Fraassen (1973) versucht sich an einer Widerlegung des *AGG*,[187] indem er die Schlussfolgerungen von Williams mit einer semantischen Interpretation anreichert. Dabei betont er die große Bedeutung moralischer Werte, die als eine Reihe von Grundsätzen oder Normen betrachtet werden, um uns bei unseren Entscheidungen zu leiten, und darauf basieren, was wir für richtig oder falsch halten. Vor diesem Hintergrund begreift er *ought* im Sinne des Versuchs, einen höchstmöglichen moralischen Wert zu realisieren bzw. den entsprechenden Zustand, der notwendige Bedingung dafür ist: »we ought to opt for the realization of the highest possible values, and, more generally, for any state of affairs that is a necessary condition for the realization of the highest attainable value.« (Ebd., S. 7) Ob ein *ought*-Statement wahr ist, bestimmen zum einen die verfügbaren Optionen und zum anderen die Werteskala, die wir anwenden. Sind die Optionen A und B in einem Dilemma inkompatibel, so kann es nicht sein, dass beide denselben Wert maximieren, und sie können daher nicht beide ein moralisches Sollen-Gebot begründen. Es gilt: Entweder ist A oder B moralisch geboten, oder die Situation ist indifferent (vgl. ebd., S. 7–8). Unter dieser Sichtweise

185 Für eine kritische Auseinandersetzung mit Williams siehe u. a. Holbo (2002, S. 261–262).

186 Williams (1987, S. 132) erläutert dazu: »I do not want to claim, […], that I have some knockdown disproof of the agglomeration principle; I want to claim only that it is not a self-evident datum of the logic of ought, and that if a more realistic picture of moral thought emerges from abandoning it, we should have no qualms in abandoning it. We can in fact see the problem the other way round: the very fact that there can be two things, each of which I ought to do and each of which I can do, but of which I cannot do both, shows the weakness of the agglomeration principle.«

187 Für eine kritische Auseinandersetzung mit Van Fraassen, vor allem mit seiner Widerlegung des *AGG*, siehe McConnell (1976).

muss also das *AGG* verworfen werden, wenn zugleich die Existenz von Dilemmata angenommen wird.[188]

An dieser Stelle können nun die Möglichkeit und Existenz moralischer Dilemmata als hinreichend belegt gelten. Wenn Dilemmata existieren, so drängt sich im Anschluss die Frage auf, wie sie entschieden werden können. Aufgrund ihrer komplexen Problemstruktur liegt es nahe anzunehmen, dass es keine triviale Entscheidungsstrategie gibt. Im folgenden Unterkapitel wird diese Vermutung anhand von Konzepten, die das Verhältnis der konfligierenden moralischen Gebote und ihrer zugrundeliegenden Werte aus metaethischer Sicht prägen, untersucht.

5.3 Lösbarkeit, Inkommensurabilität und (Un-)Vergleichbarkeit in Wertekonflikten

5.3.1 Vorrangbeziehungen und Prima-Facie-*Pflichten*

Eine weitere zentrale Streitfrage im metaethischen Dilemma-Diskurs dreht sich um die metaphysische Natur des zugrundeliegenden Wertekonflikts. Können Situationen, in denen Akteure lediglich nicht wissen, was sie tun sollen, als echte Dilemmata gelten? Und lässt sich damit die Lösbarkeit eines Konflikts überhaupt plausibel als akteurrelatives Charakteristikum denken? Die der Problematik zugrundeliegende Annahme ist, dass eine Lösung zwar existiert, Akteure diese jedoch aufgrund des Vorliegens bestimmter Umstände nicht erkennen können. Die Mehrheit des Diskurses steht einer solchen erkenntnistheoretisch begründeten Unlösbarkeit, die auch als *Argument des epistemischen Irrtums* bezeichnet wird, kritisch gegenüber und betrachtet die Realität echter Dilemmata als unabhängig vom Erkenntnisvermögen der Akteure. Nur weil ein Akteur in einer spezifischen Situation die Lösung nicht erkennen kann, bedeutet das noch nicht, dass es keine Lösung gibt, wie McConnell (1978, S. 279) formuliert: »That one does not know which hypothesis is correct does not by itself cast doubt on the claim that there is a uniquely correct answer to the question at issue.« Anders ausgedrückt: Nur

188 Für eine Kritik an Van Fraassens Aussagen, siehe z. B. Donagan (1984, S. 297–300).

weil nicht erkennbar ist, welches der konfligierenden Gebote stärker ist, heißt das noch nicht, dass dies auf keines der Gebote zutrifft. Mit einem solchen rationalistischen Argument lässt sich also noch keine zuverlässige Aussage über die Existenz echter Dilemmata treffen, weshalb epistemische Dilemmata im Diskurs sehr umstritten sind. Marcus (1980, S. 124–125) spricht ihnen gänzlich die Realität ab, indem sie epistemische Konflikte lediglich auf Unsicherheiten im Entscheidungsprozess zurückführt, die keinen Zusammenhang mit vermeintlichen Inkonsistenzen in Moraltheorien haben.[189]

Der überwiegende Teil des Diskurses ist der Meinung, dass echte Dilemmata nur ontologisch begründet werden können. Als beispielhafte Situationen werden hier häufig symmetrische Fälle wie Sophies Entscheidungsproblem aus Styrons Roman (1980) herangezogen. Sophies Dilemma besteht nicht darin, dass sie unsicher ist, für welches Kind sie sich entscheiden soll, sondern es *gibt* schlichtweg keine moralisch begründete Lösung in dem Sinne, dass eine der Alternativen eindeutig zu bevorzugen ist; es ist nicht *wahr*, dass eines der moralischen Gebote das andere aufhebt.[190] Wie bereits zuvor erläutert, unterscheiden sich bloße Konflikte und echte moralische Dilemmata im Hinblick auf das Verhältnis und die Natur der konfligierenden Gebote: Ein echtes moralisches Dilemma kann keine Gebote enthalten, die von einem anderen in moralisch relevanter Weise überwunden werden, wohingegen dies in moralischen Konflikten nicht gefordert ist; Letztere werden *gelöst*, Erstere werden *entschieden*. Über die Bedingungen dieses ›Überwindens‹ bzw. ›Vorrang-Habens‹ (*overriding*) herrscht eine rege Debatte, aus der

189 Statman (1995, S. 89–102) diskutiert die Diskrepanz zwischen epistemischen und ontologischen Dilemmata unter Bezugnahme auf Ronald Dworkin an einer analogen Thematik der Rechtsphilosophie: »In hard cases, a judge faces a practical question where it is unclear whether only one right answer exists, or whether a few possible right answers are available.« (Ebd., S. 89).

190 Ein weiteres populäres Beispiel für symmetrische Dilemma-Situationen ist das Johannes Buridan zugeschriebene Gleichnis eines Esels, der sich nicht zwischen zwei gleich großen und gleich weit entfernten Heuhaufen entscheiden kann und aufgrund seiner Entscheidungsunfähigkeit schließlich verhungert. Die vermeintliche (paradoxe) Unlösbarkeit dieser Entscheidungssituation ist nicht dadurch begründet, dass der Esel nicht *weiß*, welchen Haufen er fressen soll, sondern dass schlichtweg kein rationales Kriterium *existiert*.

vor allem Argumente contra moralische Dilemmata hervorgegangen sind.[191]

Allgemein lässt sich sagen, dass moralische Dilemmata genau dann *lösbar* sind, wenn eines der moralischen Gebote das andere überwindet. In diesem Zusammenhang werden Interpretationen von *Prima-Facie*-Pflichten, einem in der philosophischen Tradition vielfach bearbeiteten Terminus, kontrovers beurteilt. Die wohl bekannteste einschlägige Konzeption stammt von William David Ross (vgl. 1930, Kap. 2) und beruht auf einer Unterscheidung zwischen solchen Pflichten, die in einer konkreten Situation moralisch in Kraft gesetzt werden und daher tatsächlich befolgt werden sollen (*actual duties*), und solchen, für die dies nicht gilt (*prima facie duties*). In einer konkreten Situation gibt es immer eine sogenannte aktuale Pflicht, die gegenüber den konkurrierenden Pflichten Vorrang hat. Im Anschluss an Ross hat sich eine lebhafte Diskussion darüber entzündet, wie anhand von *Prima-Facie*-Pflichten die generelle Unmöglichkeit moralischer Dilemmata argumentativ gezeigt werden kann. Eine prominente Auffassung, die meist von Befürwortern moralischer Dilemmata vertreten wird, betrachtet *Prima-Facie*-Verpflichtungen als nur scheinbare Verpflichtungen (vgl. z. B. Foot, 1987, S. 257–258; van Fraassen, 1973, S. 8), die jegliche moralische Kraft verlieren, wenn sie überwunden werden. Allerdings ist anzumerken, dass diese dabei streng genommen ihren Status als moralische Gebote verlieren würden, weshalb keine Aussagen über echte moralische Dilemmata mehr möglich sind, in denen sich per definitionem moralische Gebote gegenüber stehen.

Einen bedeutenden Ansatz in diesem Zusammenhang legt Brink (1994) vor, indem er eine Unterscheidung zwischen *Prima-Facie*-Pflichten und allumfassenden Verpflichtungen (*all-things-considered obligations*) einführt, welche er in Anlehnung an Ross' einflussreiches Verständnis von *prima facie* und *sans phrase obligations* entwickelt. Demnach handelt es sich bei *Prima-Facie*-Pflichten um Eigenschaften einer Handlung, die diese als moralisch richtig aus-

191 McConnell (1978, S. 283–286) merkt zu Beginn der Debatte an, dass für die Unterscheidung zwischen scheinbaren und echten Dilemmata kein gültiges Kriterium vorliegt, auf das Befürworter ihre Argumentationen stützen könnten. Diese Lücke wird im Verlauf des andauernden metaethischen Diskurses allmählich geschlossen.

weisen. Beinhaltet eine Handlung beispielsweise das Einhalten eines Versprechens, so ist diese Handlung als richtig anzusehen.[192] Brink definiert allumfassende Verpflichtungen nun in diesem Sinne als unangefochtene *Prima-Facie*-Verpflichtungen (*undefeated prima facie obligations*), die von den stärksten moralischen Gründen gestützt werden:

> A prima facie obligation to do x means that there is a moral reason to do x or that x possesses a right-making characteristic. But prima facie obligations can be, and often are, defeated by other, weightier obligations, individually or in concert. A prima facie obligation to do x that is superior to all others constitutes an all-things-considered obligation to do x. An all-things-considered moral obligation to do x means that on balance, or in view of all morally relevant factors, x is what one ought to do or that x is supported by the strongest moral reasons. If prima facie obligations correspond to the presence of morally relevant factors or right-making characteristics, and an all-things-considered obligation is an undefeated prima facie obligation, then a natural way to understand a prima facie obligation to do x is as the claim that ceteris paribus, x is all-things-considered obligatory. (Ebd., S. 216)

Nach Brink gelten *Prima-Facie*-Pflichten im Sinne des Verständnisses von Ross nicht nur scheinbar, sondern *pro tanto*. Sie verlieren ihre moralische Geltungskraft nicht dadurch, dass sie durch ein anderes Gebot überwunden werden: »[…] prima facie obligations should be given a *metaphysical* reading that recognizes prima facie obligations as moral forces that are not canceled by the existence of other moral forces even if the latter override or defeat the former. « (Ebd., S. 218, Hervorh. i. Orig.) Ausgehend von diesem Verständnis folgert Brink, dass *Prima-Facie*-Verpflichtungen sowohl lösbare als auch unlösbare moralische Dilemmata hervorrufen können, wohingegen Dilemmata verstanden als Konflikte zwischen allumfas-

192 Ross (1930, S. 19, Hervorh. i. Orig.) erläutert im Wortlaut: »When I am in a situation, as perhaps I always am, in which more than one of these *prima facie* duties is incumbent on me, what I have to do is to study the situation as fully as I can until I form the considered opinion (it is never more) that in the circumstances one of them is more incumbent than any other; then I am bound to think that to do this *prima facie* duty is my duty *sans phrase* in the situation.«

senden Verpflichtungen nicht bestehen können.[193] Allumfassende Verpflichtungen besitzen per definitionem stets höchste moralische Kraft und unterliegen bekannten deontischen Prinzipien, weshalb sie zu Paradoxien in ethischen Theorien führen würden.[194] Nach Brinks Ansicht generieren unlösbare Konflikte lediglich disjunktive Verpflichtungen: »There is no all-things-considered obligation to do A or B, rather than the other. But there is an all-things-considered obligation to do one or the other, rather than some third thing. In particular, there is an obligation to do one or the other, rather than nothing.« (Ebd., S. 240) Somit sind unlösbare Konflikte zwischen *Prima-Facie*-Pflichten möglich und bedauerlich, aber nicht notwendigerweise paradox oder problematisch für ethische Theorien. Brink weist dennoch darauf hin, dass unlösbare Konflikte tragische oder mindestens unglückliche Folgen haben können, weshalb sie eine ernstzunehmende Thematik darstellen (vgl. ebd., S. 247).

Eine alternative Lesart des *Prima-Facie*-Konzepts, die von Ross' Verständnis abweicht, findet sich bei Hare (1987). Er betrachtet *Prima-Facie*-Prinzipien als allgemein und unspezifisch,[195] weshalb sie sich in ihrer simplen Form in spezifischen Konfliktsituationen leicht überwinden lassen. Nach Hare müssen ethische Entscheidungsstrategien immer im Rahmen eines kritischen Denkens (*critical thinking*) von einem höchsten (utilitaristischen) Prinzip Gebrauch machen:

> For Hare, prima facie principles are everyday surrogates for an ultimate oral principle, utilitarianism, from which their authority derives. Hence, when they conflict, authority reverts to the ultimate principle, from which the conflict may be resolved. For Ross, on the other hand, prima

193 Marcus (1980, S. 124–125) konstatiert, dass Vertreter der intuitionistischen Ethik in der Tradition von Ross die Realität moralischer Dilemmata verneinen. Sie gehen davon aus, dass es nie ein vollständiges moralisches System geben kann. Moralische Systeme sind nur Richtlinien, in deren Rahmen *Prima-Facie*-Prinzipien heuristischen Charakter besitzen; die ultimative Entscheidung ist letztlich eine rationale.

194 Für eine kritische Antwort auf Brink siehe Holbo (2002), der ein Argument dafür liefert, weshalb diese Art von Dilemmata nicht als Affront gegen die Moral, sondern als akzeptierte Realität angesehen werden sollte.

195 »Such principles express ›*prima* facie duties‹, and, although formally speaking they are just universal prescriptions, are associated, owing to our upbringing, with very firm and deep dispositions and feelings.« (Hare, 1987, S. 216, Hervorh. i. Orig.)

> facie duties are not surrogates for an ultimate principle: each has independent authority in the sense that no one can be derived from another or from a higher principle. Thus, when prima facie duties conflict, there can be no appeal to an ultimate principle. (Gowans, 1987, S. 13)

5.3.2 Symmetrie versus Inkommensurabilität: Kriterien und Konzeptionen

Klassische Vertreter der moralphilosophischen Tradition weisen die Möglichkeit moralischer Dilemmata mit dem Argument zurück, dass gültige Moraltheorien keine unlösbaren Fälle zulassen können. Dieser These können zwei Argumente entgegengestellt werden: Zum einen ist nicht einsichtig, dass zwingend ein Zusammenhang zwischen Existenz und Lösbarkeit (moralischer) Probleme bestehen muss: »Given the limitations on human action, it is naive to suppose that there is a solution to every moral problem with which the world can face us.« (Nagel, 1972, S. 144) Zum anderen ist grundsätzlich umstritten, ob Unlösbarkeit überhaupt eine zwingende Voraussetzung für das Vorliegen eines Dilemmas ist. Aus dem ontologischen Status von Dilemmata ergeben sich Implikationen für deren Lösbarkeit: Existiert keine ›richtige‹ Antwort, so lassen sich Dilemmata nur irrational *entscheiden*, aber nicht *lösen*. Statman (1995, S. 10–11) beschreibt diesen Sachverhalt wie folgt: »An irresolvable moral dilemma is a dilemma in which no moral consideration exists on the basis of which we could prefer one option over the other.« Doch wie muss das Verhältnis der konfligierenden moralischen Gebote beschaffen sein, sodass sie ein unlösbares Dilemma begründen?

Echte ontologische Dilemmata liegen genau dann vor, wenn beide Gebote absolute Gültigkeit besitzen. Kritiker moralischer Dilemmata sind der Ansicht, dass es solche Gebote nicht gibt, und berufen sich dabei gerne auf Ross' These, dass alle moralischen Prinzipien unter bestimmten Umständen aufgehoben werden können (vgl. Ross, 1930, Kap. 2). So schlussfolgert Donagan (1984, S. 303), dass manche moralischen Prinzipien zwar absolut gelten, aber jedes implizite Bedingungen enthält, aufgrund derer Ausnahmen geregelt sind. Allein aus der Tatsache, dass es absolut gültige Gebote geben mag, kann also noch nicht abgeleitet werden, dass es auch moralische Dilemmata gibt.

In diesem Kontext werden im Diskurs vor allem zwei Argumentationslinien verfolgt. Die erste bezieht sich auf die Annahme einer Symmetrie moralischer Gebote. Wie bereits in Kap. 5.2.2.3 ausgeführt, besagt das *Argument der Einzelwert-Konflikte*, dass ein und derselbe Wert unter bestimmten Umständen zwei inkompatible Handlungen verlangen kann. Sind deren moralisch relevante Eigenschaften identisch, so ist das Dilemma (rational) unlösbar, denn die gleichen moralischen Argumente lassen sich für beide Optionen vorbringen. Eine beispielhafte Situation ist jene der zu rettenden Zwillinge, wie sie Marcus (1980, S. 125) schildert. Allerdings erscheint es in real-lebensweltlichen Situationen tendenziell schwierig, eindeutig zu bestimmen, ob die Gebote wirklich exakt symmetrisch sind.

Ein plausibleres Argument geht daher von asymmetrischen Fällen aus. Der prominenteste Ansatz zur Begründung der Unlösbarkeit moralischer Dilemmata basiert auf der Idee der Inkommensurabilität moralischer Werte bzw. Gebote.[196] Raz (1986, S. 322) definiert folgendermaßen: »A and B are incommensurable if it is neither true that one is better than the other, nor true that they are of equal value. « In diesem Sinne ist Inkommensurabilität sowohl für symmetrische als auch asymmetrische Fälle relevant. Sie lässt sich dabei jeweils auf unterschiedliche Weise anwenden: Während bei symmetrischen Fällen kein Gebot das andere überwinden kann, weil die konfligierenden Gebote *gleich* sind, ist dies bei inkommensurablen Werten gerade deshalb nicht möglich, weil die Gebote *nicht gleich* sind. Wie lässt sich diese Aussage moralphilosophisch fassen? Statman (1995, S. 61) schlägt die Irreduzibilität von Werten als Definitionskriterium vor. Demnach sind zwei Werte inkommensurabel, wenn sie weder

196 Das Konzept der Inkommensurabilität stammt ursprünglich aus der Mathematik und besagt, dass es keinen gemeinsamen Nenner gibt, um zwei unterschiedliche Objekte hinsichtlich einer bestimmten Eigenschaft ineinander zu übersetzen. Sind zwei Zahlenwerte inkommensurabel, so ist ihr Verhältnis eine irrationale Zahl. Zur Veranschaulichung dient als klassisches Beispiel, dass es unmöglich ist, die Länge der Hypotenuse eines gleichschenkligen Dreiecks in die Länge dessen Katheten zu übersetzen. In die Wissenschaftstheorie wurde der Begriff in den 1960er-Jahren von Thomas Kuhn und Paul Feyerabend eingeführt. Hier bezieht sich Inkommensurabilität nicht auf Objekte, sondern auf Theorien. Diese können dahingehend inkommensurabel sein, dass sie eine Problemstellung aus verschiedenen Perspektiven beleuchten und dabei zu unterschiedlichen Ergebnissen kommen (vgl. Statman, 1995, S. 56–57).

aufeinander noch auf einen dritten Wert reduziert werden können, wie es z. B. bei Freiheit und Gerechtigkeit der Fall ist.

Ein anderes prominentes Konzept von Inkommensurabilität als Ausdruck der Zersplitterung moralischer Werte entwickelt Nagel (1979a). Für ihn sind Werte keine realen, unabhängig gegebenen Entitäten, sondern vielmehr Kriterien, anhand derer sich Zustände in der Welt beurteilen lassen. Nagels fünf Arten von Werten sind Ausdruck unterschiedlicher, unvergleichbarer Sichtweisen auf die Welt, die sich immer nur von einer bestimmten Perspektive aus reflektieren lässt. Dieses Konzept erscheint jedoch zu streng, denn so wären niemals irgendwelche Werte aus verschiedenen Arten vergleichbar (ebd., S. 60). Tatsächlich ist Inkommensurabilität nicht zwangsläufig gleichbedeutend mit Unvergleichbarkeit. Lässt sich ein gemeinsames Kriterium finden, anhand dessen zwei Probleme bewertet werden können, so lassen sich auch asymmetrische Fälle miteinander vergleichen:

> Moral requirements are comparable only when some comparative judgement of their strengths is true. The only comparative judgements are that the strength of one moral requirement is greater than, less than, or equal to the strength of the other. Thus, moral requirements are incomparable if and only if neither is stronger than, weaker than, or equal in strength to the other. Conflicts between incomparable moral requirements are then moral dilemmas, because neither moral requirement overrides the other. (Sinnott-Armstrong, 1988, S. 58)

Moralische Relevanz hat diese Unterscheidung insofern, als sich in manchen Fällen auch inkommensurable Werte rational miteinander vergleichen lassen und in einem Dilemma trotz Vorliegen von Irreduzibilität entsprechend Vorrang haben können. Beispielsweise dann, wenn extreme Unvergleichbarkeit durch erhebliche quantitative Unterschiede zwischen zwei Werten besteht, z. B. wenn ein triviales Versprechen gegen ein Tötungsverbot steht (vgl. Sinnott-Armstrong, 1988, S. 59). In weniger klaren Fällen kann von begrenzter Unvergleichbarkeit ausgegangen werden (vgl. ebd., S. 62–69); diese kann durch Prioritätsprinzipien (negative vor positiven Geboten) oder Ungenauigkeiten in moralischen Rankings Ausdruck finden. Gewisse Gebote sind hingegen grundsätzlich schwierig aufzuwiegen. Selbst wenn sich eine Messzahl zuordnen ließe und somit ein Vergleich theoretisch möglich wäre, ist es nicht plausibel anzunehmen, dass beispielsweise der Tod sich mit einer bestimmten Menge an

anderen Werten wie Schmerzen oder Freiheit gleichsetzen ließe. Auch eine ordinale Anordnung, die ohne kardinale Messgrößen auskommt, funktioniert nicht für alle Fälle, etwa wenn eine Option mehrere moralische Gebote enthält oder diese nach mehreren Skalen bewertet werden können, wie die Intensität und Dauer eines Schmerzes (vgl. ebd., S. 67–69).

Da inkommensurable Gebote nicht notwendigerweise unvergleichbar sein müssen, stellt Inkommensurabilität allein noch kein sicheres Kriterium dar, um die Unlösbarkeit von Dilemmata zu begründen. Sie ist jedoch hilfreich für ein besseres Verständnis spezifischer Aspekte, die im Zuge der metaethischen Dilemma-Diskussion erörtert werden. So erklärt sie zum einen, wieso es uns manchmal schwer fällt die richtige Antwort zu finden, auch wenn epistemische Unsicherheit keine echten Dilemmata begründen kann (vgl. Statman, 1995, S. 66–71). Zum anderen hilft sie zu verstehen, wie moralische Verluste zu Gefühlen des Bedauerns führen können; wären alle Optionen aufeinander reduzierbar, ließe sich durch die eine der Verlust der anderen ersetzen. Im Hinblick auf das Kritikerargument, das auf universellen Ansprüchen an Moraltheorien fußt, ist zu beachten, dass das Vorliegen von Inkommensurabilität nicht das Versagen von Moraltheorien impliziert. Sie ist vielmehr ein Fakt, der menschlichen Werten anhängt und allen Moraltheorien gleichermaßen Grenzen setzt (vgl. ebd., S. 56).

Im Anschluss an die hier erläuterten Überlegungen werden im nachfolgenden Unterkapitel die Implikationen ergründet, die mit der Inkommensurabilität spezifischer moralischer Werte einhergehen. Anhand einiger ausgewählter Ansätze wird eine Brücke geschlagen zwischen der bisherigen abstrakten metaethischen Betrachtung hin zu anwendungsnäheren ethischen Konzeptionen. Diese stellen den letzten Schritt dar, bevor die in diesem fünften Kapitel erarbeiteten Erkenntnisse auf den Anwendungskontext moralischer Unfallszenarien übertragen werden.

5.3.3 Metaethische Konzepte unvermeidbaren Scheiterns: Von unersetzbaren Verlusten und nicht-verhandelbaren moralischen Werten

Ausgangspunkt der Begründungsversuche einer Inkommensurabilität moralischer Pflichten in Bezug auf Dilemma-Strukturen ist die Annahme, dass zurückgewiesene moralische Gebote bestehen bleiben und damit unweigerlich verletzt werden. Zu den prominentesten Konzeptionen in diesem Kontext zählt Christopher Gowans' (1994) Ansatz des unausweichlichen moralischen Fehlverhaltens (*inescapable moral wrongdoing*). Diesem liegt die Idee zugrunde, dass eine Handlung auch dann moralisch falsch sein kann, wenn sie die bestmögliche Entscheidung eines spezifischen Konflikts realisiert – und zwar in dem Sinne, dass sie gegen den moralischen Wert der zurückgewiesenen Alternative verstößt, wobei sie einen moralischen Restwert auslöst. Worin besteht dieser übertretene Wert? Gowans fasst Inkommensurabilität als eine Eigenschaft auf, die dem intrinsischen Wert eines jeden Einzelnen anhaftet. Besondere Bedeutung kommt dabei einer spezifischen Kategorie moralischer Werte zu, die in der Verantwortung gegenüber anderen Personen besteht. Diese leitet sich ab aus dem intrinsischen Wert, den Gowans (wie Kant) jedem Individuum zuschreibt und der zugleich auch (anders als bei Kant) die Einzigartigkeit und damit Unersetzbarkeit jeder Person in moralisch signifikantem Sinne begründet (vgl. ebd., S. 121–128).[197]

197 Während Kant seine Idee der Selbstzweckhaftigkeit jedes Einzelnen über Rationalität und Autonomie als gemeinsame Basis aller Personen begründet, betont Gowans (1994, S. 123), dass Personen ganzheitlich betrachtet einen intrinsischen und einzigartigen Wert besitzen, der zugleich moralisch signifikant ist: »First, for Kant it is only the noumenal person, the person as a rational and free agent, that is regarded as an end in itself. On the responsibilities to persons account it is […] the whole person that is so regarded. The whole person may include rationality and autonomy, […] but it also includes much about a person that Kant would regard as being merely ›empirical‹ and hence of no moral significance in this regard – for example, a person's capacity for emotional response, a person's physical comportment in the world, and various features associated with a person's particular history. […] Second, Kant regards respect for persons as ends in themselves as a manifestation of respect for the moral law dictated by pure practical reason. […] On the responsibilities to persons account, we do not regard an intimate as intrinsically valuable by application of an a priori moral law, but through the experience of concrete interaction. […] Moreover, as we come to have a general view of all human

Verantwortung füreinander entsteht zum einen aus dem einzigartigen Wert jedes Einzelnen und zum anderen aus zwischenmenschlichen Beziehungen und sozialen Verbindungen, welche eine spezifische Verantwortung beinhalten:

> […] moral responsibilities have a twofold origin: first, in the belief that because persons are intrinsically and uniquely valuable, they are beings who are in various ways deserving and as such are beings for whom we can have responsibilities; and second, in the diverse ways in which particular persons come to be connected with one another, whether through choice or unchosen circumstance, and thereby establish a relationship of which specific responsibilities are a constitutive part. (Ebd., S. 128)

Was bedeutet das nun in Bezug auf Dilemmata? Ein zentrales Element von Gowans' Ansatz ist die Inkonvertibilität bzw. Nicht-Einlösbarkeit von Werten (*inconvertibility*). Im Konfliktfall bleiben moralische Gebote genau dann bestehen, wenn die mit den jeweiligen Geboten assoziierten Werte sich nicht vollständig durch das überwindende Gebot ersetzen lassen und daher nicht bloß *prima facie* gelten (ebd., S. 122–123). Dies ist genau dann der Fall, wenn die Werte eine spezifische Verantwortung gegenüber einer Person enthalten:

> The principal reason we sometimes have conflicting responsibilities is that these responsibilities originate in responses to the intrinsic and unique value of each of the particular persons with whom we are connected. It is the recognition of the value of, and hence the appropriateness of our specific response to, each of these persons that generate our various moral responsibilities. […] It is because they develop out of these separate responses to distinct persons that conflict cannot plausibly be eliminated. (Ebd., S. 132)

Aufgrund der Einzigartigkeit von Personen wird die auf sie gerichtete Verantwortung selbst einzigartig und schließt die Unersetzbarkeit bestimmter moralischer Verluste mit ein (vgl. ebd., S. 132-134). Echte Dilemmata sind daher stets mit einem unausweichlichen Scheitern verbunden, das durch die Nicht-Einlösbarkeit zurückgewiesener Werte verursacht wird und einen Zustand moralischer Unschuld als unerreichbares Ideal charakterisiert:

---

beings as valuable in themselves, this is determined inductively from particular cases, and not as a result of an a priori apprehension of rational nature.«

> [...] choices are inconvertible when the better choice still results in a loss, when there is something that the poorer choice would have provided that is not provided by the better choice [...]. When choices are inconvertible, it is possible to have regret without having any doubt that one made the better choice. (Ebd., S. 148)

Doch ist die von Gowans erläuterte Unersetzbarkeit als absolutes Kriterium ausreichend? Lisa Tessman (2015) entwickelt Gowans' Konzept weiter und konstatiert, dass unersetzbare Verluste manchmal schlicht einen zu akzeptierenden Teil menschlichen Lebens darstellen und deshalb nicht immer moralisch problematisch sind. Um zu bestimmen, wann eine moralische Relevanz vorliegt, zieht Tessman den Ansatz von Martha Nussbaum (2000) heran. Dieser unterscheidet die gewöhnlich bei Abwägungen anfallenden Kosten (*ordinary costs*) von tragischen Kosten; der Schwellenwert zwischen beiden Kategorien muss von einer unabhängigen ethischen Theorie bestimmt werden. Nussbaum schlägt hierfür die menschliche Würde als Orientierungskriterium vor (vgl. ebd., S. 1032). Gemäß ihrem berühmten *Capability Approach* sollten jedem Menschen bestimmte Fähigkeiten in verschiedenen Bereichen wie beispielsweise körperliche Gesundheit und Integrität, Emotionen oder praktische Vernunft zugesichert werden (vgl. Nussbaum, 1999, 2006, 2009). Wird bei einer der Fähigkeiten der Schwellenwert unterschritten, so ist dies als moralisches Fehlverhalten zu werten. Ein Dilemma liegt genau dann vor, wenn »we find that we cannot get citizens above the capability threshold in one area, without pushing them below it in another area.« (Nussbaum, 2000, S. 1025)

Aus Sicht von Tessman (2015, S. 42) lässt sich mithilfe von Nussbaums Ansatz die Frage beantworten, die bei Gowans offengeblieben ist: Ein unersetzbarer Verlust liegt vor, wenn auch die beste mögliche Entscheidung nicht gut genug ist, um eine Person hinsichtlich aller relevanten Fähigkeit oberhalb des Schwellenwerts zu halten, i. e. das entsprechende moralische Gebot, welches sich nicht eliminieren lässt, verletzt wurde. Allgemeiner bedeutet das: Moralische Gebote bleiben genau dann bestehen, wenn eine Entscheidung gegen sie Kosten auferlegen würde, die nicht kompensiert werden können. Diese sind nicht verhandelbar:

> Costs that are to be borne are costs that one can negotiate with; when they are counterbalanced with sufficient benefits, it becomes permissible to incur such costs. Costs that no one should have to bear are

> non-negotiable; there is no way to eliminate the moral requirement not to incur such costs, which means that if one decides to act on an alternative that will incur such costs, one does so in violation of a still standing moral requirement. (Ebd., S. 42)

Tessman kombiniert nun die Ansätze von Gowans und Nussbaum: Moralische Gebote können genau dann eliminiert werden, wenn der überwundene Wert eines Gebots entweder durch einen anderen substituiert (wobei kein einzigartiger Verlust entsteht) oder kompensiert wird (was zwar in einem einzigartigen Verlust resultiert, der aber in zumutbarer Weise zu tragen ist). In diesem Fall wird keines der Gebote unmöglich und es entsteht kein moralischer Restwert (vgl. ebd., S. 42–43).

Vor dem Hintergrund einer pluralistischen Position führt Tessman weiter aus, dass einerseits verschiedene Arten von Werten einander nicht kompensieren können und sich andererseits moralische Gebote anhand ihrer Geltungskraft kategorisieren lassen. Ihr Ausgangspunkt ähnelt in seiner Struktur dem phänomenologischen Argument von Williams (1987, S. 131–134), das im Fall eines bestehen bleibenden Gebots einen moralischen Restwert annimmt, geht jedoch über es hinaus. Tessman bezeichnet solche Gebote, die durch Substitution oder Kompensation im Zuge der Auflösung eines Konflikts anderen untergeordnet werden können, als verhandelbare (*negotiable*) Gebote. In Fällen hingegen, in denen weder Substitution noch Kompensation möglich ist, bleiben die Gebote bestehen, wobei das zurückgewiesene Gebot durch die Wahl des anderen unausweichlich verletzt und dabei unmöglich wird (vgl. 2015, S. 25–27). Im Zuge eines »unavoidable moral failure« kommt es zu einem einzigartigen Verlust, der nicht getragen werden muss; das entsprechende Gebot ist nicht verhandelbar (*non-negotiable*) (vgl. ebd., S. 43).[198]

Aus der Unmöglichkeit des zurückgewiesenen Gebots ergeben sich Implikationen auf logischer Ebene, die von Dilemma-Kritikern

198 Wann solche nicht-verhandelbaren Gebote vorliegen, begründet Tessman moral- und kognitionspsychologisch, wobei sie sich v. a. auf den Ansatz von Greene und Haidt (2002) bezieht: *non-negotiable requirements* liegen genau dann vor, wenn Emotionen, die intuitive Urteile hervorrufen, uns sagen, dass etwas verboten ist. Dagegen sind Emotionen, die durch *reasoning* entstehen, nur *prima facie* erforderliche Aktionen (vgl. Tessman, 2015, Kap. 2 bzw. S. 57–98). Neben Nussbaums *Capability Approach* diskutiert Tessman auch Harry Frankfurts Konzept der »volitional necessity« näher (vgl. ebd., S. 45–55).

allerdings im Zuge ihrer Fokussierung auf rationale Strategien wie Kosten-Nutzen-Abwägungen unbeachtet bleiben. So gehorcht ein unmöglich gewordenes Gebot nicht länger dem *OIC*-Axiom: »Non-negotiable moral requirements – those that cannot be absorbed into an all-things-considered ›ought‹ through either substitution or compensation – remain requirements, contravening the principle that ›ought implies can‹.« (Ebd., S. 44) Vielmehr sind die moralischen Restwerte, die derartige Gebote hervorrufen, verantwortlich für die Entstehung dilemmatischer Strukturen:

> I define a moral dilemma as a situation of conflict in which there is a moral requirement to do A and a moral requirement to do B, where one cannot do both A and B, and where neither moral requirement ceases to be a moral requirement just because it conflicts with another moral requirement, even if for the purpose of action-guidance it is overridden. In a dilemma, whichever action one chooses to perform, one violates what has become, through one's choice, the impossible moral requirement to do the other action. (Ebd., S. 15)

An dieser Stelle ist der metaethischen Analyse moralischer Dilemma-Strukturen in dem Maße, wie sie das in dieser Forschungsarbeit betrachtete Anwendungsproblem erfordert, Genüge getan. Nachfolgend werden die erarbeiteten Ergebnisse nun auf praktische Dilemma-Szenarien des autonomen Fahrens angewandt, um diese zum einen näher zu charakterisieren und zum anderen in einem weiteren Schritt konkrete Implikationen hinsichtlich möglicher Entscheidungsstrategien zu erschließen.

5.4 Anwendungsfall Unfalldilemmata: Interpretation aus metaethischer Sicht

5.4.1 Dilemmatische Unfallsituationen als Konflikte inkommensurabler Werte

In ihren ethischen Richtlinien zum autonomen Fahren plädiert die Ethik-Kommission für ein rigoroses Verbot der Aufrechnung persönlicher Schäden (vgl. Di Fabio et al., 2017, S. 18). Im Rahmen der Rekonstruktion des metaethischen Diskurses in den vorhergehenden Unterkapiteln wurde eine theoretisch-metaethische Grundlage gelegt, mittels derer sich nun eine Begründung für diese praktische

Forderung formulieren lässt: Die spezifische ethische Problematik von Entscheidungsdilemmata im Kontext von Unfallalgorithmen besteht in der inkompatiblen Verschränkung moralisch gleichrangiger legitimer Interessen von Individuen. Aus diesen lassen sich moralische Gebote ableiten, die in der absoluten Pflicht zum Schutz des Lebens bzw. dem Verbot bestehen, anderen Schaden zuzufügen.

Auch wenn es auf den ersten Blick so erscheinen mag, handelt es sich hierbei jedoch nicht um symmetrische Dilemmata. Zwar sind die Interessen der Beteiligten inhaltlich gleich und prinzipiell in gleichem Maße schützenswert, doch sind Dilemma-Situationen in konkreten Anwendungsfragen jenseits von metaethischen Betrachtungen selten wirklich symmetrisch. Beispielsweise kann es Unterschiede im Hinblick auf die zu erwartenden Folgen geben. In welchem Maße individuelle Interessen bzw. Werte in einer konkreten Dilemma-Situation gefährdet sind, hängt von verschiedenen Faktoren ab; das Ausmaß und die Wahrscheinlichkeit möglicher Schäden können für jede beteiligte Person unterschiedlich sein und sind sowohl von physikalischen bzw. mechanischen Aspekten wie Geschwindigkeit oder Aufprallwinkel als auch von persönlichen Merkmalen wie Alter, gesundheitlicher Konstitution usw. abhängig. Besonders umstritten ist in diesem Zusammenhang die Vulnerabilität von Individuen (siehe Kap. 7.3.3.2). Moralisch relevante Unterschiede, z. B. die familiäre Situation der Betroffenen, bestehen häufig einfach deshalb, weil betroffene Personen unvergleichbare Individuen sind. Das Diskriminierungsverbot, welches exemplarisch in den Empfehlungen der Ethik-Kommission verankert ist, widerspricht jedoch einer Berücksichtigung von Persönlichkeitsmerkmalen im Rahmen einer algorithmischen Entscheidungsfindung. Doch selbst wenn eine solche Vorgehensweise zulässig wäre, müssten für das Vorliegen symmetrischer Dilemmata die tangierten Werte sowohl quantitativ als auch qualitativ annähernd gleich sein. Gewisse Werte sind allerdings grundsätzlich schwierig gegeneinander aufzuwiegen; zu diesen zählen beispielsweise Freiheit, Gerechtigkeit oder Autonomie.

Wie zuvor gezeigt, folgt aus der Separatheit von Personen und der Einzigartigkeit menschlichen Lebens, dass dessen Wert nicht identisch, sondern inkommensurabel ist; ein Leben lässt sich nicht durch ein anderes ersetzen. Die zentrale Eigenschaft, die komplexe moralische Dilemmata im Kontext von Unfallalgorithmen kennzeichnet,

besteht daher nicht in der Symmetrie, sondern in der Inkommensurabilität involvierter Werte. Wie lässt sich diese inhaltlich begründen? Stehen die Leben verschiedener Personen gegeneinander, so liegt sowohl gemäß Tessman als auch gemäß Gowans ein Konflikt zwischen moralischen Geboten vor, die sich nicht überwinden lassen. Daraus folgt, dass sich der Konflikt zwischen den entsprechenden moralischen Geboten nur im Zuge moralischen Fehlverhaltens bzw. Scheiterns entscheiden lässt. Mit dem Gedankengang von Gowans (1994) gründet sich die Unlösbarkeit des Dilemmas darauf, dass das Verbot der Verletzung legitimer individueller Interessen mit einer einzigartigen Verantwortung gegenüber jeder der involvierten Personen verknüpft ist. Die konfligierenden Werte sind nicht ineinander überführbar, bei jeder möglichen Entscheidung käme es zu einem unersetzbaren Verlust in Bezug auf eine der betroffenen Personen. Welche Alternative der Steuerungsalgorithmus des autonomen Fahrsystems auch wählt, er kann seiner Verantwortung gegenüber jedem Einzelnen nicht gerecht werden.

Auch nach Tessmans (2015) moralpsychologischem Ansatz kommt es in diesem spezifischen Anwendungsfall zu einem unersetzbaren Verlust. Das Gebot des Schutzes des Lebens bzw. der körperlichen Unversehrtheit beinhaltet einen hochrangigen moralischen Wert, der sich nicht durch einen anderen ersetzen lässt. Auch die aus Sicht einer bestimmten ethischen Position beste mögliche Entscheidung[199] verursacht tragische Kosten im Sinne Nussbaums, die vor dem Hintergrund einer pluralistischen Sichtweise nicht kompensiert werden können. In der Folge bleiben die konfligierenden moralischen Gebote erhalten; das im Zuge einer praktischen Entscheidung zurückgewiesene Gebot würde durch die Wahl des jeweils anderen unausweichlich verletzt und dabei unmöglich. Der entstehende moralische Restwert zeigt einen einzigartigen Verlust an, der nicht getragen werden muss – die entsprechenden moralischen Gebote sind nicht verhandelbar:

> [...] if one chooses to kill one person in order that five others be saved, one imposes on the person who is killed (and on that person's loved ones) a cost that no one should have to bear (even if the best

199 Bei dilemmatischen Quantifizierungsproblemen könnte eine solche ›beste mögliche Entscheidung‹ beispielsweise in einer utilitaristischen Strategie bestehen, derzufolge eine möglichst geringe Anzahl an Personen geschädigt wird.

> action-guiding decision for someone who faces such a conflict is to kill that one person and thus impose a tragic cost). (Ebd., S. 43)

Zusammenfassend lässt sich festhalten: Das zentrale Problem, das die Komplexität moralischer Dilemma-Situationen im Kontext des autonomen Fahrens theoretisch-formal begründet, besteht in der inkompatiblen Verschränkung moralisch gleichrangiger legitimer Interessen von Individuen und der gleichzeitigen *Nicht-Verrechenbarkeit* deontologischer Pflichten zum Schutz des Lebens. Das entscheidende Argument besteht dabei in der Inkommensurabilität unersetzbarer und einzigartiger Werte. Aus metaethischer Sicht gibt es keine systematischen Strategien zur Entscheidung oder Vermeidung von Dilemmata inkommensurabler Werte. Im Hinblick auf praktische Anwendungsprobleme ist es jedoch erforderlich, *dass* Handlungsorientierung für diese Situationen bereitgestellt wird, auch wenn an diesem Punkt noch nicht klar ist, *wie* dies erfolgen kann. Im Folgenden werden mögliche Entscheidungsstrategien kritisch beleuchtet: Zum einen wird dem Vorschlag, Unfalldilemmata mittels eines Zufallsgenerators zu entscheiden, eine Absage erteilt; zum anderen wird aus der Warte einer pragmatischen Ethik ein möglicher (praktischer) Ausweg aus der theoretischen Unlösbarkeit skizziert.

5.4.2 Entscheidungsperspektiven für inkommensurable Wertekonflikte

5.4.2.1 Zurückweisung des Zufallsprinzips

Eine Konfliktkonstellation, in der sich inkommensurable Werte gegenüberstehen, impliziert, dass keine Abwägung zwischen den Alternativen möglich ist. Es gibt keinen moralischen Grund, auf dessen Basis für die eine oder die andere entschieden werden könnte. Einige Philosophen sind daher der Ansicht, daraus ließe sich folgern, dass in solchen Fällen per Zufallsprinzip entschieden werden sollte. Speziell im Kontext von Algorithmen erscheint es attraktiv, dilemmatische Unfallentscheidungen anhand eines softwaretechnisch unkompliziert zu realisierenden Zufallsgenerators durchführen zu lassen. Auch im Diskurs um Unfallalgorithmen wurde diese Methode bereits vorgeschlagen (vgl. z. B. Gantsho, 2022; Zhao & Li, 2020). Ein Zufallsgenerator würde in Dilemma-Situationen über mehrere Handlungsoptionen randomisieren und so beispielsweise immer

dann nach links ausweichen, wenn eine gerade Zahl generiert wird, bzw. bei einer ungeraden Zahl entsprechend nach rechts.[200] Auf diese Weise ließe sich zuverlässig ausschließen, dass persönliche Vorteile in die Entscheidungsfindung einfließen, das Vorgehen wäre weitgehend unparteiisch. Doch ist das Fehlen guter (moralischer) Gründe, welche es rechtfertigen würden, argumentativ für eine der Alternativen zu entscheiden, bereits ausreichend, um die Entscheidung legitimerweise an den Zufall zu delegieren? In Bezug auf symmetrische Fälle wird häufig argumentiert, dass es hierbei aus moralischer Sicht egal sei, was man wählt (vgl. Donagan, 1984, S. 307). Lassen wir im Fall der Zwillinge den Zufall entscheiden, so behandeln wir beide mit gleicher Gerechtigkeit und als Selbstzweck, sofern beide die gleiche Überlebenswahrscheinlichkeit haben. In diesem Sinne argumentiert z. B. Gantsho (2022), dass nur ein nicht-deterministischer, zufallsbasierter Algorithmus der Prämisse der gleichen Unverletzlichkeit jeder Person gerecht werden kann.[201]

Inkommensurable Fälle per Zufallsprinzip zu entscheiden ist hingegen eine in der Moralphilosophie sehr umstrittene Methode. Das zentrale kritische Argument lautet, dass randomisierte Entscheidungen im Kern willkürlich[202] sind und auf einer indifferenten Haltung

200 Eine alternative Implementierung des Zufallsprinzips könnte auch darin bestehen, das System zufallsbasiert zwischen unterschiedlichen ethischen Prinzipien wählen zu lassen, beispielsweise zwischen utilitaristischen und deontologischen, die dann im jeweiligen Fall zur Anwendung kommen.

201 Gantshos (2022, S. 181) zentrales Argument lautet folgendermaßen: »In my argument for the equal inviolability of persons, I have conceded that the inviolability of a person's rights can be violated in dire situations where tragic choices of life and death must be made. I argued that the best way to reconcile the view that inviolability of moral status is a threshold concept with the view that persons can be sacrificed for others in catastrophic situations is for autonomous vehicles to select who ought to be sacrificed with a random nondeterministic programme. Taking any capability, interest or good into consideration when deciding which person should die would be to contradict the respect-based account with its threshold concept of moral status and inviolability. Some innocent persons may be sacrificed for the many, but such a decision must be blind to any of the differences that set persons apart from each other.«

202 Im philosophischen Kontext ist Zufall ein komplexes Konzept, das »im weiteren Sinne alles [bezeichnet], was nicht als notwendig oder beabsichtigt erscheint und für dessen unvermutetes Eintreten wir keinen Grund angeben können.« (Regenbogen & Meyer, 2013, S. 751) Das hier gegebene Argument basiert auf einem Verständnis des Zufallsbegriffs im Sinne eines absoluten

fußen, die den gebotenen Respekt gegenüber Leben und Rechten der Betroffenen vermissen lässt.[203] Dies ist in zweierlei Hinsicht problematisch: Zum einen sind zum Zeitpunkt der Implementierung eines möglichen Zufallsgenerators noch keine konkreten Ergebnisse mit den möglichen Handlungsoptionen verknüpft. Daher sind sie auch nicht moralisch bewertbar. Definitive Entscheidungen werden so vermieden, für die folglich auch keine Verantwortung entstehen kann. Als moralfähige Wesen schulden wir einander jedoch eine gewisse Sorgfalt bei ethischen Entscheidungen. Wie Gowans bemerkt, ist es die einzigartige Verantwortung gegenüber jedem Individuum, die eine Verletzung derselben zu einem *irreplaceable loss* macht. Wenn wir die Entscheidung an eine andere Instanz, den Zufall, übergeben, so sind wir für dieses Delegieren ebenfalls verantwortlich. Hier sind neue Verantwortungskonzepte vonnöten, die das Zusammenspiel von Menschen und autonomen Systemen in den Blick nehmen. Bis dato kann ein Zufallsgenerator die menschliche Entscheidungsverantwortung grundsätzlich nicht auflösen. Aus ver-

---

Zufalls: »Dementsprechend hat das Wort Z. die drei Bedeutungen des Nichtwesentlichen, des Nichtnotwendigen oder des Nichtbeabsichtigten. Wird unter Z. das Nichtnotwendige verstanden, ist also absoluter Z. gemeint, so bedeutet Z. eine Durchbrechung des Kausalgesetzes und setzt die Möglichkeit teilweise freien, willkürlichen Geschehens voraus […]« (Ebd., S. 751).

203 Statman (1995, S. 79–80) versteht unter bestimmten Bedingungen eine Zufallsentscheidung gerade als Ausdruck des Respekts gegenüber Individuen, indem sie impliziert, dass es keinen Grund gibt, eine der beiden Alternativen zu bevorzugen. Grundlage ist, dass er die Grenze zwischen symmetrischen und inkommensurablen Werten als weniger trennscharf betrachtet. Für ihn ist die Symmetrie eines Konflikts nicht einfach gegeben, sondern ergibt sich aus einer normativen Position, gemäß derer menschlichem Leben ein absoluter Wert beigemessen wird. Dieser erlaubt es uns, alle anderen Überlegungen zu ignorieren, die möglicherweise faktisch vorhandene moralisch relevante Unterschiede begründen könnten. Nur unter diesen Prämissen liegt Statmans Ansicht nach echte Symmetrie und damit Unlösbarkeit vor, der mithilfe einer Zufallsentscheidung begegnet werden sollte: »Though flipping a coin is usually a sign of the insignificance of the options, in the present context it might serve as an expression of the equal respect we have for all human beings.« (Ebd., S. 80).

antwortungsethischen Gründen erscheint es daher fragwürdig, Entscheidungen über Menschenleben dem Zufall zu überlassen.[204]

Zum anderen kann der Einsatz eines Zufallsgenerators als Versuch gewertet werden, sich einer bewussten, begründeten Entscheidung entziehen zu wollen. Dies ist jedoch nur ein vermeintlicher Ausweg aus der Entscheidungsnotwendigkeit, denn auch auf diese Weise wird eine Entscheidung getroffen; nämlich die, sich nicht im Einzelfall ethisch damit auseinandersetzen zu wollen. Eine Anwendung des Zufallsprinzips nimmt den Akteuren zwar scheinbar die Last des überlegten ›Entscheiden-Müssens‹ ab (vgl. Raz, 1986, S. 331–332), widerspricht im Kern jedoch dem ureigenen Wesen moralischer Entscheidungen. Diese müssen aus dem jeweiligen gesellschaftlichen Kontext heraus sorgfältig begründet und argumentativ verteidigt werden. Mit einer an einen Zufallsgenerator delegierten Entscheidung geht ein hohes Maß an menschlichem Kontrollverlust und Transparenzeinbußen bezüglich der gewählten Handlung einher (vgl. Lin, 2014a), denn der Zufallsgenerator entscheidet im Sinne einer *Black Box* letztlich rein auf Basis technisch generierter Zufallszahlen. Eine Zufallsentscheidung versäumt es, im Einzelfall relevante Gründe angemessen zu würdigen, und untergräbt somit den moralisch gebotenen Respekt gegenüber den Einzelnen.

Schlussendlich lässt sich eine auf Zufall basierende Entscheidung von Unfalldilemmata nicht plausibel rechtfertigen. Dilemma-Situationen im Kontext von Unfallalgorithmen bedürfen einer bewussten Auseinandersetzung, der wir uns als moralische Akteure trotz der Unvermeidlichkeit moralischen Versagens nicht entziehen können: »Despite the inevitability of moral failure, the people who confront these dilemmas experience themselves as choosing, however coerced the choices may be.« (Tessman, 2015, S. 163)

Welche Perspektiven stehen nun zur Entscheidung unlösbarer Dilemmata zur Verfügung? Im Folgenden wird dargelegt, wie die Bewältigung von Unfalldilemmata durch eine Orientierung an praktischen Zwecken und Gegebenheiten im Kontext der jeweiligen Entscheidungssituation gelingen kann, ohne formale Limitationen außer Acht zu lassen.

---

204 Für eine tiefergehende philosophische Diskussion von (dilemmatischen) Zufallsentscheidungen zwischen Menschen siehe z. B. Broome (1984) oder Jacquette (1991).

5.4.2.2 Pragmatische Ethik: Ein möglicher Ausweg aus dem Dilemma?

Im metaethischen Diskurs herrscht dahingehend Einigkeit, dass es keine systematischen Entscheidungsstrategien für Dilemmata inkommensurabler Werte gibt. Was folgt daraus für den Status der Moralphilosophie als systematische normative Disziplin? Mit seinem Entwurf einer neuen Vision für die Moralphilosophie zeigt Nagel (1979b, S. 180–184) einen Weg auf, wie sich diese neu definieren kann, ohne angesichts der Möglichkeit moralischer Dilemmata vor ihren eigenen Ansprüchen kapitulieren zu müssen. Voraussetzung dafür ist, dass die Moralphilosophie im Zuge einer Weiterentwicklung zu einem veränderten Selbstverständnis gelangt. Anstatt ein allgemeingültiges Entscheidungsverfahren bereitzustellen, nimmt sie im Rahmen von Nagels Konzept lediglich die Rolle eines Ratgebers ein, »*indem sie Akteuren und Institutionen beratend zur Seite steht, damit deren moralische Entscheidung möglichst vernünftig und mit möglichst guten Gründen, sprich: möglichst objektiv, ausfällt.*« (Raters, 2013, S. 348–349, Hervorh. i. Orig.) Die Wandlung der Moralphilosophie vollzieht sich in zwei Schritten. Zunächst muss sie ein Verfahren anbieten, das in der Herangehensweise an ein moralisches Problem alle relevanten Positionen berücksichtigt:

> What we need most is a method of breaking up or analyzing practical problems to say what evaluative principles apply, and how. This is not a method of decision. Perhaps in special cases it would yield a decision, but more usually it would simply indicate the points at which different kinds of ethical considerations needed to be introduced to supply the basis for a responsible and intelligent decision. (Nagel, 1979b, S. 184)

Im Anschluss erfolgt eine Zerlegung des Problems in kleinere Teilprobleme, die sich u. a. über Ansätze aus verschiedenen Wissenschaften aufgreifen lassen und in eine finale Abwägung einfließen; wie diese konkret erfolgen soll, beschreibt Nagel allerdings nicht explizit. Die auf diese Weise sich selbst neu definierende Moralphilosophie kann zwar den Widerspruch zwischen konfligierenden Normen, der durch heterogene Ursprünge moralischer Handlungsgrün-

de hervorgerufen wird, nicht überwinden, aber immerhin Perspektiven einer wohlbegründeten Entscheidung aufzeigen.[205]

Auch wenn Nagels Ansatz in seiner Argumentation prinzipiell überzeugend ist, bleibt er jedoch im Hinblick auf die Bewältigung von Dilemmata in praktischen Anwendungskontexten defizitär. Dadurch, dass die Moralphilosophie sich lediglich als Beraterdisziplin definiert, verbleibt die Verantwortung für getroffene Entscheidungen vollständig bei den moralischen Akteuren. Die Moralphilosophie gibt hier keine handlungsleitende Orientierung, sondern artikuliert nur mögliche Einwände gegen jede der Optionen und erzeugt somit Schuldgefühle bei denjenigen, die schlussendlich Entscheidungen treffen (müssen) (vgl. Nagel, 1979c, S. 327; Raters, 2013, S. 353–355). In ihrer visionären Gestalt im Sinne Nagels bahnt die Moralphilosophie nicht den Weg zu einer Entscheidung, welche die Akteure entlastet; im Gegenteil, sie betont, dass es bei Vorliegen dilemmatischer Strukturen keine Möglichkeit gibt, der Schuld zu entrinnen – vor allem dann nicht, wenn die beste Entscheidung nicht offenkundig ist (vgl. Raters, 2013, S. 381–382). Dem Akteur bleiben somit immer Restzweifel, die handlungshemmend wirken können und ebenso wie subjektive Schuldgefühle ein ernstzunehmendes Problem eröffnen.[206]

Schließlich bleibt festzuhalten, dass Nagels Konzept zwar die Existenz moralischer Dilemmata plausibel integrieren kann, hinsichtlich möglicher Strategien im Umgang mit Letzteren aber zu unkonkret bleibt und insbesondere zentrale Fragen im Kontext von Schuld und Verantwortung aufwirft. Zudem bleibt offen, wie ausgehend von einem »überpersönliche[n] […] Standpunkt moralischer Objektivität«, der im Zuge einer Moralphilosophie als Beraterdisziplin erreicht wird, ein konkretes Dilemma in letzter Instanz zu entscheiden ist (vgl. ebd., S. 380).

Im Allgemeinen findet der metaethische Diskurs in weiten Teilen auf einer abstrakten Ebene statt, die Dilemmata zumeist isoliert betrachtet. Zur Veranschaulichung werden zwar auch real-lebensweltli-

205 Siehe hierzu die Darstellung von Nagels Argument in Kap. 5.2.2.2.

206 Um dem Problem der subjektiven Schuldgefühle zu begegnen, entwirft Raters ein »Prinzip der subjektiven Minimierung der moralischen Verfehlung« (vgl. 2013, S. 383–390) sowie ein Schema, das pragmatische Hilfestellung geben kann, um eine gut begründete Entscheidung zu treffen (vgl. ebd., S. 390–422).

che Beispiele herangezogen, die aber in der Regel in relativ simple Zusammenhänge eingebettet sind. Dilemma-Situationen hingegen, die im Kontext von komplexen Forschungsproblemen aus der realen Lebenswelt auftreten, sind weithin anspruchsvoller, da sie in spezifischen Strukturen des jeweiligen Problemkontextes verhaftet sind. Wie Nagel bereits mit der vorgeschlagenen Zerlegung moralischer Probleme in kontextabhängige, disziplinbezogene Teilprobleme andeutet, können Dilemmata in praktischen Fragen nicht entschieden werden, ohne deren situative Umstände in die Überlegungen miteinzubeziehen.[207] Da Anwendungskontexte allerdings divers und in hohem Maße spezifisch sind, ist die Anwendbarkeit verallgemeinerter Ansätze fraglich. Auch wenn die Moralphilosophie keine eindeutige Handlungsorientierung offerieren kann, bedeutet das jedoch nicht, dass keine Entscheidungsnotwendigkeit bestünde:

> The unavailability of a single, reductive method or a clear set of priorities for settling them does not remove the necessity for making decisions in such cases. When faced with conflicting and incommensurable claims we still have to do something – even if it is only to do nothing. (Nagel, 1979b, S. 180)

Vielmehr gilt, dass praktische Probleme unbedingt klärende Entscheidungen erfordern, um gesetzte Ziele zu erreichen. Wenn diese auf rein theoretischer Ebene nicht zu finden sind, sind sie dann in Verbindung mit praktischen Überlegungen möglich? Dieser Gedanke liegt Ansätzen zugrunde, die eine pragmatische Ausrichtung in der Bewältigung echter moralischer Dilemmata propagieren.[208] Die pragmatische Ethik als normative Theorie besagt dabei, dass zur Entscheidung herangezogene moralische Gründe durch ihre Sachbezogenheit legimitiert werden.[209] Sie abstrahiert von theoretischen

---

207 Beispielhaft bemängelt Gowans (1994, S. 132) das Fehlen genereller Strategien bei Konflikten zwischen Verantwortungspflichten; es müssten vielmehr stets Einzelfallbetrachtungen stattfinden, die die Besonderheiten der interpersonellen Beziehungen zwischen Verantwortungsträgern berücksichtigen.

208 Die Definition von ›pragmatisch‹ im *Wörterbuch der philosophischen Begriffe* lautet: »1. Zum Handeln befähigt, praktisch, der Praxis dienend, 2. der Wohlfahrt, dem allg. Nutzen dienend« (Regenbogen & Meyer, 2013, S. 518).

209 Die Bezeichnung ›pragmatisch‹, wie sie im Kontext der Argumentation dieser Forschungsarbeit verwendet wird, ist ausdrücklich nicht auf den sprachphilosophischen Begriff der Pragmatik bezogen, der die Theorie sprachlichen Handelns beschreibt.

Analysen und orientiert sich wesentlich an den Besonderheiten des jeweiligen Anwendungskontextes, indem sie disziplinübergreifend und situativ abwägt.[210] Zur Bedeutung des Begriffs ›pragmatisch‹ existieren verschiedene Interpretationen. Eine der prominentesten stammt von Jürgen Habermas, der den Kern pragmatischer Ethik als wertorientierte Zweckrationalität versteht. Im Rahmen seiner *Erläuterungen zur Diskursethik* (1991a) differenziert Habermas zwischen moralischen, ethischen und pragmatischen Prägungen der Frage ›Was soll ich tun?‹. ›Pragmatisch‹ heißt für ihn so viel wie ›wertorientiert zweckrational‹, d. h. geeignet, ein bestimmtes Ziel zu erreichen:

> Je nach Problemstellung gewinnt also die Frage ›Was soll ich tun?‹ eine pragmatische, ethische oder moralische Bedeutung. In allen Fällen geht es um die Begründung von Entscheidungen zwischen alternativen Handlungsmöglichkeiten; aber pragmatische Aufgaben erfordern einen anderen *Typus von Handlungen*, die entsprechenden Fragen einen anderen *Typus von Antworten* als ethische und moralische. Die wertorientierte Abwägung von Zwecken und die zweckrationale Abwägung von verfügbaren Mitteln dient der vernünftigen Entscheidung darüber, wie wir in die objektive Welt eingreifen müssen, um einen erwünschten Zustand herbeizuführen. (Habermas, 1991b, S. 108, Hervorh. i. Orig.)

Als methodische Mittel stehen im Hinblick auf pragmatische Aufgaben vor allem Abwägungen und Beobachtungen zur Verfügung, um geeignete Strategien zu finden, »die in einfachen Fällen die semantische Form bedingter Imperative haben. Der imperativische Sinn, den sie ausdrücken, läßt sich als ein *relatives Sollen* verstehen.« (Ebd., S. 102, Hervorh. i. Orig.)[211] Kompromisse sind ein inhärentes Merkmal ethisch-politischer Debatten, in deren Zuge es zu einer »Klärung einer kollektiven Identität [kommt], die Raum lassen

---

210 Dies entspricht der von Raters (2013, S. 378–379) skizzierten Ausdifferenzierung der Angewandten Ethik in spezialisierte (Bereichs-)Ethiken wie die Technik- oder Medizinethik.

211 Die imperativische Zweckrationalität gilt dabei nur so lange, wie die zugrundeliegenden Werte an sich nicht fragwürdig sind: »Die Handlungsanweisungen sagen, was man im Hinblick auf ein bestimmtes Problem tun ›soll‹ oder tun ›muß‹, wenn man bestimmte Werte oder Zwecke realisieren will. Sobald freilich die Werte selber problematisch werden, weist die Frage: Was soll ich tun? über den Horizont der Zweckrationalität hinaus.« (Habermas, 1991b, S. 102–103)

muß für die Mannigfaltigkeit individueller Lebensentwürfe.« (Ebd., S. 117) Auch jenseits von diskursethisch geprägten Konzeptionen sind pragmatische Ansätze vertreten. So entwickelt beispielsweise Christoph Hubig (1999), motiviert durch das verstärkte Hervortreten eines ethischen Pluralismus einerseits und als kritische Auseinandersetzung mit Habermas andererseits, eine pragmatisch-provisorische Moral in der Tradition der Klugheitsethik, die auf ›letztbegründende‹ Regeln verzichtet und stets den Einzelfall in den Blick nimmt.

Ohne sich auf eine bestimmte Interpretationsweise festlegen zu wollen, bedeutet ›pragmatisch‹ bezogen auf die praktische Problemstellung dieser Forschungsarbeit in jedem Fall nicht zwangsläufig ›politisch‹, auch wenn es um Fragen von kollektivem Interesse geht. Wie in Kap. 4.2.2 anhand der Positionen von Brändle und Schmidt (2021), Himmelreich (2018) und Keeling (2020) bereits diskutiert wurde, erweist sich die Gestaltung von Unfallalgorithmen durchaus als politisches Problem, das jedoch nicht völlig losgelöst von ethischen Betrachtungen zu bewältigen ist. Vielmehr gehen ethische Begründbarkeit und gesellschaftliche Akzeptanz hier Hand in Hand, die Ethik geht der Politik quasi voraus: »Ethics is not being recommended as a decision procedure, but as an essential resource for making decisions.« (Nagel, 1979b, S. 186). Eine pragmatische Ausrichtung intensiviert diese in der Konstitution einer öffentlichen Vernunft mündende Verbindung schließlich durch das Setzen eines gemeinsamen, praxisdienlichen Orientierungsziels.

Auch tatsächlich beobachtbares Verhalten deutet darauf hin, dass wir in alltäglichen Situationen tatsächlich oft intuitiv pragmatisch handeln, auch wenn wir grundsätzlich die absolute Nicht-Verrechenbarkeit spezifischer Pflichten anerkennen. In konkreten Konfliktsituationen räumen wir anderen Überlegungen Priorität ein, obwohl wir dem menschlichen Leben einzigartigen und absoluten Wert zuschreiben (vgl. Statman, 1995, S. 80). So lassen wir uns als Autofahrer wider besseres Wissen um die Gefahren leicht ablenken, um beispielsweise Nachrichten auf dem Mobiltelefon zu lesen, und nehmen auf diese Weise ein hohes Unfallrisiko in Kauf. In diesen Fällen suggeriert ›absolut‹ also gerade nicht, dass das entsprechende moralische Gebot niemals überwunden werden kann, sondern lediglich, dass mit einer Zurückweisung verbundene Verluste nicht vollständig kompensiert werden können und daher stets bedauerns-

wert sind (vgl. ebd., S. 82). Wenn es darum geht, zwischen Grundrechtsverletzungen abzuwägen, stellt z. B. der Grundsatz der Proportionalität ein zentrales inhaltliches Kriterium nach pragmatischer Lesart dar.[212]

5.4.3 Zwischenergebnis: Argumentative Relevanz der metaethischen Analyse

Die Ergebnisse dieses fünften Kapitels haben in zweierlei Hinsicht eine große Bedeutung für die Gesamtargumentation des Buches. Zum einen wurde offengelegt, dass sich auch aus der metaethischen Struktur des zugrundeliegenden Entscheidungsproblems offene Fragen dahingehend ergeben, wie die Problematik der Nicht-Verrechenbarkeit spezifischer deontologischer Pflichten aus theoretisch-formaler Sicht bewältigt werden kann. Vor diesem Hintergrund wird ein tieferes Verständnis der Problemstellung moralischer Unfalldilemmata ermöglicht, aus dem sich eine ganzheitliche Perspektive ableiten lässt, die im bisherigen Forschungsdiskurs weitgehend unberücksichtigt geblieben ist. Es wurde argumentiert, dass potenzielle Strategien zur Bewältigung von Unfalldilemmata durch die metaethische Struktur derselben dahingehend limitiert werden, dass sie sich nur pragmatisch entscheiden lassen, i. e. dass sie anhand ihrer Einbettung in praktische gesellschaftliche Kontexte evaluiert werden müssen. Letztere sind, wie in Kap. 4.2 gezeigt, wesentlich durch eine gesellschaftlich-soziale Dimension sowie risikobehaftete

212 Auch wenn in dieser Forschungsarbeit ausdrücklich auf ethischer Basis argumentiert wird, kann an dieser Stelle ein Blick auf den verwandten Diskurs verfassungsrechtlicher Abwägungen von Grundrechten zunächst hilfreich sein, um sich den komplexeren (meta-)ethischen Ansätzen anzunähern. Grundrechte sind zwar vorbehaltlos, aber nicht schrankenlos gewährleistet; zum Schutz individueller Freiheit und öffentlicher Interessen müssen sie (begründet) einschränkbar sein, wenn z. B. Grundrechte kollidieren. Jedes Grundrecht verfügt über spezifische immanente Schranken, durch die Eingriffe in Form verfassungsmäßiger Gesetze wirksam werden können. Verfassungsrechtlich gerechtfertigt sind Grundrechtseingriffe aufgrund von Abwägungen dann, wenn sie verhältnismäßig sind. Das hier implizierte Konzept der Verhältnismäßigkeit umfasst die Grundsätze der Geeignetheit, Erforderlichkeit und Verhältnismäßigkeit im engeren Sinne. Letzteres wird auch als Proportionalität bezeichnet und erfordert, dass Ziel und Schwere eines Eingriffs in einem angemessenen Verhältnis stehen, i. e. zumutbar sind (vgl. Degenhart, 2023).

Entscheidungen geprägt. Um derartige Dilemmata entscheiden zu können, sind Abwägungen unumgänglich, die dem theoretischen Problem gerecht werden, zugleich aber auch der (praktischen) Sache dienen. Das in Kap. 4 erarbeitete Argument, welches die erste These der Forschungsarbeit stützt, wird auf diese Weise vervollständigt.

Zum anderen wurde erörtert, dass das Fehlen systematischer Strategien für unlösbare Konflikte zwischen inkommensurablen Werten bedeutende Implikationen für eine ethische Auseinandersetzung mit der Gestaltung von Unfallalgorithmen hat. Die Problematik bewegt sich im Spannungsfeld zwischen der formalen Unlösbarkeit einerseits und der praktischen Notwendigkeit einer Regulierung andererseits; begründbare Entscheidungsstrategien müssen beiden Aspekten Rechnung tragen. Die Sachbezogenheit auf die Spezifika von Dilemma-Szenarien ist dabei der entscheidende Blickwinkel, von dem ausgehend offene Fragen im Sinne eines pragmatischen Ansatzes beantwortet werden können. Letzterer impliziert u. a. eine explizite ethische Auseinandersetzung mit dem Unsicherheitsaspekt hinsichtlich tatsächlich zu erwartender Schäden, der in diesem Buch im Fokus steht. Eine solche kann die Risikoethik leisten, welche sich mit der moralischen Bewertung unsicherer Handlungsfolgen beschäftigt und im weiteren Verlauf der Forschungsarbeit als alternativer Problemzugang zur Thematik moralischer Dilemma-Szenarien vorgeschlagen wird. Hierbei liefern die Ergebnisse des fünften Kapitels ein entscheidendes Puzzleteil zur Vorbereitung der zweiten These, die besagt, dass sich auf der Grundlage einer risikoethischen Interpretation zentrale Fragen des Anwendungsproblems klären und neue Entscheidungsperspektiven entwickeln lassen. Eine solche wird im nun folgenden dritten Teil des Buches präsentiert.

III.
Risikoethische Auseinandersetzung: Entwurf eines alternativen Problemzugangs

In diesem dritten Teil des Buches wird auf der Basis der bisherigen Ergebnisse schließlich ein alternativer Zugang zum Anwendungsproblem entworfen. Im Zuge einer risikoethischen Auseinandersetzung mit der spezifischen Problematik von Unfallalgorithmen wird das vierte Teilziel der Forschungsarbeit realisiert. Dabei liegt die zweite These zugrunde, dass sich im Rahmen eines risikoethischen Zugangs bis dato ungeklärte, zentrale ethische Fragen klären lassen, die, gemäß der ersten These, unter bisher dominanten Forschungszugängen offengeblieben sind. Zudem versteht sich der vorgelegte risikoethische Entwurf als Antwort auf die metaethische Darstellung aus Kap. 5, derzufolge sich pragmatische Entscheidungsstrategien für Unfalldilemmata als vielversprechend herausgestellt haben. Beide Aspekte werden im Folgenden zusammengeführt, um normative Implikationen für neue, risikoethisch-basierte Entscheidungsperspektiven zu ermitteln.

In Kap. 6 werden zunächst historische, begriffliche und konzeptionelle Grundlagen gelegt und reflektiert, die für das Verständnis der erarbeiteten risikoethischen Darstellung essenziell sind. Zu die-

sem Zweck erfolgt in Kap. 6.1 eine systematische wissenschaftliche Einordnung der Risikoethik als Teilgebiet der Angewandten Ethik, das sich mit der moralischen Bewertung von Handlungen beschäftigt, die hinsichtlich ihrer Folgen risikobehaftet sind. Ausgehend von einem Abriss der historischen Evolution der Risikoethik als Synthese verschiedener Diskurse der Risikoforschung wird die Frage nach der Zulässigkeit von Risikoübertragungen als eigenständiges Problem der Moraltheorie begründet. In Kap. 6.2 werden risikoethische Grundbegriffe erläutert sowie verschiedene Risikosituationen und -konstellationen systematisch vorgestellt, auf die im weiteren Verlauf des dritten Teils zurückgegriffen wird. Es werden Grundfragen und Problemfelder der Risikoethik motiviert und kategorisiert. Darauf aufbauend werden die Grundzüge einer rationalen Risikopraxis, die eine Verzahnung von Risikoethik und Risikopolitik darstellt, in Kap. 6.3 geschildert. Zentrale risikopraktische Paradigmen und Entscheidungsprinzipien werden überblicksartig skizziert und kritisch beleuchtet. Anschließend wird vor dem Hintergrund der ethischen Kritik an einer konsequentialistischen Grundorientierung verdeutlicht, weshalb eine rationale Risikopraxis angesichts der spezifischen Problematik von Unfallalgorithmen unglaubwürdig erscheint.

In Kap. 7 werden schließlich die Grundzüge einer risikoethischen Auseinandersetzung systematisch entwickelt, die – verstanden als Skizze einer kohärenten Risikopraxis – einen alternativen Zugang zum Anwendungsproblem ermöglicht. Einführend wird in Kap. 7.1 die Frage nach der Zulässigkeit risikobehafteter Handlungen im Spannungsfeld zwischen Risikoakzeptanz und Risikoakzeptabilität verortet. Es werden Ansätze bisheriger Forschung evaluiert, die sowohl implizit als auch explizit risikoethisch vorgehen. Bezugnehmend auf die dabei identifizierte Forschungslücke werden Gegenstand und Ziele des im Folgenden vorgelegten risikoethischen Entwurfs eruiert. Kap. 7.2 präsentiert eine Analyse der spezifischen Risikokonstellationen, welche in Dilemma-Szenarien des autonomen Fahrens denkbar sind und auf einer Interpretation von Unfallalgorithmen als risikoethisches Verteilungsproblem beruhen. Diese werden anhand zentraler risikoethischer Konzepte und etablierter Kriterien der Risikoakzeptabilität kritisch diskutiert, bevor schließlich zugunsten einer deontologischen Perspektive argumentiert wird. In Kap. 7.3 werden nach dem Vorbild der Konzeption einer kohärenten Risikopraxis von Nida-Rümelin et al. (2012) sodann die Eckpfeiler

einer deontologischen Risikoethik für Unfalldilemmata entworfen. Dabei lassen sich normative Implikationen freilegen, die den Gestaltungsauftrag ethischer Unfallalgorithmen als risikoethisches Optimierungsproblem adressieren, dem durch deontologische Grenzkriterien unverhandelbare Beschränkungen auferlegt sind.

6. Theoretische Grundlagen, begriffliche Reflexion und Ziele einer Risikoethik für Unfalldilemmata

6.1 Systematische wissenschaftliche Einordnung der Risikoethik

Kontingente Folgen stellen einen der zentralen Ausgangspunkte der Risikoforschung dar. Diese ist in vielerlei Hinsicht eine »Grenzwissenschaft« (Bohle & Pohl, 2014); sie bewegt sich im Spannungsfeld von dem, was wir wissen, und dem, was wir nicht sicher oder gar nicht wissen (können). Dabei stellt sie sich komplexen Fragen, die die Grenzen wissenschaftlicher Disziplinen aus empirischen Wissenschaften einerseits und Geisteswissenschaften andererseits transzendieren. Die inhärente Interdisziplinarität vor allem praktischer Risikofragen ist für eine definitorische Vielfalt verantwortlich, in deren Rahmen Begriffe und Konzepte vor dem Hintergrund der Methoden und Weltsichten der jeweiligen Disziplinen interpretiert werden. Der für die Fragestellung dieser Arbeit relevante Risikobegriff bewegt sich an der Schnittstelle von sozialwissenschaftlichem, technischem und ethischem Risikodiskurs. Als Einführung in die risikoethische Perspektive, die in diesem dritten Teil des Buches entwickelt wird, soll daher im Folgenden zunächst der Zusammenhang zwischen diesen dreien kurz skizziert werden. Dabei wird auf Forschungsliteratur aus der Angewandten Ethik, aber auch der Soziologie sowie den Rechts- und Technikwissenschaften zurückgegriffen. Es stehen jene Aspekte im Fokus, die als Grundlage für den zu entwickelnden risikoethischen Entwurf relevant sind.

6.1.1 Sozialwissenschaftlicher und sozio-technischer Diskurs

Das Leben in modernen Gesellschaften ist geprägt durch ein Netzwerk von Abhängigkeiten und das Zusammenwirken komplexer sozialer, ökonomischer und ökologischer Systeme. Das Ergebnis

unserer Handlungen wird durch viele Faktoren bestimmt, die sich oftmals nicht voraussehen lassen. Analog zum gesellschaftlichen Wandel durch transformative Kräfte wie die fortschreitende Globalisierung und Digitalisierung erlebt auch der sozialwissenschaftliche Risikodiskurs eine Dynamik, in der zentrale Begriffe wie ›Sicherheit‹, ›Unsicherheit‹ und ›Risiko‹ stetig neu zu definieren sind. Das moderne Sicherheits- und Risikoverständnis, das in Deutschland seine Ursprünge in den 1980er-Jahren hat, ist bestimmt von einer breiten gesellschaftlichen Verunsicherung. Diese ist einerseits auf den Einfluss allgegenwärtiger Großrisiken auf das alltägliche Leben und andererseits auf eine grundlegende Unsicherheit hinsichtlich der Folgen eigener Handlungen zurückzuführen (vgl. Lippert et al., 2013). In seinem prominenten Buch *Risikogesellschaft. Auf dem Weg in eine andere Moderne*, das 1986 kurz nach der Atomreaktorkatastrophe von Tschernobyl erschien und als Klassiker der modernen Soziologie gilt, beschreibt der Soziologe Ulrich Beck einen radikalen Bruch als Kennzeichen der Moderne. Durch den wachsenden (technischen) Fortschritt sieht sich die Industriegesellschaft zunehmend bedrohlichen Risiken ausgesetzt, die statusunabhängig alle Individuen betreffen:

> In der fortgeschrittenen Moderne geht die gesellschaftliche Produktion von Reichtum systematisch einher mit der gesellschaftlichen Produktion von Risiken. [...] Die Verteilungsprobleme und -konflikte der Mangelgesellschaft [werden] überlagert durch die Probleme und Konflikte, die aus der [...] Verteilung wissenschaftlich-technisch produzierter Risiken entstehen. (Beck, 1986, S. 25)

Die moderne Gesellschaft ist im Wesentlichen eine Risikogesellschaft. Ausgehend von einer kritischen Auseinandersetzung mit Becks Thesen verzeichnet die sozialwissenschaftliche Risikoforschung der Gegenwart eine zunehmende Hinwendung zu einem kultursoziologischen Risikokonstruktivismus (vgl. Krohn & Krücken, 1993), der auf einem Verständnis von Sicherheit und Unsicherheit als »gesellschaftliche Konstruktionen« (Bonß, 1997, S. 21) fußt.[213] Seit

213 Die zentrale Reflexion des sozialwissenschaftlichen Risikodiskurses bewegt sich im Spannungsfeld zwischen risiko-objektivistischen und risiko-konstruktivistischen Positionen. Ein Meilenstein in diesem Kontext ist die Monografie von Douglas und Wildavsky (1983). Allerdings lässt sich weder die objektivistische noch die konstruktivistische Position final theoretisch klären, weshalb

seiner Initiierung erfährt der sozialwissenschaftliche Risikodiskurs, in dessen Rahmen die Klärung der Konzepte ›Sicherheit‹ und ›Risiko‹ sowohl als soziologisches als auch sozialpolitisches Problem betrachtet wird, eine kontinuierliche Politisierung. Gegenwärtige Debatten sind zudem stark techniksoziologisch geprägt (vgl. Nassehi, 2019). Dominiert wird die Risikoforschung durch das sogenannte psychometrische Paradigma, welches auf der Annahme beruht, dass Risiken subjektiv sind und ihre Wahrnehmung mit qualitativen Gefahrenmerkmalen verknüpft ist.[214] Für den Psychologen Paul Slovic sind Risiken subjektive, soziale Konstrukte:

> One of the most important assumptions [...] is that risk is inherently subjective. Risk does not exist ›out there‹, independent of our minds and cultures, waiting to be measured. Human beings have invented the concept risk to help them understand and cope with the dangers and uncertainties of life. There is no such thing as ›real risk‹ or ›objective risk‹. (Slovic, 1992, S. 119)

In den Technikwissenschaften hingegen steht der Begriff des Risikos im Kontext der Sicherheitstechnik. Er bezeichnet eine numerische, stochastisch ermittelte Kennzahl, die das Gefahrenausmaß quantifiziert, welches von einem technischen System als Ganzes ausgeht. Die dominierende technische Zielgröße ist dabei der Erwartungswert: Risiko ist definiert als das Produkt aus der Eintrittswahrscheinlichkeit eines Schaden verursachenden Ereignisses und dem zu erwartenden Schadensausmaß. Da völlige Risikovermeidung ein utopisches Ziel darstellt, impliziert der technische Sicherheitsbegriff, dass jedes System eine Restgefahr birgt, wobei das Risiko unter dem Grenzrisiko liegt. Dieses wiederum gibt das maximal vertretbare Risiko an und wird meist implizit durch Richtlinien und spezifische sicherheitstechnische Regeln definiert. Für den Kontext dieser Arbeit greift insbesondere die Norm ISO 26262, die Anforderungen an die

---

sich pragmatische Ansätze derzeit auf dem Vormarsch befinden (vgl. Krohn & Krücken, 1993). Für traditionelle Grundpositionen des sozialwissenschaftlichen Risikodiskurses siehe z. B. Bechmann (1993), Bonß (1996), Nassehi (1997a, 1997b, 1997c) und Luhmann (1991b).

214 Das psychometrische Paradigma hat durch seine empirisch-quantitative Ausrichtung und die Verwendung multivariater statistischer Verfahren zur Analyse kognitiver Urteilsstrukturen die Risikosoziologie nachhaltig beeinflusst. Zur psychometrischen Methode siehe z. B. Jungermann und Slovic (1997), Rohrmann und Renn (2000) sowie Slovic (1992).

funktionale Sicherheit von Fahrzeugen stellt. Ein sicherer Zustand im Sinne dieser Norm liegt dann vor, wenn sich das System in einem Betriebsmodus befindet, in dem kein unzumutbares Risiko besteht, d. h. dieses unterhalb einer gesellschaftlich akzeptierten Schwelle liegt (vgl. International Organisation for Standardisation, 2018, Teil I, 1.102 bzw. 136; Reschka, 2015, S. 490–496).

6.1.2 Von der Technikanalyse zur Technikbewertung: Technikfolgenabschätzung und technikethischer Diskurs

In den letzten vierzig Jahren hat die Risikoforschung insbesondere in Bezug auf technologiegetriebene Risiken Fahrt aufgenommen. Die zunehmende gesellschaftliche Durchdringung mit Großtechnologien bringt qualitativ neuartige Risiken mit sich, die diverse Disziplinen vor komplexe Herausforderungen stellen.[215] Ausgelöst durch eine zunehmende Technikkritik im Zuge des eingeläuteten Endes des Fortschrittsglaubens beginnen theoretische Reflexionen über Technik im letzten Drittel des 20. Jahrhunderts über die Grenzen traditioneller Ingenieurwissenschaften hinauszureichen und Geistes- und Sozialwissenschaften in einen öffentlichen Diskurs einzubeziehen (vgl. Ott, 2005, S. 576–577; Ropohl, 1996, S. 24). Während informelle Technikbewertung durch Einzelne schon seit der frühesten Verwendung technischer Artefakte stattfindet, kam im Zusammenhang mit der Neuorientierung der Risikoforschung auch das Bedürfnis nach formellen und systematischen Vorgehensweisen auf (vgl. Ropohl, 1996, S. 160–164). Die Bewertung von Technik*folgen* bildet seitdem den thematischen Schwerpunkt der Technikbewertung, die sich zunächst in Form der Technikfolgenabschätzung (TFA) organisiert hat. Diese lässt sich als heterogenes Forschungsfeld charakterisieren, das an der Schnittstelle von Techniksoziologie und -philosophie liegt und primär sozialphilosophische Diskurse zum Gegenstand hat. Ziel der TFA ist es, politische Handlungsempfehlungen oder Richtlinien für die Vermeidung von Risiken einerseits und

215 Konstitutive Faktoren für diese Entwicklung sind u. a. Globalisierung, Komplexität, Ausmaß, Irreversibilität und Langzeitwirkungen von Risiken (vgl. Banse, 1996, S. 33).

die effektive Nutzung der Potenziale neuer Technik andererseits zu entwickeln.[216]

Neben ingenieurwissenschaftlichen Herausforderungen in der Implementierung sicherer Systeme und der Analyse von deren soziologischen Effekten treten zunehmend auch ethische Fragen in den Vordergrund: Sind die Veränderungen, die durch technische Innovation entstehen, tatsächlich wünschenswert? Sollen wir alles, was technisch möglich ist, auch umsetzen? Wie kann eine Balance zwischen technikinduzierten Vorteilen und Risiken geschaffen werden? Fragen der Risikoakzeptabilität beinhalten ethische Werturteile und können nicht allein auf der Basis von Zahlen entschieden werden (vgl. Hansson, 1993). Das Spannungsfeld zwischen Fortschrittszielen und nicht-intendierten Folgen generiert eine grundlegende Technikambivalenz, die es notwendig macht, moderne Technik auch aus ethischer Sicht einer kritischen Bewertung zu unterziehen (vgl. Jonas, 1993, S. 81–82).

Die Auseinandersetzung mit moralischen Fragen, die sich im Zuge der Entwicklung, des Gebrauchs und der Entsorgung neuer Technologien stellen, steht im Zentrum der Technikethik. Deren Ursprünge in Deutschland lassen sich zurückverfolgen in die 1960er-Jahre. Im Zuge der sogenannten Technokratiedebatte (vgl. z. B. Habermas, 1968; Schelsky, 1961) befassten sich Sozialwissenschaftler und Philosophen erstmals mit normativen Fragen der Technik, bevor der Diskurs in den 1970er-Jahren im Zuge der ökologischen Bewegung schließlich an Intensität zunahm.[217] Die moderne Technikethik betrachtet Technik nicht länger als isolierte Artefakte, sondern begreift sie als integralen Bestandteil einer Gesellschaft, als Teil eines sozio-technischen Zusammenhangs (vgl. Ropohl, 1979).[218] Ihren endgültigen Durchbruch erlebte die Technikethik infolge der viel

216 Zur systematischen Einführung in die TFA siehe Grunwald (2002).

217 Zur Einführung in Definition, Gegenstandsbereich, historische Entwicklung und Bewertung der Technikethik siehe Grunwald (2013, S. 3–8, 2016, S. 26). Für traditionelle Grundpositionen der Technikethik siehe Höffe (1993), Lenk und Maring (1998), Lenk und Ropohl (1987), Ott (2005), Ropohl (1996) sowie Skorupinski und Ott (2000).

218 Als zentrales konstitutives Element in der Begründung und Herausbildung der Technikethik als wissenschaftliches Teilgebiet ist die Zurückweisung der Wertneutralitätsthese zu nennen. Diese galt noch bis in die 1990er-Jahre und besagt, dass Technik an sich wertneutral ist und sich moralische Probleme erst aus ihrem Gebrauch ergeben (vgl. Grunwald, 2013, S. 2). Mit der zuneh-

beachteten Technikkritik von Hans Jonas aus dem Jahre 1979, dessen ethisches Hauptwerk *Das Prinzip Verantwortung* als erste umfassende philosophische Antwort auf die Gefahren der modernen Technik gilt. In seinem wirkungsmächtigen Entwurf einer ›Zukunftsethik‹ rückt er die Bedeutung langfristiger Folgen von Technik und damit die Notwendigkeit eines nachhaltigen Handelns in den Blick. Dabei betont er, dass technische Machbarkeit einen tiefgreifenden Widerspruch zu ethischer Vertretbarkeit darstellt:

> Der endgültig entfesselte Prometheus, dem die Wissenschaft nie gekannte Kräfte und die Wirtschaft den rastlosen Antrieb gibt, ruft nach einer Ethik, die durch freiwillige Zügel seine Macht davor zurückhält, dem Menschen zum Unheil zu werden. […] Keine überlieferte Ethik belehrt uns daher über die Normen von ›Gut‹ und ›Böse‹, denen die ganz neuen Modalitäten der Macht und ihrer möglichen Schöpfungen zu unterstellen sind. Das Neuland kollektiver Praxis, das wir mit der Hochtechnologie betreten haben, ist für die ethische Theorie noch ein Niemandsland. […] Was kann als Kompaß dienen? Die vorausgedachte Gefahr selber! (Ebd., S. 7)

In der Folge reifte das Verständnis, dass der Technikethik eine normative Orientierungs- und Gestaltungsfunktion auch in frühen Phasen der Technologieentwicklung zukommen muss; die Technikethik darf nicht länger nur »Reparaturethik« (Mittelstraß, 1991) sein. In diesem Sinne diskutiert sie einerseits spezielle Fragestellungen aus menschlichen Handlungskontexten, wobei sie in einer pragmati-

---

menden Komplexität technischer – vor allem autonomer – Systeme ist die Werthaltigkeit von Technik allerdings neu zu hinterfragen (vgl. Hubig, 1993, S. 21). Dazu Ropohl (1996, S. 159): »Herstellung und Verwendung technischer Sachsysteme setzen sich […] aus einer Vielzahl von menschlichen Handlungen zusammen, und menschliches Handeln ist immer an Zielen orientiert, die letzten Endes auf Werte Bezug nehmen.« Technische Systeme sind nicht länger nur Werkzeuge, sondern transportieren immer Werte, beispielsweise die soziale oder ökologische Verträglichkeit bzw. Robustheit, Privatsphäre – oder sie besitzen regulativen Charakter. Dadurch gestalten sie unsere menschlichen Lebensformen aktiv mit (vgl. Grunwald, 2016, S. 27–28; Hubig, 1993, S. 55–58). Im Kontext des Forschungsfelds *Values in Design* »werden Artefakte und Technologien als inhärent moralisch vorprogrammiert verstanden, insofern sie bestimmte moralische Werte und Normen fördern oder behindern.« (Simon, 2016, S. 359) Nach Brey (2010) konstituieren Werte sozio-technische Realitäten, die sich durch Freiheitsgrade in der Nutzung eine gewisse Flexibilität bewahren.

schen Relation zu praktisch auftretenden Problemen steht und daher als Teilbereich der Angewandten Ethik zu sehen ist. Gegenwärtig werden technikethische Diskurse vor allem in den Kontext von politik- und sozialwissenschaftlichen Fragen gestellt:

> Im wissenschaftlich-technischen Fortschritt werden neue Handlungsoptionen entwickelt und Eingriffsmöglichkeiten des Menschen verfügbar gemacht, wodurch vielfach ethische Fragen aufgeworfen und traditionelle Selbstverständlichkeiten aufgelöst werden. Aufgabe der Technikethik ist es, die normativen Hintergründe dieser Fragen in der Gestaltung des wissenschaftlich-technischen Fortschritts und im Umgang mit seinen Folgen nach Maßstäben rationaler Argumentation zu rekonstruieren, um auf diese Weise zu ethisch reflektierten und verantwortbaren Entscheidungen beizutragen. (Grunwald, 2016, S. 25)

Andererseits widmet sich die Technikethik ebenso Fragen aus dem Gegenstandsbereich der allgemeinen Ethik, wobei sie in diesem Sinne eher eine hermeneutische Aufklärungsfunktion innehat, die ethischen Strategien für konkrete Anwendungskontexte vorausgeht (vgl. Grunwald, 2013, S. 3–4).[219]

Der gegenwärtige wissenschaftlich-technische Fortschritt wird von einer stetigen Erweiterung der Grenzen des menschlichen Macht- und Handlungsbereichs begleitet. Mit wachsenden Eingriffsmöglichkeiten steigt jedoch auch die Entscheidungsnotwendigkeit dort, wo der Mensch über wenig oder unzureichendes Wissen verfügt. Spätestens seit Hans Jonas' philosophischem Meilenstein der Technikethik ist die Bedeutung kontingenter und langfristiger Folgen von Technik und damit die Notwendigkeit eines nachhaltigen Handelns ins gesellschaftliche Bewusstsein vorgedrungen:

> Eben diese Ungewißheit nun aber, welche die ethische Einsicht für die [...] Zukunftsverantwortung unwirksam zu machen droht und natürlich nicht auf die Unheilsprophetie beschränkt ist, muß selber in die ethische Theorie einbezogen und in ihr zum Anlaß eines neuen Grundsatzes genommen werden, der nun seinerseits als praktische Vorschrift wirksam werden kann. Es ist die Vorschrift, primitiv gesagt, daß der Unheilsprophezeiung mehr Gehör zu geben ist als der Heilsprophezeiung. (Jonas, 1979, S. 70)

219 In diesem Sinne setzt sich die Technikethik mit anthropologischen, natur- und technikphilosophischen Fragen auseinander, die technikinduzierte Veränderungen des menschlichen Selbstverständnisses oder allgemeine Folgen einer zunehmenden Technologisierung der Gesellschaft berühren.

Die in diesem Kontext in den Vordergrund getretene Technikethik nimmt eine antizipative Ausrichtung ein, indem sie sich der proaktiven Gestaltung von Technik verschreibt. In diesem Sinne muss sie sich neben gewünschten Effekten auch mit nicht-intendierten Folgen befassen, welche ex ante mit Unsicherheiten behaftet sind (vgl. Grunwald, 2013, S. 5); auch eine »Beurteilung des Nichtwissens [muss] in die Folgendiskussion einfließen« (Decker, 2013, S. 37). Wie Höhn (1996, S. 29) konstatiert, werden »Technikdebatten […] heute durchgängig als Risikodebatten geführt, in deren Zentrum Fragen der technisch-ökologischen Selbstgefährdung moderner Gesellschaften stehen.« Treten die Folgen einer Handlung nur noch mit gewisser Wahrscheinlichkeit ein, sind sie kontingent und es reicht nicht mehr aus, die entsprechende Handlung allein anhand ihrer Konsequenzen oder ihrer zugrundeliegenden Motivation zu bewerten. Rehmann-Sutter (1998, S. 48) fasst treffend zusammen: »Das Zufügen eines Schadens hat die Wirklichkeit des Schadens zur Folge; das Auferlegen eines Risikos hat die Möglichkeit des Schadens zur Folge.«

Angesichts kontingenter Handlungsfolgen ist es notwendig, Kants ethische Grundfrage ›Was soll ich tun?‹ umzuformulieren: ›Welchen Risiken darf ich mich und andere aussetzen?‹. Der Versuch, die Frage nach der Zulässigkeit von Risikoübertragungen unter diejenige nach erlaubten direkten Schädigungen zu subsumieren, ist jedoch weder pragmatisch plausibel noch moraltheoretisch gerechtfertigt. Zwei ausgewählte Argumente aus der einschlägigen risikoethischen Literatur verdeutlichen dies. So geht Thomson (1985a, S. 127) zunächst von der positiven Formulierung erlaubter Handlungen aus: Wenn es gerechtfertigt ist, einer Person einen Schaden direkt zuzufügen, dann ist es auch gerechtfertigt, sie einem entsprechenden Schadensrisiko auszusetzen. Auch wenn diese Faustregel schlüssig erscheint, so ist sie doch wenig hilfreich, wenn es um das Verbot bzw. die Zulässigkeit risikobehafteter Handlungen geht. Denn um mit ihrer Hilfe eine Aussage über Risikoübertragungen treffen zu können, müssten zunächst die Bedingungen definiert werden, unter denen das Zufügen eines direkten Schadens erlaubt ist (vgl. Nida-Rümelin et al., 2012, S. 32). Dies stellt jedoch eine der schwierigsten Fragestellungen in der Moraltheorie überhaupt dar. Auch die Argumentation von McCarthy (1997, S. 205–208) legt nahe, dass zwischen Problemstellungen hypothetischer und direkter Schäden

strikt zu trennen ist. Er formuliert das Recht, keinen Schaden durch andere zu erleiden, als separate (Schadens-)These. Im Rahmen einer ausführlichen Diskussion, inwiefern infolge einer Risikoübertragung die Pflicht zur Kompensation derselben besteht, zeigt er, dass beide Thesen grundlegend verschiedene ethische Problemstellungen zum Gegenstand haben:

> [...] any plausible theory of rights will assign to us something at least very much like the right that others not harm us. Call the claim that we have exactly that right the Harm Thesis, and suppose it is correct. Suppose I impose a risk of harm on you, and this causes you to be harmed. I have thereby infringed your right that I not impose a risk of harm on you, and I have also infringed that I not harm you. (Ebd., S. 224)

Für den Fall, dass sich ein Schadensrisiko tatsächlich materialisiert, wäre der Verursacher demnach zur Kompensation beider Rechtsverletzungen verpflichtet, und damit zweifach. Das mutet unplausibel an und deutet darauf hin, dass die Rechtsverletzung, die die Kompensationspflicht begründet, bereits in der Risikoexposition besteht. Risikoübertragungen und direkte Schäden sind daher voneinander unabhängige moralische Gegenstände.

Beide Argumente zeigen letztlich, dass die Frage nach der Zulässigkeit von Risikoübertragungen ein eigenständiges Problem der Moraltheorie begründet, das nicht auf die moralische Bewertung direkter Schäden rückführbar ist. Derartigen Problemstellungen widmet sich die Risikoethik, die sich als Teilbereich der Angewandten Ethik mit der moralischen Bewertung von unsicheren und risikobehafteten Handlungen im Kontext gesellschaftlicher Risiken beschäftigt. Ihre Natur ist dabei inhärent pragmatisch; im Hinblick auf konkrete Anwendungsfragen setzt sie sich mit den ethischen Kriterien der Rechtfertigung zweckmäßiger Ansätze der Risikopraxis auseinander. Sie konzentriert sich darauf, praktische Lösungen und Entscheidungen im Umgang mit Risiken zu finden, die uns auch in komplexen Problemlagen handlungsfähig machen. Bei der Abwägung von Schaden und Nutzen geht sie stets kontextbezogen und interdisziplinär vor. Grundlage jeder risikoethischen Betrachtung ist die Annahme, dass Handlungen, die in Risikosituationen zu einem potenziellen Schaden führen, dennoch ex ante moralisch zulässig sein können – nämlich genau dann, wenn Umstände vorliegen, unter denen es aus ethischer Sicht vertretbar oder gar wünschenswert

erscheint, ein spezifisches Risiko einzugehen. Um diese Umstände konkreter bestimmen zu können, werden im Folgenden zunächst zentrale Begrifflichkeiten, Konzepte, Themen und Fragen erläutert, die zum grundlegenden Handwerkszeug der Risikoethik zählen.

6.2 Risikoethische Grundlagen und Begriffe

6.2.1 Risikoethische Grundbegriffe: Unsicherheit, Ungewissheit und Risiko

Das Konzept des Risikos ist im Allgemeinen definiert durch zwei Risikovariablen: Konsequenz und Eintrittswahrscheinlichkeit. In den Diskursen der Risikoforschung wird der Begriff ›Risiko‹ nicht einheitlich gebraucht.[220] Vielen risikoethischen Abhandlungen wird daher ein Überblick über Begriffsdefinitionen und Verwendungsweisen des Risikokonzepts in relevanten Debatten vorangestellt.[221] Ebenso heterogen wie der Risikobegriff wird auch der Terminus der Unsicherheit verwendet. Unsicherheit in Bezug auf Handlungsfolgen ist zunächst kein exklusives Merkmal von Risikosituationen, denn aufgrund »der epistemischen und naturalen Unterbestimmtheit der Umweltbedingungen menschlicher Praxis« (Nida-Rümelin et al., 2012, S. 9) ist die Zukunft immer unsicher. Doch wieso wird Unsicherheit in manchen Entscheidungssituationen problematisiert und in anderen nicht? Die Antwort liegt in der qualitativen Beschaffenheit der jeweiligen Unsicherheit und ihren Implikationen für die ethische Bewertung einer Handlung. Um diese systematisch zu untersuchen, ist die Klärung relevanter Begrifflichkeiten notwendig.

Ein häufiger Ansatz zur Differenzierung von Risiko- und Unsicherheitsbegriff stellt diese in eine hierarchische Relation zueinan-

220 Für eine überblicksartige Darstellung verbreiteter Verwendungsweisen siehe Hansson (2005, S. 7–8).

221 Die in diesem Unterkapitel skizzierten risikoethischen Grundlagen, theoretischen Elemente sowie Termini und Begrifflichkeiten stützen sich im Wesentlichen auf die Darstellungen in deutschsprachigen Standardwerken der Risikoethik wie beispielsweise Nida-Rümelin (2005), Nida-Rümelin et al. (2012), Nida-Rümelin und Schulenburg (2013) sowie Rath (2011).

der.[222] So schlägt z. B. Harsanyi (1977a, S. 320) vor, dass man Situationen, in denen unvollständige Informationen über Eintrittswahrscheinlichkeiten vorliegen, als unsicher bezeichnen sollte. Liegen hingegen ausreichend bekannte Wahrscheinlichkeiten vor, spricht man von Risiko im Sinne messbarer Unsicherheit (vgl. Hansson, 2005, S. 8, 2009, S. 426). Im Falle, dass die Informationen sowohl über Eintrittswahrscheinlichkeiten als auch über Konsequenzen unvollständig sind, liegt Ungewissheit vor (vgl. Rath, 2011, S. 24). Auch Nida-Rümelin et al. (2012, S. 9–10) entwerfen Unsicherheit als ein Kontinuum epistemischer Zustände in Bezug auf die Abschätzbarkeit von Eintrittswahrscheinlichkeiten möglicher Folgen, das sich zwischen zwei antonymen Extremen erstreckt: Auf der einen Seite befindet sich das »reine Risiko«, bei dem exakte Erwartungswerte gebildet werden können.[223] Dem gegenüber liegt ein Zustand, in dem keinerlei begründete Aussagen über Eintrittswahrscheinlichkeiten möglich sind – die Ungewissheit.[224] Beide Extreme sind fiktiv, denn meistens sind wenigstens Schätzungen auf der Basis von begründeten Annahmen möglich.

Ausgehend von dieser Begriffsklärung lässt sich nun auch der Risikobegriff definitorisch präzisieren. Goodall (2016b, S. 31) definiert Risiko allgemein als »the magnitude of misfortune associated with the feared event multiplied by its likelihood.« In der Regel wird zwischen einem allgemeinen und einem spezifischen Risikobegriff differenziert, die jedoch unterschiedlich weit gefasst werden. Eine Verwendung im umfassenden Sinne findet sich z. B. bei Nida-Rümelin et al. (2012, S. 5), die risikobehaftete Entscheidungssituationen dadurch definieren, dass »eine mögliche Handlung ex ante, also

222 Grundlage dieser Ansätze ist die entscheidungstheoretische Unterscheidung des Wirtschaftswissenschaftlers Frank Knight (1921). Er beschreibt Unsicherheit als Konzept mit zwei Ausprägungen, wobei ›Risiko‹ den Zustand vollständig messbarer Variablen bezeichnet, während ›uncertainty‹ das nicht-messbare Pendant darstellt. Letzteres kann sowohl kognitiv als auch sozial bedingt sein.

223 Dies entspricht der in Fußnote Nr. 227 beschriebenen zweiten Form eines verengten Risikobegriffs, der hier als Spezialfall unter dem umfassenden Risikobegriff subsumiert wird.

224 Zusätzlich zu unsicheren Wahrscheinlichkeiten können auch Wissensmängel in Bezug auf qualitative Merkmale der Folgen selbst existieren; in diesem Fall spricht man von »vollständiger Ungewissheit« (Nida-Rümelin et al., 2012, S. 9).

zum Entscheidungszeitpunkt, zu mindestens zwei verschiedenen Konsequenzen führen kann, wobei ex post nur eine dieser möglichen Konsequenzen tatsächlich eintreten kann.« Dabei kann jeder möglichen Konsequenz prinzipiell eine positive Eintrittswahrscheinlichkeit zugeordnet werden. Weitere begriffliche Differenzierungen nimmt Rath (2011, S. 18–20) in Bezug auf die beiden Risikovariablen vor. Können allen relevanten Variablen valide numerische Werte zugeordnet werden, lässt sich die Eintrittswahrscheinlichkeit objektiv und exakt bestimmen. Ist dagegen ein Wert einer Variablen unsicher und muss dahingehend eine Annahme getroffen werden, so spricht man von subjektiver Wahrscheinlichkeit.[225] Zudem unterscheidet Rath zwischen sicherer und erwarteter Konsequenz, wobei Letztere dann vorliegt, wenn die verfügbaren Informationen unvollständig sind. Anstelle einer Differenzierung verschiedener Risikobegriffe betont Hansson (2009, S. 424) zwei Gemeinsamkeiten aller Risikokonzepte: Zum einen impliziert Risiko immer einen epistemischen Mangel an Wissen, es geht um »knowledge about the unknown«; zum anderen sind stets unerwünschte Ereignisse involviert. Folglich ist der Risikobegriff prinzipiell wertgeladen.

Wichtig für das Verständnis des umfassenden Risikobegriffs ist die Tatsache, dass dieser sich stets auf Entscheidungen und Handlungen bezieht. Während der Unsicherheitsbegriff epistemische Zustände benennt, die spezifische Handlungsbedingungen darstellen, setzt der Risikobegriff diese in einen Handlungszusammenhang: »Unsicherheit als Beschreibung eines epistemischen Zustandes ist somit deskriptive Voraussetzung für die normativ relevante Kennzeichnung einer Entscheidung oder Entscheidungssituation als risikobehaftet.« (Nida-Rümelin et al., 2012, S. 10). Daraus folgt, dass Risiko im umfassenden Sinne immer einen Akteursbezug aufweist, d. h. dass die Manifestation einer Konsequenz immer (zumindest teilweise) ursächlich durch eine Handlung bzw. Entscheidung ausgelöst wird oder die Konsequenz in Ausmaß oder Wahrscheinlichkeit durch

225 Eine subjektive Wahrscheinlichkeit liegt auch dann vor, wenn begründete Zweifel bestehen, ob alle relevanten Variablen hinsichtlich ihres numerischen Werts oder ihrer Relevanz korrekt berücksichtigt wurden (vgl. Rath, 2011, S. 20).

eine Handlung beeinflusst werden kann (vgl. ebd., S. 5–7).[226] Dabei ist zu beachten, dass die Konsequenzen einer Handlung neben dem Risikourheber auch andere Unbeteiligte betreffen können; dies macht eine weitere begriffliche Unterscheidung zwischen individuellen und übertragenen Risiken notwendig. Während im Fall eines individuellen Risikos dieses nur der verursachenden Instanz zufällt, wobei diese eine Einzelperson oder ein gemeinsam handelndes Kollektiv sein kann, sind bei einem übertragenen Risiko unbeteiligte Dritte potenziell davon betroffen (vgl. Rath, 2011, S. 19).

Dem umfassenden Begriffsverständnis steht eine Verwendungsweise gegenüber, die einer Verengung des Risikobegriffs gleichkommt.[227] Diese ist jedoch aus ethischer Sicht unergiebig und hat für real-lebensweltliche Probleme wenig Relevanz. Es erscheint plausibel anzunehmen, dass bei realen Entscheidungen oft nur unvoll-

226 Für ein tieferes Verständnis ist weiterhin der Begriff der ›Gefahr‹ hilfreich. Rath (2011, S. 24) beschreibt diese als labilen Zustand, der mit gewisser Wahrscheinlichkeit eine potenzielle negative Konsequenz freisetzt. Gefahr stellt in ethischer Hinsicht den Gegenbegriff zu Risiko dar. Eine Gefahr an sich ist ethisch irrelevant, nicht aber das Wissen um eine solche (vgl. Nida-Rümelin et al., 2012, S. 22). Risiken sind entscheidungsbezogen, Gefahren hingegen werden durch externe (Umwelt-)Bedingungen verursacht, die auch in sozialen Konstrukten bestehen können (vgl. Luhmann, 1991a, S. 89; 1991b, S. 30–31).

227 Der Begriff des Risikos wird dabei auf zwei Weisen verengt. Zum einen bezeichnet er den Umstand, dass infolge einer Entscheidungssituation grundsätzlich negative Konsequenzen auftreten können. In diesem Zusammenhang wird Risiko häufig als Gegenbegriff zu ›Chance‹ konzipiert; risikobehaftet sind Situationen mit mindestens zwei potenziellen Konsequenzen, von denen mindestens eine negativ und die andere nicht positiv (im Sinne von Nutzen) ist (vgl. Rath, 2011, S. 22). ›Risiko‹ in diesem Sinne würde sich nur auf die negativen Konsequenzen einer Handlung beziehen. Nun kann eine Handlung aber nur dann rational in Betracht gezogen werden, wenn sie neben Risiken auch Chancen bietet. Eine risikobehaftete Entscheidung ist daher immer »das Ergebnis einer Abwägung von Nutzen und Schaden« (Nida-Rümelin et al., 2012, S. 7). Es erscheint nicht sinnvoll, mögliche positive und negative Konsequenzen isoliert zu betrachten. Zum anderen wird der Risikobegriff teilweise auch verwendet, um einen spezifischen Fall von Unsicherheit zu beschreiben, bei dem Wahrscheinlichkeiten und Konsequenzen präzise quantifizierbar sind. Eine solche Verengung des Risikobegriffs ist aus praktischer Sicht nicht plausibel, da er nur auf sehr wenige Fälle in der Lebenswelt angewandt werden könnte und eine entsprechende Risikoethik kaum Relevanz entfalten würde (vgl. ebd., S. 6). In der lebensweltlichen Praxis sind Folgen nicht nur selten gewiss, sondern auch kaum in Form exakter Erwartungswerte bestimmbar.

ständige Informationen vorliegen und sich daher keine exakten Erwartungswerte bestimmen lassen.[228] In dieser Forschungsarbeit wird daher der umfassende Risikobegriff zugrunde gelegt, wie er z. B. von Nida-Rümelin et al. (2012) und Rath (2011) verwendet wird.

6.2.2 Risiken im Handlungskontext: Risikosituationen und Risikokonstellationen

Der Begriff des Risikos ist zunächst abstrakt; er beschreibt »lediglich die unterschiedlichen epistemischen Niveaus eines Akteurs in Bezug auf die Folgen seiner Handlungen« (Nida-Rümelin et al., 2012, S. 25), ist aber nicht ausreichend, um eine Entscheidungs*situation* zu charakterisieren. In Risikosituationen kommt zu den beiden Risikovariablen der Konsequenz und der Eintrittswahrscheinlichkeit, welche den Risikobegriff kennzeichnen, noch eine dritte hinzu – die sogenannte Risikokonstellation. Diese hat im Hinblick auf das zentrale risikoethische Anliegen, die Zulässigkeit von Risikoübertragungen zu klären, eine bedeutende Rolle inne, da sie die risikoethisch relevanten Umstände und Rahmenbedingungen reflektiert, unter denen ein Risiko besteht. Sie lässt sich entlang zweier Dimensionen spezifizieren: zum einen der Beziehung zwischen einem Risiko und denjenigen, die von diesem betroffen sind, zum anderen der Tragweite bzw. des Ausmaßes eines Risikos. Entsprechend der jeweiligen Ausprägungen dieser beiden Faktoren lassen sich vier verschiedene Grundtypen von Risikosituationen differenzieren.

6.2.2.1 Individuelle vs. soziale Risikosituationen

Mit einer individuellen Risikosituation haben wir es dann zu tun, wenn Urheber und Betroffene eines Risikos identisch sind, sodass keine Externalitäten vorliegen (vgl. Rath, 2011, S. 27). Das entsprechende private Risiko muss zudem willentlich eingegangen worden sein (vgl. Shrader-Frechette, 1991, S. 105). Es muss sich jedoch nicht

228 Wie Hansson (2009, S. 427) anschaulich beschreibt, sind die Unsicherheiten der realen Lebenswelt mehr mit einer Abenteuerexpedition in den Dschungel vergleichbar als mit einem Besuch im Casino: In aller Regel sind uns weder drohende Gefahren noch deren Wahrscheinlichkeiten bekannt.

zwingend um eine Einzelperson handeln, auch ein Kollektiv kann gemeinschaftlich agieren. Die Zugehörigkeit zu diesem muss freiwillig erfolgen und die Konsequenzen, die sich aus einer Risikosituation für jeden Einzelnen ergeben, müssen akzeptiert sein. Unklar ist, ob das Zweite immer gegeben ist, wenn nur ein Entscheidungsverfahren akzeptiert wurde. Unter diesen Umständen kann eine Situation nur dann als individuelle Risikosituation bezeichnet werden, wenn diejenigen, die einer konkreten Verteilung von Konsequenzen eines Risikos nicht zustimmen, dennoch als Risikourheber gelten können. Dies ist genau dann der Fall, wenn »das Akzeptieren eines Entscheidungsverfahrens zumindest *prima facie* die Verpflichtung impliziert, im Einzelfall unerwünschte Entscheidungen mitzutragen.« (Nida-Rümelin et al., 2012, S. 29, Hervorh. i. Orig.)[229]

In der realen Lebenswelt scheint es eher selten der Fall zu sein, dass gar keine Auswirkungen für andere zu erwarten sind (vgl. Rath, 2011, S. 28–29). Dennoch besitzen individuelle Risikosituationen insofern ethische Relevanz, als sie aufzeigen, dass auch Pflichten gegenüber der eigenen Person ethisch bedeutsam sein können (vgl. Nida-Rümelin et al., 2012, S. 27–28). Einige (Risiko-)Forscher sind der Ansicht, dass individuelle Risikosituationen sich durch ein Abwägen von Vor- und Nachteilen angemessen entscheiden lassen (vgl. Hacking, 1986, S. 141; Shrader-Frechette, 1991, S. 105–106). In diesem Sinne handelt es sich nicht länger um ethische Fragen der Zulässigkeit, sondern um rationale, individuelle Kosten-Nutzen-Analysen. Diese sind solange ethisch unproblematisch und damit risikoethisch irrelevant, wie sie keinen Konflikt in Bezug auf Pflichten gegenüber der eigenen Person kreieren (vgl. Rath, 2011, S. 28–29). Liegt ein solcher allerdings vor, fordern u. a. Nida-Rümelin et al. (2012, S. 27–28), diese Pflichten in den Abwägungsprozess zu integrieren. Andere wie beispielsweise Rawls (1971) vertreten die Ansicht, dass es unabhängig von bestehenden Pflichten gegen sich selbst moralisch zulässig ist, sich willentlich einem Risiko auszusetzen, solange andere davon nicht tangiert werden:

> […] each man in realizing his own interests is certainly free to balance his own losses against his own gains. We may impose a sacrifice on ourselves now for the sake of a greater advantage later. A person quite

229 Trifft diese Bedingung nicht zu, liegt ein Fall von Risikoübertragung durch Zustimmung vor (vgl. Nida-Rümelin, 2005, S. 881–884).

> properly acts, at least when others are not affected, to achieve his own greatest good, to advance his rational ends as far as possible. (Ebd., S. 23)

Werden Pflichten gegen sich selbst hingegen prinzipiell verneint, gibt es kein Kriterium mehr, um derartige Situationen ethisch zu bewerten. Dies würde implizieren, dass alle Handlungen moralisch erlaubt sind, die individuelle Risiken beinhalten; zudem wäre risikobehaftetes Handeln in diesem Fall ebenso zu bewerten wie Handeln mit sicheren Folgen (vgl. Gethmann, 2000, S. 62).[230]

Ist mindestens eines der betroffenen Individuen nicht zugleich Verursacher des Risikos, spricht man von einer sozialen Risikosituation bzw. einem übertragenen Risiko, wobei die Übertragung willentlich oder unwillentlich erfolgen kann (vgl. Rath, 2011, S. 29–30). Für die risikoethische Bewertung der (moralischen) Zulässigkeit von Risikoübertragungen ist Letzteres von großer Bedeutung (vgl. Nida-Rümelin et al., 2012, S. 30). Deutlich wird dies beispielsweise anhand von Thomsons (1985a, S. 124–126) Unterscheidung zwischen verschiedenen Typen übertragener Risiken. Fügt eine Person einer anderen eine unerwünschte Konsequenz[231] zu und legt ihr zusätzlich noch ein Risiko für eine weitere auf, so nennt Thomson dies eine unreine Risikoübertragung (*impure risk imposition*). Als Beispiel führt sie eine Situation an, in der eine Person einer anderen durch einen Schuss in den Bauch eine Verletzung beibringt, die zusätzlich mit dem Risiko einhergeht, mittelfristig daran zu sterben. Solche Fälle sind insbesondere dann risikoethisch relevant, wenn die unerwünschte Konsequenz, die zugefügt wurde, relativ trivial ist im Vergleich zu derjenigen, die durch die Risikoübertragung droht. Erstere kann unter bestimmten Umständen als moralisch unbedeutend betrachtet werden: »That is because the ground for the complaint lying in the risk imposed is so much more grave than the ground for

230 Da in diesem Buch gesellschaftlich-soziale Problematiken im Sinne übertragener Risiken im Fokus stehen, wird die Diskussion individueller Risiken an dieser Stelle nicht weiter vertieft.

231 Thomson (1985a) gibt einige Beispiele dafür an, was als unerwünscht gilt, nimmt jedoch bewusst keine scharfe Abgrenzung vor. Grob gesagt ist eine Konsequenz dann unerwünscht, wenn sie mindestens ein Recht eines Betroffenen verletzt: »It may help, however, to say that what I have in mind is very roughly characterizable as follows: other things being equal, to cause a person an outcome of the kind I mean is to infringe a right of his.« (Ebd., S. 125).

complaint lying in the unwanted outcome actually caused.« (Ebd., S. 126) Die zentrale ethische Problemstellung verbleibt also in Bezug auf das übertragene Risiko. Ein solches ist Gegenstand einer reinen Risikoübertragung (*pure risk imposition*). Hierbei setzt eine Person eine andere einem Risiko aus, wobei mindestens eine der möglichen Konsequenzen unerwünscht ist. Das Paradebeispiel ist hier das Russische Roulette: Unabhängig davon, ob eine Person tatsächlich im Verlauf des Spiels von einer Kugel tödlich getroffen wird, hat sie aufgrund der gravierenden potenziellen Konsequenz guten Grund, eine moralische Beschwerde vorzubringen. Das trifft auch dann zu, wenn die Betroffenen überhaupt nicht wussten, welchem Risiko sie ausgesetzt waren.

6.2.2.2 Triviale vs. katastrophale Risikosituationen

Die zweite Dimension von Risikokonstellationen tritt als Kontinuum diverser möglicher Ausprägungen des Ausmaßes von Risiken auf, das von zwei Extremfällen eingerahmt wird. Ein Risiko gilt als trivial, wenn entweder Eintrittswahrscheinlichkeiten oder Konsequenzen – oder beides – sehr gering sind. Welches Risikoausmaß im konkreten Fall jedoch noch unter diese Definition fällt, ist umstritten.[232] Nida-Rümelin et al. (2012, S. 47) schlagen vor, von trivialen Risiken zu sprechen, »wenn die beteiligten Individuen nahezu indif-

232 Ein bekanntes Beispiel für eine triviale, aber dennoch risikoethisch kontroverse Risikosituation stammt von Thomson (1985a, S. 128–131): Mit dem Anschalten eines Gasherds zum Zweck des morgendlichen Kaffeekochens ist ein prinzipielles Schadensrisiko für den Nachbarn verbunden, beispielsweise durch eine Explosion oder das Austreten von Gas. Es gibt gute Gründe dafür anzunehmen, dass die Wahrscheinlichkeit, dass es dazu kommt, sehr gering ist. Thomson diskutiert eingehend, ob das Risiko trotz des theoretischen Schadens trivial bleibt, auch wenn die mögliche Konsequenz den Tod eines Betroffenen bedeuten würde. Problematisch ist eine Definition von Fällen mit geringen Wahrscheinlichkeiten als trivial dann, wenn der mögliche Schaden sehr hoch ist: »What counts as a low or high risk of this or that harm is presumably a function, not merely of the probability of the harm, but also of the nature of the harm.« (Ebd., S. 131) In diesen Fällen stellt sich die Frage, wie gering die Wahrscheinlichkeit sein muss, damit die Situation insgesamt noch als trivial gelten kann. Im risikoethischen Diskurs ist die Frage nach der Abgrenzung trivialer Risiken umstritten; Rath (2011, S. 44) spricht von einer »Sisyphos-Aufgabe«.

ferent sind zwischen dem Erwartungswert der relevanten Risikosituation und dem Status quo.« In ähnlicher Weise vermerkt Posner (2004, S. 168): »People may not demand any compensation at all for bearing risks that are only trivial greater than zero […]«. Aus risikoethischer Sicht sind triviale Risiken deshalb von Bedeutung, weil allgemeine Entscheidungskriterien hier nicht greifen, sondern die Entwicklung spezieller Kriterien erforderlich ist (vgl. Rath, 2011, S. 43).

Katastrophale Risiken hingegen zeichnen sich dadurch aus, dass mindestens eine mögliche Konsequenz eine sehr große Zahl an Individuen betrifft; die Auswirkungen sind grundsätzlich von sozialer Relevanz.[233] Die Eintrittswahrscheinlichkeit ist dabei aus definitorischer Sicht kein konstitutiver Faktor, kann jedoch für Regulierungsentscheidungen von Bedeutung sein (vgl. Nida-Rümelin et al., 2012, S. 48–49; Rath, 2011, S. 45). Betroffene Individuen sind dabei über die gemeinsame Risikoursache miteinander verbunden. Für Zeckhauser (1996, S. 113) können bei katastrophalen Risiken auch Betroffene zugleich Urheber in dem Sinne sein, dass ohne sie kein Risiko bestünde. Er spricht in diesem Kontext von Risikokonsumenten. So kann z. B. den Bewohnern einer Küstenregion eine implizite Zustimmung unterstellt werden, sich einem potenziell katastrophalen Hochwasserrisiko auszusetzen (vgl. Rath, 2011, S. 46).

6.2.2.3 Kombination von Risikosituationen

Individuelle und übertragene Risiken einerseits sowie triviale und katastrophale Risiken andererseits treten jeweils als antonyme Begriffspaare zueinander auf. Während Risiken immer entweder individuell oder übertragen sind – Externalitäten liegen entweder vor oder nicht –, stellen triviale und katastrophale Risiken Extrempunkte dar, zwischen denen sich ein Kontinuum an vielen möglichen Ausprägungen erstreckt. Um Risikosituationen in Praxisproblemen ethisch bewerten zu können, ist es sinnvoll, die beiden Kategorien miteinander zu kombinieren, was jedoch nicht in allen Fällen mög-

233 Rath (2011, S. 45) weist darauf hin, dass Risiken, die von einzelnen Individuen oder kleineren Gruppen als persönlich katastrophal angesehen werden, nicht als katastrophal im Sinne der obigen Definition gelten können. Diese werden vielmehr als beträchtliche Risiken bezeichnet.

lich ist (vgl. Nida-Rümelin et al., 2012, S. 53). So ist es beispielsweise plausibel anzunehmen, dass viele alltägliche Handlungen individuelle Risiken beinhalten, die zugleich trivial sind. Katastrophale Risiken beziehen sich hingegen definitionsgemäß auf eine größere Anzahl an Individuen und sind daher nicht ohne Widerspruch als individuelle Risiken denkbar. Übertragene Risiken können hingegen sowohl trivial als auch katastrophal sein.

Darüber hinaus sind auch Sonderfälle denkbar. Einen solchen beschreibt Thomson (1985a, S. 125): Durch eine Handlung werden zunächst triviale oder auch gar keine Risiken übertragen, durch deren wiederholte Ausführung – in Form von Handeln in Serie derselben bzw. paralleles Handeln einer anderen Person – wird jedoch aufgrund der aggregierten Wirkung ein kritischer Wert überschritten, der ein nicht-triviales Risiko begründet. Thomson spricht in diesem Kontext von Schwellenwert-Effekten (*threshold effects*), welche sie am Beispiel eines Fischteichs veranschaulicht: Während das erste Hineinschütten einer giftigen Substanz nur zu einer Trübung des Wassers führt, wird mit dem zweiten Mal schließlich eine für die Fische tödliche Gesamtdosis erreicht. Nida-Rümelin et al. (2012, S. 41–45) gehen näher auf die risikoethischen Implikationen dieses Sonderfalls ein, indem sie sich die Frage stellen, wie die letzte Handlung in der Handlungskette zu bewerten ist, welche die Materialisierung des größeren Schadens schlussendlich auslöst. Wird jede einzelne Handlung für sich genommen als moralisch vernachlässigbar angesehen, muss das auch für die letzte in der Kette gelten. Sind die einzelnen Handlungen, die in ihrer Gesamtheit zum Schaden führen, koordiniert, so sind die Handelnden als ein Kollektiv zu verstehen, das ein Risiko an Dritte überträgt. Dies gilt sowohl dann, wenn mehrere Personen parallel handeln, als auch wenn Handlungen in Serie erfolgen. Sind die Handlungen dagegen unkoordiniert, ist jedes Individuum isoliert als handelnde Person anzusehen. In diesem Fall ist es ethisch implausibel, lediglich demjenigen, dessen Handeln letztlich unbeabsichtigt den aggregierten Grenzwert überschreitet, eine willkürliche Verantwortungszuschreibung aufzuerlegen, die als »unbegründete […] Vermischung von kausaler und moralischer Verantwortung« daherkommt (ebd., S. 44). In der Konsequenz verbleibt damit als ethisch relevantes Kriterium sowohl für koordinierte als auch unkoordinierte Handlungen lediglich, ob die handelnden Per-

sonen um ihren Beitrag zur drohenden Grenzwertüberschreitung zu wissen verpflichtet gewesen wären.

Während die Schätzung von Wahrscheinlichkeiten primär methodologische Probleme aufwirft, stellt die Bewertung der Schadensfolgen eine ethische Aufgabe dar. In Bezug auf die lebensweltliche Praxis ist es plausibel anzunehmen, dass die meisten relevanten Risiken weder trivial noch katastrophal sind und damit irgendwo zwischen den Extrempunkten liegen. Nach Ansicht von Nida-Rümelin et al. (2012, S. 53) greift das zuvor beschriebene idealtypische Kategorisierungsschema dahingehend zu kurz: Reale Risikosituationen sind meist nicht »typenrein[…]«, sondern es müssen die spezifischen Umstände mitbedacht werden. Rath (2011, S. 47–48) weist zudem darauf hin, dass bei der Implementierung von Risikostrategien auch zeitliche Aspekte eine Rolle spielen. So könnten Risiken katastrophalen Ausmaßes, die weit in der Zukunft liegen, zum gegenwärtigen Zeitpunkt als triviale Risiken behandelt werden, was sich auf formaler Ebene über die Verwendung von Diskontraten in der Risikobewertung berücksichtigen ließe. Hieraus ergeben sich jedoch neue ethische Probleme hinsichtlich der Bewertung zukünftigen menschlichen Lebens. Rath verweist hier auf einen Ansatz von Sunstein (2002, S. 226–227), der sich mit potenziellen Auswirkungen risikoethischer Maßnahmen auf zukünftige Generationen auseinandersetzt.

Wie sichtbar wurde, kommt der Kategorisierung von Risikosituationen im Hinblick auf die risikoethische Bewertung von Handlungen eine entscheidende Bedeutung zu, denn je nach vorliegendem Typus greifen unterschiedliche risikoethische Kriterien. Auch wenn sich eine exakte Abgrenzung in der Praxis nicht immer als möglich erweist, so ist immerhin eine Eingrenzung hilfreich.

6.2.3 Grundfragen der Risikoethik: Zulässigkeit, Fairness und Verantwortung im Kontext von Risikoübertragungen

Der Gegenstandsbereich der Risikoethik umfasst u. a. die ethische Bewertung von Ausmaß und Wahrscheinlichkeit von Risiken sowie des (sozialen) Beziehungsgefüges, das zwischen Risikoverursachern und Risikobetroffenen besteht. Die philosophische Diskussion risikoethischer Fragestellungen lässt sich im Wesentlichen in fünf zentrale Problemfelder kategorisieren, die sich teilweise gegenseitig bedingen und beeinflussen (vgl. Hayenhjelm & Wolff, 2012, e47); re-

al-lebensweltliche Fragen tangieren in der Regel mehrere der nachfolgend beschriebenen Themen.

Die erste Frage, die im Rahmen risikoethischer Auseinandersetzungen diskutiert wird, thematisiert die moralische Rechtfertigung bzw. Zulässigkeit von Risikoübertragungen. Während es ethisch unproblematisch erscheint, Personen einem Risiko auszusetzen, sofern die mögliche Konsequenz auch direkt zugefügt werden dürfte (vgl. Scheffler, 1985, S. 83; Thomson, 1985a, S. 127), greift die Risikoethik insbesondere Situationen auf, die den negativen Fall beschreiben: Auch wenn das Zufügen einer direkten, ungewollten Konsequenz moralisch verboten ist, können dennoch Umstände vorliegen, unter denen eine entsprechende Risikoübertragung zulässig ist. Der Umkehrschluss gilt allerdings nicht: Mit der Zulässigkeit von Handlungen, durch die Risiken übertragen werden, kann nicht das direkte Zufügen einer ungewollten Konsequenz gerechtfertigt werden (vgl. Nida-Rümelin et al., 2012, S. 32). Wird das Recht, keinem Risiko durch andere ausgesetzt zu werden, als absolutes moralisches Gebot betrachtet, resultiert daraus eine prinzipielle Unzulässigkeit jeglicher Art von Risikoübertragung. Hieraus ergibt sich das zweite zentrale Problem, dem sich die Risikoethik widmet: Das sogenannte *Problem of Paralysis* impliziert, dass ein absoluter Geltungsanspruch des Nichtschadensgebots den Raum erlaubter Handlungen derart unplausibel einschränkt, sodass letztlich jegliches Handeln an sich untersagt wäre.

Das dritte Problemfeld der Risikoethik thematisiert Fragen einer fairen Verteilung von Vor- und Nachteilen, die durch Risiken entstehen. Nicht immer sind gesamtgesellschaftlich getragene Risiken auch vorteilhaft für alle; häufig betreffen die Nachteile risikopolitischer Entscheidungen lediglich Teile der Bevölkerung, während andere weitgehend von den auferlegten Risikoübertragungen profitieren. Eine zentrale Bedeutung in risikoethischen Fragen kommt daher distributiven Aspekten der Verteilung von Vor- und Nachteilen risikobehafteter Entscheidungen und der damit verbundenen Unterscheidung zwischen Entscheidern und Betroffenen zu. Das vierte, stärker konzeptionell orientierte risikoethische Grundproblem besteht in der Risikobewertung bzw. der numerischen Abschätzung eines Risikos. Immaterielle Werte wie Leben und Gesundheit von Individuen lassen sich ebenso wenig in Zahlenwerten ausdrücken wie konstituierende Faktoren individueller Risikoeinstellungen, z. B.

die Wahrscheinlichkeiten, welche seltenen Ereignissen zugewiesen werden. Das fünfte risikoethische Themenfeld betrifft schließlich die Bewertung moralischer Verantwortung für Risikoübertragungen im Hinblick auf nicht klar differenzierbare Ursachen, unzureichendes Wissen über mögliche Folgen, kollektive Entscheidungen sowie spezifische Verantwortlichkeiten gegenüber zukünftigen Generationen.

Traditionelle ethische Systeme geraten bei dem Versuch, normative Orientierung in Risikosituationen zu geben, regelmäßig an ihre Grenzen. So wirken konsequentialistische Theorien oft zu tolerant in dem Sinne, dass sie Risiken stets für zulässig erklären, sofern der erwartete Nutzen die Kosten übersteigt (vgl. Hayenhjelm & Wolff, 2012, e26). Zudem stützen sie sich auf aggregierte Nutzen- und Kostenwerte, die gegenüber den Folgen für einzelne Personen indifferent sind (siehe Kap. 6.3.1). Rechtebasierte Moraltheorien hingegen schränken den Raum erlaubter Handlungen zu stark ein; es erscheint impraktikabel, Individuen das Recht zuzuschreiben, niemals einem Risiko ausgesetzt zu werden. Risikosituationen, die nur negative Folgen haben, sind anhand von tradierten Moralprinzipien nicht final begründbar (siehe Kap. 4.4). Gleichzeitig ist eine systematische Nichtberücksichtigung von Unsicherheits- und Risikoaspekten in ethischen Theoriegebäuden nicht glaubwürdig, denn Risiken sind bei Handlungen in der realen Lebenswelt allgegenwärtig:

> ›Risk‹ as an integrated part of our understanding of moral actions would bring much of moral theory closer to the inherent complexity of moral issues. There are relevant discussions about uncertainty and risk in decision theory and epistemology but these stay within the realm of rational, rather than moral, actions. In moral and political philosophy there have been interesting debates about moral luck and lotteries, but such topics have typically been addressed in terms of reasons for exemptions from moral responsibility, or in any case as special cases rather than as an essential part of moral actions. (Ebd., e46–e47)

Um risikobehaftete Entscheidungen ethisch bewerten zu können, müssen Moraltheorien für Anwendungskontexte mit unsicheren Folgen dahingehend konzeptionell erweitert werden, dass sie Unsicherheiten angemessen berücksichtigen können (vgl. Hansson, 2003, S. 291–292). Die Risikoethik bietet eine solche Erweiterungsmöglichkeit an. Jedoch geht es in risikoethischen Argumentationen *nicht* darum, alternative ethische Ansätze zu begründen, sondern tradierte Prinzipien lediglich da zu präzisieren, wo sie keine ausreichenden

Antworten für risikobehaftete Entscheidungssituationen geben können. Gemäß Rath (2011, S. 12) kann die Risikoethik »immer nur als Ergänzung zu einer etablierten Moral verstanden werden.« Sie behandelt nicht nur konkrete angewandte Probleme, sondern trägt auch zur Erweiterung der Moraltheorie an sich bei:

> It is in part ethics, decision theory, and epistemology, applied to cases of danger with the aim to inform the normative discussion on distinct issues and problems. But it is also in part a re-shaping of moral philosophy to reconcile moral theory with circumstances of epistemic uncertainty and high stakes. (Hayenhjelm & Wolff, 2012, S. e46)

Im Wesentlichen identifiziert die Risikoethik drei Begründungsansätze hinsichtlich der moralischen Zulässigkeit von Risikoübertragungen: Der erste Ansatz beruht auf einer quantitativen Optimierung bzw. auf Kosten-Nutzen-Abwägungen, der zweite rückt kontraktualistische Elemente in den Fokus und der dritte argumentiert schließlich deontologisch auf der Grundlage individueller Rechte und allgemein anerkannter Gerechtigkeitsprinzipien. Ersterer bildet die Basis einer rationalen Risikopraxis, welche die politische Entscheidungsfindung in praktischen Risikofragen seit geraumer Zeit dominiert. Im folgenden Unterkapitel wird sie in ihren Grundzügen dargestellt.

6.3 Grundzüge der (rationalen) Risikopraxis: Paradigmen und entscheidungstheoretische Ansätze

6.3.1 Risikopraktische Paradigmen

Praktische Fragen des Umgangs mit Risiken in Gesellschaft und Politik stehen im Spannungsfeld vielschichtiger Anforderungen an tragfähige Konzeptionen, die sowohl kontextadäquate als auch normethische Aspekte miteinbeziehen. Die Art und Weise, wie Risiken vor dem Hintergrund gesellschaftlicher und politischer Fragen seit den 1980er-Jahren problematisiert werden, ist gekennzeichnet durch spezifische Ansätze politischer und ethischer Risikobewältigung, die in weiten Teilen konträr zueinander stehen. Nida-Rümelin et al. (2012, S. 55) beschreiben dies als Paradigmendualismus: Auf der einen Seite befindet sich das sogenannte konsequentialistisch-objektivistische Paradigma, das sich in seinen Grundzügen an der utilita-

ristischen Tradition ethischen Denkens einerseits sowie der ökonomischen Entscheidungstheorie andererseits orientiert. Es fußt auf der Annahme, dass Risiken als mathematische Größe aus Schaden und Wahrscheinlichkeit objektiv bestimmbar sind, und damit auch optimierbar. Zentrales Element ist die Aggregation von Einzelrisiken zu einer Gesamtrisikosumme, die der Zielvorgabe einer ökonomisch geprägten Risikooptimierung im Sinne einer »Maximierung des über verschiedene Personen hinweg aggregierten Erwartungswertes unter Bedingungen der Unsicherheit« (Nida-Rümelin et al., 2012, S. 57) unterliegt. Zu diesem Zweck wird im Rahmen des konsequentialistischen Paradigmas auf Nutzenfunktionen zurückgegriffen, die normative Implikationen für die Wahl von bestimmten Handlungsalternativen aufweisen: nämlich derjenigen, die den höchsten Nutzenwert einer spezifischen Nutzenfunktion realisieren.

Aus ethischer Sicht ist eine Aggregation in Bezug auf Einzelinteressen grundsätzlich fragwürdig (vgl. Nida-Rümelin, 2005, S. 877; Nida-Rümelin et al., 2012, S. 57–59).[234] Individuelle Rechte, Freiheit und Autonomie sowie etablierte Gerechtigkeitsvorstellungen stellen ethisch begründete und verfassungsrechtlich gesicherte Grenzen einer ökonomisierten Risikoaggregation dar. Personen dürfen nicht ungefragt Risiken ausgesetzt werden, die ihnen im Zuge paternalistischer Strategien zur Risikooptimierung auferlegt werden. Dabei wird sich nicht nur über individuelle, subjektive Risikowahrnehmungen hinweggesetzt. Vielmehr bedeutet die mangelnde Unterscheidung von Risikoakteuren zwischen Entscheidern und Betroffenen auch eine Verletzung des autonomen, individuellen Zustimmungsvorbehalts, was die generelle Indifferenz utilitaristischer Konzeptionen gegenüber den Interessen der Einzelnen bei einer ausschließlichen Konzentration auf aggregierte Folgen und Wahrscheinlichkeiten widerspiegelt (vgl. Nida-Rümelin, 2005, S. 874–876). Die u. a. von Rawls begründete Separatheit von Personen fordert, dass interpersonelle Nutzenabwägungen unzulässig bleiben. Da der Konsequentialismus stets auf Optimierung ausgerichtet ist, können nur solche Projekte Bestand haben, die zur Maximierung eines Gesamtnutzens beitragen. Jeder Einzelne muss dem Gemeinwohl alles unterordnen,

234 Kritikpunkte, die sich auf die konsequentialistische Ausrichtung des Paradigmas beziehen, stehen im Kontext einer generellen Kritik eines ethischen Konsequentialismus, wie sie beispielsweise Nida-Rümelin (1993) vertritt und die auch bereits in Kap. 4.4.2 in Grundzügen dargelegt wurde.

wodurch sich persönliche Projekte auflösen und die Integrität der Person verletzt zu werden droht. Angesichts sich tatsächlich manifestierender Schädigungen infolge von Risikoübertragungen und damit einhergehender Verletzungen von Individualrechten wird deutlich, dass eine Maximierung des aggregierten Nutzens in jedem Fall begründungsbedürftig ist, da sie sich über das traditionell kantianisch begründete Instrumentalisierungsverbot hinwegsetzt. Gerechtigkeitsethische Problematiken ergeben sich aus der Fokussierung auf eine interpersonelle Nutzensumme insofern, als das Auseinanderfallen von Vor- und Nachteilen für Individuen aufgrund von Risikoübertragungen grundsätzlich unberücksichtigt bleibt. Es gibt also Einschränkungen, die man auch dann nicht übertreten darf, wenn sie zur Optimierung eines gegebenen (konsequentialistischen) Wertmaßstabs führen würden:

> Die Übersetzung […] deontologische[r][…] Einschränkungen in die Werteterminologie führt dann etwa zur Theorie der Unvergleichbarkeit menschlichen Lebens, was in der Konsequenz, wenn wir uns auch die anderen oben genannten Aspekte vor Augen führen, zu einer mehrdimensionalen Bewertungsmatrix führt, die Handeln im Sinne des Optimierungsparadigmas nur in engen Grenzen zuläßt. (Ebd., S. 878)

Als Gegenpol und kritische Antwort auf die Defizite des konsequentialistischen Paradigmas begreift sich das sogenannte postmodern-subjektivistische Paradigma (vgl. Nida-Rümelin et al., 2012, S. 60–63). Dessen Vertreter[235] gehen im Anschluss an den insbesondere soziologisch getriebenen Risikokonstruktivismus davon aus, dass Risiken stets subjektiv sind und Einzelinteressen daher eine große Bedeutung erlangen. Das zentrale Problem einer solchen Sichtweise besteht in der Prämisse, dass Unsicherheiten nicht mehr als gegebener Teil der Realität angesehen werden, sondern lediglich als gesellschaftliche, kulturrelative und intersubjektiv konstituierte Konstrukte, für die sich keine allgemeingültigen Kriterien definieren lassen. Damit wird die normative Dimension jeglicher Relevanz für risikopolitische Entscheidungen beraubt. Es bleibt unklar, wie ein wohlbegründeter Umgang mit Risiken auf gesellschaftlicher Ebene unter dem postmodern-subjektivistischen Paradigma erfolgen kann, was sich angesichts stetig steigender Entscheidungsnotwendigkeiten als

235 Als solche können beispielsweise Ulrich Beck als Verfasser der *Risikogesellschaft* (1986), Wolfgang Bonß (1995) oder Niklas Luhmann (1991b) gelten.

problematisch erweist. Die Schwächen des konsequentialistisch-objektivistischen wie auch des postmodern-subjektivistischen Paradigmas münden letztlich in der Begründungsunfähigkeit hinsichtlich einer umsetzbaren Risikopraxis.

In der Folge kam es zur Herausbildung eines dritten, alternativen Paradigmas, das risikopolitische Fragen der neueren Zeit prägt. Das sogenannte partizipatorische Paradigma (vgl. ebd., S. 63–70) ist durch eine primär politische Grundausrichtung charakterisiert. Im Gegensatz zum postmodernen Pendant, das normative Elemente vollständig ausblendet, setzt es an der defizitären normativen Perspektive des konsequentialistischen Paradigmas an. Ausgehend von einer Kritik an dessen zugrundeliegender Risikoobjektivierung distanziert es sich insbesondere von der damit assoziierten quantitativen Ausrichtung der Risikoanalyse, welche in indirekter Weise auf normative Aspekte risikopolitischer Entscheidungen einwirkt. Im Zentrum steht der Entwurf explizit normativer Konzeptionen zur Gestaltung öffentlicher Risikodiskurse. Vertreter des partizipatorischen Paradigmas gehen dabei von dem Grundgedanken aus, dass sich risikobehaftete Entscheidungen durch die Partizipation Betroffener am Entscheidungsprozess legitimieren lassen. Herausforderungen stellen sich für sie vor allem im Hinblick auf die Klärung der normativen Grundpfeiler, mit denen die Partizipation begründet wird; konkret hinsichtlich ihrer diskursethischen und demokratietheoretischen Grundlagen sowie der Elemente, die aus der Logik kollektiver Entscheidungen resultieren, z. B. Effekte aus strategischer Interaktion etc.

Inzwischen sind diverse normative Basisintuitionen in Bezug auf eine ethische Herangehensweise an die Problematik allgemein anerkannt (vgl. Ott, 1998, S. 139), die eine Zurückweisung eines konsequentialistisch-objektivistischen Paradigmas in seiner radikalen Form implizieren. Dennoch wird die gesellschaftspolitische Risikopraxis bis heute von entscheidungstheoretischen Kriterien dominiert, die konsequentialistisch geprägte Optimierungsziele – meist im Rahmen der TFA – verfolgen. Diese rationale Risikopraxis ist aufgrund ihrer Problemorientierung eng an politische und gesellschaftliche Prozesse gebunden und damit nicht gänzlich ohne Werturteile denkbar (vgl. Grunwald, 2002, S. 215–216). Die prominentesten und aufgrund ihrer ausgereiften theoretischen Fundierung be-

deutsamsten rationalen Entscheidungskriterien werden im Folgenden aus einer kritischen Perspektive umrissen.

6.3.2 Entscheidungstheoretische Kriterien rationaler Risikopraxis

6.3.2.1 Das Bayesianische Prinzip

Die auf Thomas Bayes zurückgehende *Bayesianische Entscheidungstheorie* basiert auf einem Verständnis des Rationalitätsbegriffs als spezifische Handlungsstruktur, welche die Grundform des späteren *Rational-Choice*-Paradigmas bildet (vgl. Nida-Rümelin et al., 2012, S. 73–92; Rath, 2011, S. 51–69). Ihr Grundmodell verbindet Elemente der utilitaristischen und ökonomischen Theorie zum Konzept einer Entscheidungsrationalität, in dessen Zentrum Maximierungsziele hinsichtlich des Erwartungsnutzens stehen, die soziale Präferenzen einer Gesamtgesellschaft widerspiegeln.

Neben der Anwendung im Rahmen von Kosten-Nutzen-Analysen entstammt die für die Risikoethik bedeutsamste Version der *Bayesianischen Entscheidungstheorie* der Konzeption des Wirtschaftswissenschaftlers John Harsanyi. Nach dessen Verständnis ist die Ethik eine Teildisziplin der Theorie rationalen Handelns – genauer eine Theorie des rationalen moralischen Urteilens. Ihre Aufgabe ist es, die Erreichung gesellschaftlicher Interessen über die Formulierung und Begründung von Axiomen sicherzustellen. Ein solches neo-utilitaristisches Ethikverständnis schließt Normen aus, die unabhängig von konkreten Konsequenzen gut oder schlecht sein können (vgl. Harsanyi, 1977a, S. 322–324). Nutzenwerte treten innerhalb von Harsanyis Theorie als Von-Neumann-Morgenstern-Nutzenfunktionen auf (vgl. Harsanyi, 1977b, S. 642–644). Vor deren Hintergrund entwickelt Harsanyi ein System, über das kardinale Vergleiche individueller Nutzenniveaus möglich werden. Mit der Abgrenzung vom klassischen Handlungsutilitarismus und seiner gleichzeitigen Hinwendung zum Regelutilitarismus versucht Harsanyi grundlegender Kritik zu entgehen, was allerdings nur teilweise gelingt; fundamentale Annahmen bleiben fragwürdig. So setzt die Argumentation zugunsten einer utilitaristischen Grundlage für risikoethische Entscheidungssituationen die interpersonelle Vergleichbarkeit von Nutzenwerten voraus. Auch die Annahme, dass rationale Individuen

allen möglichen sozialen Rollen *a priori* dieselbe Wahrscheinlichkeit zuordnen,[236] ist sowohl empirisch als auch normativ nicht glaubhaft – ebenso wie die Nichtbeachtung individueller Risikoeinstellungen.

6.3.2.2 Das Maximin-Kriterium

Da die *Bayesianische Entscheidungstheorie* rationale Präferenzen in den Mittelpunkt stellt, ist sie nur auf solche Situationen anwendbar, in denen Informationen vorliegen, die rationalitätstheoretische Mindestanforderungen erfüllen. Viele lebensweltliche Situationen sind hingegen durch Ungewissheit gekennzeichnet, sodass (subjektive) Wahrscheinlichkeiten nicht sicher angenommen werden können. Der geläufigste entscheidungstheoretische Ansatz für Situationen unter Ungewissheit ist das *Maximin*-Kriterium als klassische Strategie zur Risikovermeidung (vgl. Nida-Rümelin et al., 2012, S. 95–99; Rath, 2011, S. 69–88). Es impliziert in sozialen Risikosituationen die Wahl derjenigen Alternative, deren schlechtestmögliche erwartete Konsequenz besser ist als die aller anderen Optionen. In Form einer auf ordinalen Präferenzen basierenden, risikoaversen Strategie der Vermeidung des größten Übels bzw. der Wahl des höchsten Sicherheitsniveaus findet es prominente Anwendung beispielsweise im Rahmen der Verantwortungsethik von Hans Jonas (1979) sowie der politisch-ethischen Theorie von John Rawls (1971). Letzterer leitet die *Maximin*-Regel kontraktualistisch als Ergebnis der im Urzustand und unter dem *Schleier des Nichtwissens* getroffenen Wahl der Individuen her. Das *Maximin*-Prinzip fordert dabei, dass diejenige Alternative gewählt wird, die für die am wenigsten begünstigte Person den höchsten Nutzen verwirklicht.[237] Dies setzt ordinale Vergleiche in zweierlei Hinsicht voraus: Wie Rawls erläutert, lässt sich das am schlechtesten gestellte Individuum aus dem Urzustand heraus ermitteln; das relevante Maß zur Bestimmung des qualitativen Niveaus, anhand dessen sich die Vorteilhaftigkeit von Ergebnissituationen be-

236 Diese Annahme beruht auf einer bayesianisch geprägten Interpretation des Wahrscheinlichkeitsbegriffs, welche diesen nicht als relative Häufigkeit versteht, sondern als Ausdruck rationaler Präferenzen. Diese Auffassung liegt auch der Laplace'schen Wahrscheinlichkeitsverteilung zugrunde.

237 Diese Handlungsregel bildet die Grundlage von Rawls' *Differenzprinzip*, das in Kap. 7.3.3 näher ausgeführt wird.

werten lässt, besteht in der Aussicht auf spezifische gesellschaftliche Grundgüter (vgl. ebd., S. 92).

Auch wenn das *Maximin*-Kriterium in spezifischen Situationen sinnvoll sein mag, ist es als allgemeine Regel unplausibel, was einerseits auf die Verkörperung einer höchst risikoaversen Einstellung und andererseits die ausschließliche Berücksichtigung negativer Folgen zurückzuführen ist. So kommt es teilweise zu kontraintuitiven Ergebnissen, wenn diejenige Option gewählt wird, deren schlechtestmögliche Konsequenz sich nur geringfügig von derjenigen möglicher Alternativen unterscheidet, diese hinsichtlich der erwarteten Vorteile bzw. positiven Effekte jedoch weitaus schlechter ist. In diesem Sinne setzt das *Maximin*-Prinzip zwar nicht direkt kardinale Nutzenwerte voraus, jedoch sind diese bei sehr großen Unterschieden zwischen Alternativen implizit entscheidungsrelevant.[238] Neben kontraktualistischen Begründungsversuchen gibt es auch alternative Ansätze, die versuchen das *Maximin*-Prinzip praktisch zu begründen. So konstatiert beispielsweise Shrader-Frechette im Rahmen ihres *Scientific Proceduralism* (1991, S. 46–50), dass das *Maximin*-Prinzip – im Gegensatz zum Bayesianischen Kriterium – eine hohe Legitimität innerhalb einer gesellschaftlichen Ordnung genießt, in der Risikoevaluation als ein politischer Prozess verstanden wird.

6.3.2.3 Das Prinzip der Vorsicht

Das *Prinzip der Vorsicht* (*precautionary principle*) (vgl. Nida-Rümelin et al., 2012, S. 105–122; Rath, 2011, S. 88–103) ist eine (internationale) Weiterentwicklung des Vorsorgeprinzips und dient traditionell als institutionalisierte Grundlage verschiedener, vor allem umweltbezogener Gesetzgebungen und Bestimmungen. Tatsächlich handelt es sich weniger um ein Entscheidungsprinzip im engeren Sinne als vielmehr um ein Entscheidungsverfahren bzw. einen organisierten Entscheidungsprozess (vgl. Graham & Hsia, 2002, S. 373), der auf verschiedene Entscheidungslogiken zurückgreift, welche an das *Maximin*- bzw. Bayes-Kriterium angelehnt sind. Daher stellt es streng

238 Das Gegenstück ist die *Maximax*-Regel, welche die Wahl der Alternative mit der besten der bestmöglichen erwarteten Folgen fordert und somit eine risikofreudige Einstellung widerspiegelt. Hier treffen analoge Kritikpunkte zu.

genommen kein alternatives Entscheidungsprinzip dar, sondern ein integratives Verfahren, in dessen Rahmen risikoaverse Vermeidungsstrategien und utilitaristische Elemente eine variable Entscheidungstheorie konstituieren. Diese versteht sich als Methodik verantwortungsethischer Risikooptimierung, in deren Zuge anfänglich von einer temporären Risikoaversion ausgegangen wird.

In der Literatur findet sich keine fixe Definition des *Prinzips der Vorsicht*; vielmehr existieren verschiedene Varianten auf der Basis eines gemeinsamen Anwendungsgrunds; ein solcher ist durch das Vorliegen einer Risikosituation mit inakzeptablen Konsequenzen einerseits und unzureichenden wissenschaftlichen Informationen über diese andererseits gegeben. Das *Prinzip der Vorsicht* steht naturgemäß immer in Zusammenhang mit der politischen Implementierung entsprechender Risikostrategien. Es beinhaltet einen dreistufigen Prozess, für dessen Beschreibung sich Rath (2011, S. 93–95) an einer von der Commission of the EC (2000, S. 13–20) herausgegebenen Richtlinie hinsichtlich des *Prinzips der Vorsicht* orientiert. Die drei Stufen sind: Risikobewertung (wissenschaftliche Erfassung der Risikosituation), Risikomanagement (Festlegung der Maßnahmen mit dem Ziel der Erhaltung des Status Quo oder der Erreichung eines bestimmten Zielzustands aus zunächst neutraler Position, d. h. es ist auch denkbar, keine Maßnahme zu implementieren) und Risikoinformation (politische Kommunikation der Implementierung). Liegen (wissenschaftliche) Unsicherheiten in der ersten Phase der Risikoerfassung vor, so kommt das *Prinzip der Vorsicht* zum Einsatz; dies ist insbesondere bei katastrophalen Risiken der Fall.

Die Kernidee des Prinzips besteht in einem Strategiewechsel, welcher im zeitlichen Verlauf und aufgrund von Veränderungen in der Informationsbasis vollzogen wird: Besteht ein Zustand weitreichender Unsicherheit,[239] so wird zunächst eine risikoaverse Strategie der Risikovermeidung gewählt. Ergeben sich zu einem späteren Zeitpunkt, beispielsweise durch Forschung oder neue gesellschaftliche

239 Der Begriff der wissenschaftlichen Unsicherheit spielt eine bedeutende Rolle für das *Prinzip der Vorsicht*, wobei sich unterschiedliche Interpretationen finden lassen. Vor dem Hintergrund einer allgemeinen und möglichst umfassenden Definition wird »tendenziell jede Situation, in der subjektive Annahmen hinsichtlich der Konsequenz oder der Eintrittswahrscheinlichkeit zu treffen sind, ein Kandidat für die Anwendung des Precautionary Principle.« (Rath, 2011, S. 93)

Entwicklungen, weitere Informationen über das betreffende Risiko, so fehlt fortan die Rechtfertigungsgrundlage für die Fortführung eines strikt risikoaversen Vorgehens. Es erfolgt eine Strategieanpassung auf der Basis einer risikoneutralen bzw. risikofreudigen Einstellung. Das *Prinzip der Vorsicht* erfordert notwendigerweise eine stetige Neubewertung der gegebenen Risikosituationen, um kontinuierlich Anpassungen an veränderte Rahmenbedingungen vornehmen zu können.

Kritik am *Prinzip der Vorsicht* als risikoethische Entscheidungstheorie richtet sich primär gegen die spezifischen Herausforderungen, mit denen es sich konfrontiert sieht. Rath (2011, S. 99–102) merkt an, dass aufgrund der Flexibilität des Prinzips besondere Anforderungen für seine plausible Anwendung formuliert werden müssen. Die Commission of the EC (2000, S. 18–21) definiert Verhältnismäßigkeit, Diskriminierungsverbot, Konsistenz, Bewertung von Vor- und Nachteilen sowie die Beachtung wissenschaftlicher Entwicklungen als fünf Eckpunkte entsprechender Leitlinien, die zudem die Thematik der Zuordnung der Beweislast klären müssen. Nida-Rümelin et al. (2012, S. 119–122) skizzieren zwei weitere Argumente. Erstens ist das Prinzip zur Abbildung systemischen Denkens ungeeignet; durch die Fokussierung auf eine isolierte Risikosituation kommt es zu einer Vernachlässigung von Systemeffekten, die beispielsweise entstehen, wenn im Rahmen der Umsetzung von Risikovermeidungsstrategien neue Risiken generiert werden oder Ressourcen fehlen, um andere (zukünftige oder parallele) Risiken effektiv zu bewältigen.[240] Die Folge kann ein höheres Gesamtrisiko sein, das eine Anpassung der Zielvorgabe hin zu einer Minimierung des »aggregierten Systemrisikos« (ebd., S. 120) erforderlich macht. Zweitens bleibt unklar, welche Mechanismen bzw. Instrumente geeignet sind, um gewünschte Verteilungseffekte zu erreichen, wenn Unsicherheit hinsichtlich erwarteter Wirkungen besteht.

6.3.2.4 Weitere Entscheidungsprinzipien

Neben den drei vorgestellten Kriterien gibt es noch weitere Entscheidungsprinzipien (vgl. Nida-Rümelin et al., 2012, S. 99–103; Rath,

240 Diese Konzentration von Ressourcen zur Risikoreduktion bezeichnet Sunstein (2005, S. 32) als »substitute risk«.

2011, S. 105–107), die aufgrund ihrer fehlenden systematischen Ausarbeitung im Hinblick auf risikoethische Fragen allerdings kaum eine Rolle spielen. Sie seien an dieser Stelle lediglich der Vollständigkeit halber erwähnt: 1. Das Hurwicz-Kriterium basiert auf einer Kombination von *Maximin* und *Maximax*, indem es sowohl das bestmögliche als auch das schlechtestmögliche Ergebnis einer Handlung berücksichtigt, welche jeweils mit einem Pessimismus-Optimismus-Index gewichtet werden. Der Fokus auf Extreme ist hier ebenso strittig wie die Nichtbeachtung von Wahrscheinlichkeiten. 2. Das Laplace-Kriterium nimmt gleichverteilte Wahrscheinlichkeiten an, was willkürlich und unplausibel erscheint. 3. Die *Bootstrapping*-Methode basiert auf einer Entscheidungslogik, nach der in ethisch fragwürdiger Weise von vergangenem Risikoverhalten auf die Akzeptabilität zukünftiger Risiken geschlossen wird. 4. Eine kritische ethische Auseinandersetzung mit dem Zufallsprinzip wurde bereits in Kap. 5.4.2.1 skizziert.

6.3.3 Zur Kritik traditioneller Risikopraxis

Die in diesem Kapitel vorgestellten, durchweg konsequentialistischen Entscheidungskriterien rationaler Risikopraxis sehen sich einer zunehmend kritischen Haltung gegenüber, die ihre zentralen Argumente aus der Warte einer generellen Kritik am ethischen Konsequentialismus speist. Zusammenfassend lassen sich gemäß der Darstellung von Rath (2011, S. 109–119) fünf zentrale Problemfelder identifizieren, die in allen drei dominanten Entscheidungsprinzipien präsent sind.[241] Erstens setzen die skizzierten Kriterien implizit oder explizit die Vergleichbarkeit risikorelevanter Zustände hinsichtlich individueller Nutzenwerte (Bayes), gesellschaftlicher Positionen (*Maximin*) oder Risikodimensionen in Systemeffekten (*Prinzip der Vorsicht*) voraus. Zweitens werden Verteilungseffekte entweder, wie im Rahmen des Bayes-Kriteriums als Maximierung

241 Rath nennt noch weitere kriterienspezifische Themen, hinsichtlich derer sich die Prinzipien kritisieren lassen. Auf eine ausführliche Darstellung derselben wird an dieser Stelle jedoch verzichtet, da es hier weniger um eine Widerlegung einzelner spezifischer Prinzipien als vielmehr um eine generelle Kritik an der konsequentialistischen Basis geht, welche die genannten Entscheidungskriterien gemeinsam haben.

des Durchschnittsnutzens, unglaubwürdig begründet oder einem höheren Ziel, beispielsweise der Vermeidung besonders schlechter Konsequenzen bei *Maximin* und Vorsichtsprinzip, untergeordnet. Drittens stellt die Verletzung individueller Rechte eine grundsätzlich kontroverse Problematik angewandter ethischer Fragen dar, die einer komplexen Begründung bedarf. Ein überzeugendes Kriterium der Risikobeurteilung kann niemals lediglich aggregativ sein, sondern würdigt stets individuelle Interessen der Betroffenen. Viertens werfen die Kriterien Fragen zur Legitimität von individuellen bzw. kollektiven Entscheidungsträgern und spezifischen Entscheidungskategorien auf. Fünftens verzichten sie allesamt auf die Legitimationsgrundlage expliziter Zustimmung, die nicht nur eines der mächtigsten risikoethischen Instrumente ist, sondern als lebensnahes Element die ethische Komplexität angewandter Fragen zu reduzieren vermag.

Aus der ethischen Kritik an einer konsequentialistischen Grundorientierung rationaler Entscheidungskriterien folgt somit, dass diese als finale Grundlage einer glaubwürdigen Risikopraxis normethisch nicht zu rechtfertigen sind. In diesem Zuge muss auch die konsequentialistisch geprägte Denkweise in Bezug auf die ethische Dimension von Entscheidungsstrategien für Risikofragestellungen, wie sie insbesondere im Kontext der TFA vorherrschend ist, in Frage gestellt werden. Eine der prominentesten kritischen Positionen vertritt hier Roeser (2018, 2020). Sie argumentiert sowohl aus Sicht einer demokratischen Legitimität als auch aus der Warte allgemein anerkannter Grundwerte wie Fairness, Autonomie und Gleichheit u. a. gegen die technokratische Fokussierung auf rein formale, quantitative Ansätze der Risikobewertung. Dabei betont sie insbesondere die Rolle von Emotionen, welche sie als Ausdruck verinnerlichter Werte interpretiert.[242] Auch Hanssons (2007b, S. 27) Forderung nach der Ergänzung quantitativer Methoden der Risikoanalyse mit spezifisch ethischen Aspekten wie Freiwilligkeit, Zustimmung, Absicht und gerechtigkeitsethischen Überlegungen unterstreicht die Notwendigkeit eines Paradigmenwechsels von einem konsequentialistischen hin zu einem nicht-konsequentialistischen Ansatz für den

242 Hilfreich für das Verständnis von Roesers Thesen ist ein Artikel von Nyholm (2020b), in dem er sich vor dem Hintergrund technologischer Risiken erläuternd mit Roesers Prämissen auseinandersetzt.

ethischen und gesellschaftspolitischen Umgang mit Risiken. Dieser ist unabhängig von der Begründbarkeit und Plausibilität postmodern-subjektivistischer und partizipatorischer Konzeptionen an sich geboten (vgl. Nida-Rümelin, 2005, S. 864–865; Nida-Rümelin et al., 2012, S. 125–133). Dies gilt umso mehr für risikobehaftete Technologien, bei denen beispielsweise deontologische Aspekte bisher weitgehend unterrepräsentiert sind. Die im folgenden siebten Kapitel entwickelte deontologische Perspektive versteht sich daher als Antwort auf die Feststellung, dass die rationalen Grundstrukturen einer traditionellen Risikopraxis der praktischen Problemstellung moralischer Unfalldilemmata nicht gerecht werden können.

7. Unfallalgorithmen als risikoethisches Verteilungsproblem

Bevor in diesem siebten Kapitel schließlich zur risikoethischen Analyse übergegangen wird, ist an dieser Stelle zunächst eine klärende Vorbemerkung zu einer zentralen Begrifflichkeit notwendig. In den bisher einschlägigen risikoethischen Forschungsbeiträgen ist im Hinblick auf das zugrundeliegende Entscheidungsproblem beinahe ausschließlich von einer »Risikoverteilung« (*risk distribution*) die Rede (vgl. z. B. Berkey, 2022; Dietrich, 2021; Dietrich & Weisswange, 2019; Geisslinger et al., 2021; Smith, 2022). »The ethics of automated vehicles may be better framed as the fair distribution of risks, both during and before forced-choice situations«, stellt auch Goodall (2017, S. 496) fest. Aus strikt definitorischer Sicht ist die Verwendung dieses Begriffs jedoch nicht korrekt. Der Terminus der ›Distribution‹ bzw. ›Verteilung‹ ist ökonomischen Ursprungs und wird insbesondere im Kontext von Konzepten der distributiven Gerechtigkeit, z. B. bei Rawls, mit einer Verteilung von Gütern bzw. Ressourcen assoziiert. Mit ihm eng verwandt und häufig verwechselt ist der inhaltlich weiter gefasste Begriff der ›Allokation‹, der sich ebenfalls mit distributiven Fragen auseinandersetzt. In den Wirtschaftswissenschaften meint ›Allokation‹ die Verteilung bzw. Zuordnung knapper Ressourcen bzw. Produktionsfaktoren mit dem Ziel, diese denjenigen Produktionsmöglichkeiten einer Volkswirtschaft im Sinne von Gütern und Dienstleistungen zuzuführen, welche den bestmöglichen Einsatz der Produktionsfaktoren verwirklichen (vgl. Bundeszentrale für politische Bildung, 2016). ›Allokation‹ bezieht sich also auf *Input*faktoren, mittels derer *Outputs* erzeugt werden; die Verteilung Letzterer auf Wirtschaftssubjekte wird als ›Distribution‹ bezeichnet.

Diese begriffliche Differenzierung hat bedeutende Implikationen für ein korrektes Verständnis der Entscheidungsproblematik, die der Programmierung von Unfallalgorithmen zugrunde liegt. Um was geht es eigentlich genau? Der Versuch, entstehende Schäden anhand

ethischer Kriterien zu verteilen, ist schon deshalb unplausibel, weil die Entscheidungssituation sich gerade dadurch auszeichnet, dass die sich tatsächlich manifestierenden Folgen kontingent sind. Es geht vielmehr um die Inputfaktoren, die Schäden hervorbringen – um etwas, aus dem potenziell eine negative Konsequenz entstehen kann, aber nicht zwangsläufig muss: um Risiken. Streng genommen wäre es also angemessener, von einem Allokationsproblem von Risiken zu sprechen:

> The distribution of the risks of operating autonomous vehicles […] may fall short of the targeting paradigm specific individuals, yet the centralised and coordinated nature of algorithmic risk transforms this process to one of risk allocation rather than mere risk distribution. This is because allocation infers direction, edging towards intentionality, and emphasises the perspective of the patient-victim in the accident relationship by connoting the imposition of risk. (Liu, 2017, S. 201)

Der Begriff der ›Verteilung‹ ist hingegen dann adäquat, wenn es um die Outputs geht, die durch Risiken generiert werden. Diese lassen sich allgemein als die Vor- und Nachteile beschreiben, die aus einer Risikokonstellation resultieren und anhand von Fairnesskriterien verteilt werden sollten (siehe Kap. 7.3.3): »Distributive justice, […] is the fairness between those who directly and indirectly benefit and those who directly and indirectly carry the burdens«. (Goër Herve et al., 2023, S. 4) Tatsächlich werden Fragen einer fairen Risikoverteilung im Kontext von Unfallalgorithmen zumeist anhand einer Diskussion von Nutzen und Lasten im Zuge von Risikoübertragungen erörtert. Das akzentuierte Problem ist damit im Kern durchaus ein Verteilungsproblem, wenn auch nur indirekt eines von Risiken. Aus diesem Grund wird in der in diesem Buch vorgelegten risikoethischen Abhandlung von einem *risikoethischen Verteilungsproblem* bzw. einer *Verteilung von Vor- und Nachteilen*, nicht aber von einer Risikoverteilung gesprochen. Wann immer in direkter oder indirekter Weise auf Literatur verwiesen wird, die den Begriff der ›Risikoverteilung‹ verwendet, sei angemerkt, dass zitierte Inhalte sinngemäß vor dem Hintergrund der erläuterten Begriffsreflexion zu verstehen sind.

7.1 Die (risiko-)ethische Problematisierung von Mobilitätsrisiken im Kontext autonomer Fahrsysteme

7.1.1 Autonome Fahrzeuge im Spannungsfeld zwischen soziologischer Risikoakzeptanz und ethischer Risikoakzeptabilität

Obwohl Mobilitätsrisiken gegenwärtig vor allem im Zusammenhang mit neuen Technologien diskutiert werden, sind sie kein Spezifikum der Verkehrsautomatisierung, sondern vielmehr seit jeher Bestandteil des Systems Mobilität. Jedoch nimmt ihre Komplexität seit den Anfangstagen des motorisierten Verkehrs und damit verbundenen höheren Fahrgeschwindigkeiten stetig zu. Am Verkehrsgeschehen beteiligte Personen sind zum einen jenen Risiken ausgesetzt, die durch andere Verkehrsteilnehmer herbeigeführt werden, und zum anderen verursachen sie selbst Risiken, die wiederum andere betreffen. Dabei sind Risiken und Gefahren im Kontext von Mobilitätssystemen, wie auch in vielen anderen Bereichen unseres alltäglichen Lebens, weitgehend gesellschaftlich akzeptiert. Sie werden aufgrund ihres bedeutenden gesellschaftlichen Nutzens oftmals bedenkenlos in Kauf genommen. Das lässt sich einerseits durch psychologische Effekte erklären, denn aus psychologischer Sicht klaffen Risikorealität und Risikowahrnehmung oft auseinander. Sichtbar wird dies zum einen hinsichtlich der subjektiven Wahrnehmung von Eintrittswahrscheinlichkeiten. Bei sehr seltenen Ereignissen bzw. sinkenden Wahrscheinlichkeiten rückt das Wissen um die kausale Verknüpfung zwischen einer Handlung und deren Folgen tendenziell in den Hintergrund (vgl. Hansson, 2003, S. 292), während unmittelbare Vorteile der Mobilität in den Vordergrund treten. Zum anderen werden neuartige Risiken tendenziell eher überschätzt, etablierte hingegen unterschätzt. Risiken, die man glaubt steuern zu können, werden als weniger bedrohlich empfunden; stehen Risiken mehr in der öffentlichen Diskussion, verändert dies das subjektive Risikobewusstsein ebenfalls (vgl. Nida-Rümelin, 2005, S. 868–871; Rath, 2011, S. 120).

Andererseits wird auf gesamtgesellschaftlicher Ebene zwischen Kosten bzw. Risiken und Nutzen eines Mobilitätssystems abgewogen, wobei Sicherheit (*safety*) und Komfort (*convenience*) die zentralen Aspekte darstellen. Ist der gesellschaftliche Nutzen höher als die Kosten, werden Risikoübertragungen innerhalb des betreffenden Systems als zulässig bewertet und ein gewisses Risiko ist grund-

sätzlich zu tolerieren (vgl. Gasser, 2015, S. 548). Mobilität als gesellschaftliches System kann damit auf quasi-kontraktualistische Weise als im Kern gerechtfertigt erachtet werden (vgl. Müller & Gogoll, 2020, S. 1554). Dies muss umso mehr für automatisierte Fahrsysteme gelten, von deren Einführung man sich gerade eine signifikante Erhöhung der Verkehrssicherheit erhofft.

Derartige Argumentationsstrukturen sind zwar logisch valide und damit gesellschaftlich akzeptiert, aus (risiko-)ethischer Sicht jedoch prinzipiell problematisch. Philosophische Kritik rückt vor allem die Diskrepanz zwischen gesellschaftlichem und individuellem Nutzen in den Vordergrund. Sie richtet sich dabei gegen den konsequentialistischen Kern des Gedankengangs, unter dessen Primat individuelle Interessen im Rahmen einer gesamtgesellschaftlichen Nutzenabwägung dem großen Ganzen untergeordnet werden, wie es auch im Grundkonzept rationaler Risikopraxis verankert ist. Wie bereits die kritische Auseinandersetzung in Kap. 6.3.3 gezeigt hat, ist ein solches Vorgehen nicht haltbar. Verkehr ist ein soziales System, das sich an den Bedürfnissen des Einzelnen ausrichten muss und letztlich nur durch diese legitimiert werden kann.

Dennoch ist es plausibel anzunehmen, dass gesamtgesellschaftlich getragene Risiken des automatisierten Verkehrs nicht notwendigerweise für alle Individuen vorteilhaft sind; vielmehr ist davon auszugehen, dass gewisse Gruppen – i. e. die Halter bzw. Insassen eines selbstfahrenden Fahrzeugs – in den Genuss der Vorteile kommen, während andere vor allem die resultierenden Nachteile tragen. Auch das auf den ersten Blick einschlägige Argument, dass aufgrund des Sicherheitspotenzials autonomer Fahrzeuge alle Mitglieder einer Gesellschaft von einer erhöhten Verkehrssicherheit profitieren würden, kann die Skepsis gegenüber einer ethisch fragwürdigen Verteilung von Vor- und Nachteilen nicht aufheben. Erstens ist zweifelhaft, ob sich das visionäre Sicherheitsversprechen autonomer Fahrzeuge überhaupt erfüllen wird (siehe Kap. 2.2.3). Zweitens schließt eine allgemeine Verbesserung des Sicherheitsniveaus nicht aus, dass Individuen im Einzelfall schlechter gestellt werden können, z. B. in Situationen, die durch die rezeptiven und kognitiven Limitationen technischer Systeme erst verursacht werden (vgl. Brändle & Grunwald, 2019, S. 289).

An dieser Problematik setzt die Risikoethik an, die die Legitimität eines risikobehafteten Systems stets unter Bezugnahme auf Einzelin-

teressen begründet: Ein System, das Risiken produziert, kann nur unter Berücksichtigung distributiver Effekte ethisch gerechtfertigt werden, d. h. wenn die entstandenen Vor- und Nachteile für alle Betroffenen fair verteilt sind. So argumentiert Ferretti (2010, S. 506) von einem institutionalistischen Standpunkt aus, dass eine bestimmte Risikohandlung, die andere betrifft, unter diesen Voraussetzungen für eine gerechte Gesellschaft akzeptabel sein kann. Dies ist beispielsweise dann der Fall, wenn ein »equitable social system of risk-taking« (Hansson, 2003, S. 305) vorliegt, das auf der folgenden zentralen These basiert: Das *Prima-Facie*-Recht, keinen Risiken ausgesetzt zu werden, kann dann überschrieben werden, wenn die Risikoübertragung Teil eines gerechten Systems von Risiken ist, welches dem Einzelnen Vorteile bringt.[243]

Auch wenn die soziologisch begründete Akzeptanz aggregierter Nutzenabwägungen unabhängig von ihrem konkreten Anwendungskontext prinzipiell ethisch fragwürdig ist, wird sie zumeist nicht gesondert ethisch thematisiert, sondern als etabliertes gesellschaftsstrukturelles Phänomen betrachtet.[244] Ein möglicher Grund dafür mag darin liegen, dass die Risiken des herkömmlichen Straßenverkehrs durch die Annahme impliziter Zustimmung der Einzelnen als legitimiert angesehen werden, um das *Problem of Paralysis* – und damit einen drohenden Verlust der durch das Mobilitätssystem generierten Vorteile – zu umgehen. Ungeachtet dessen, dass eine solche Argumentationsbasis generell ethisch fragwürdig ist, muss sie angesichts des autonomen Fahrens erst recht kritisch adressiert werden. Mögliche Zustimmungsvorbehalte sind angesichts neuartiger technikinduzierter Risiken neu zu verhandeln (siehe Kap. 7.2.3.3); insbesondere dann, wenn die Nutzung autonomer Fahrzeuge tatsächlich eines Tages verpflichtend werden sollte.

---

243 Diese These Hanssons wird im weiteren Verlauf des siebten Kapitels noch intensiv erörtert.

244 Als einflussreich gilt in diesem Kontext das Modell des Ingenieurs Chauncey Starr (1969), der sich als einer der ersten Wissenschaftler mit der gesellschaftlichen Akzeptanz neuer technischer Risiken auseinandergesetzt hat. Ausgehend von einer Analyse der in Gesellschaften historisch nachweisbaren und traditionell akzeptierten Risiken hat er eine Theorie gesellschaftlicher Risikoakzeptanz entwickelt, in deren Rahmen Letztere nicht auf eine moralische Bewertung, sondern auf soziologisch feststellbare Gewohnheiten hinsichtlich der Risikoakzeptanz zurückgeführt wird.

Autonome Fahrsysteme verändern die Risikolandschaft des zukünftigen Straßenverkehrs auf disruptive Weise. Ein zentraler Grund, weshalb Mobilitätsrisiken im Kontext des autonomen Fahrens im Vergleich zu herkömmlichen Fahrzeugen ethisch stärker problematisiert werden, liegt in der Natur der betreffenden Risiken. Die ›neuen‹ Risiken sind technikinduziert, d. h. sie werden durch spezifische Funktionsweisen technischer Systeme in einer Weise transformiert, die ethisch problematisch ist: Diese Risiken sind systemisch und verändern das Wesen risikobehafteter Entscheidungen – aus einer intuitiv-situativen Reaktion in einer spontan auftretenden Verkehrssituation wird eine überlegte, bewusste Entscheidung, die sich in Algorithmen niederschlägt (vgl. Lin, 2015, S. 74; Nyholm & Smids, 2016, S. 1278–1279). Die im Rahmen der Fahrautomatisierung herbeigeführten Risiken sind in hohem Maße abhängig von einem fehlerfreien Betrieb der entsprechenden Systeme, der jedoch angesichts der generellen Fehleranfälligkeit von Maschinen nie zu vollständiger Sicherheit führen kann.

Wie Sahlin und Persson (1994, S. 38–53) betonen, müssen neben Risiken, die sich direkt auf zu erwartende Konsequenzen beziehen, auch epistemische Risiken Bestandteil einer validen Risikobewertung sein, welche die Zuverlässigkeit des vorausgesetzten Wissens hinterfragt. Hinzu kommt schließlich, dass Technik an sich grundsätzlich nicht wertneutral ist, sondern vielmehr geeignet, wesentliche Aspekte des menschlichen Lebens zu konditionieren (vgl. Mladenovic & McPherson, 2016, S. 1132–1133). Die Erkenntnis, dass eine explizite Risikobetrachung bei KI-Systemen unerlässlich ist, reifte spätestens seit dem Inkrafttreten der Datenschutz-Grundverordnung (DSGVO) im Jahr 2018 und der damit verbundenen Forderung nach der Etablierung von *Privacy-by-Design*-Ansätzen. Der im Herbst 2021 vom Ingenieurverband *IEEE* herausgegebene Ethikstandard für intelligente und autonome Systeme (*IEEE 7000*) mit dem Titel »IEEE Standard Model Process for Addressing Ethical Concerns during System Design« trägt dem zunehmenden Bewusstsein Rechnung, dass Ethik bereits von Beginn eines Entwicklungsprojekts an mitgedacht werden muss, um spätere negative Auswirkungen zu verhindern. Grundlage dieses Standards ist das von Sarah Spiekermann erarbeitete Prinzip des *Value Based Engineering*, welches das Ziel verfolgt, ethische Werte und Normen in allen Phasen der Systementwicklung zu berücksichtigen (*Ethics by Design*). Der Standard

enthält zehn Handlungsempfehlungen, zu denen auch eine risikobewusste und verantwortungsvolle Entwicklung von KI-Systemen zählt (vgl. Institute of Electrical and Electronics Engineers, 2021; Spiekermann & Winkler, 2022).

Vor diesem Hintergrund ist festzuhalten, dass die Aufgabe der ethischen Gestaltung von Unfallalgorithmen verschiedene Problemstellungen mit sich bringt. Ziel des dritten und letzten Teils der Forschungsarbeit ist es, ausgewählte Aspekte aus dem Blickwinkel einer risikoethischen Perspektive zu beleuchten und damit über den Horizont einer rein theoretischen Ebene hinauszugehen, auf welcher der bisherige Forschungsdiskurs in weiten Teilen stattfindet (vgl. Poszler et al., 2023, S. 19). Als Pointierung des inkompatiblen Aufeinandertreffens individueller Interessen werden Dilemma-Szenarien als besonders geeignet erachtet, dieses Ziel zu erreichen. Wenn sich Situationen, in denen es unweigerlich zu Schäden kommt – mindestens in dem Sinne, dass das Recht, keinem Risiko ausgesetzt zu werden, verletzt wird –, nicht völlig ausschließen lassen, dann verbleiben im Hinblick auf die automatisierte Mobilität nur zwei Möglichkeiten: Entweder sind Risikoübertragungen in jeglicher Form und rigoros unzulässig, was ein Verbot der Einführung des autonomen Fahrens an sich zur Folge hätte. Das Postulat eines Nullrisikos würde eine kategorische Zurückweisung jeglicher Risikoübertragungen implizieren und somit eine zu starke und absurde Einschränkung des Raums erlaubter Handlungen darstellen. Dies wiederum stünde dem Zweck der Moral als Ermöglichung guten Handelns zuwider (vgl. Hansson, 2003, S. 297–299; Hayenhjelm & Wolff, 2012, e37; Rippe, 2013, S. 534) und erscheint aus pragmatischer Sicht unplausibel. Oder es gibt spezifische Umstände, unter denen dieses Recht überschrieben werden darf und die es zu identifizieren gilt; Hansson beschreibt dies als »exemption problem« (2003, S. 302). Letztgenanntem Auftrag widmen sich die Ausführungen in diesem dritten Teil des Buches im Hinblick auf das Anwendungsproblem. Zunächst wird näher erläutert, inwiefern diese einen originären Beitrag zum Forschungsdiskurs leisten.

7.1.2 Unfallalgorithmen und Risikoethik: Ansätze bisheriger (risikoethischer) Forschung

Im Zuge dominanter Forschungszugänge, wie sie in Kap. 4 expliziert wurden, wird die Frage nach der Programmierung von Unfallalgorithmen als moralisches Designproblem konzipiert, das die praktische Gegebenheit unsicherer Handlungsfolgen zum entscheidungsrelevanten Zeitpunkt vernachlässigt. Zwar wird stellenweise angedeutet, dass Unsicherheit in Dilemma-Szenarien eine Rolle spielt;[245] konkrete Forschungsdesiderate werden daraus zunächst jedoch nicht abgeleitet. Erst im Zuge einer kritischen Auseinandersetzung mit den bis dato vorgeschlagenen Ansätzen einerseits sowie einer stärkeren, auch technischen Anwendungsorientierung andererseits rücken neuere Beiträge zunehmend Risikoaspekte in den Vordergrund. Diverse von Bund und Ländern finanzierte Forschungsprojekte haben die Bildung interdisziplinärer Forschungsteams forciert, die anwendungsnäher forschen als noch zu Beginn des Diskurses. Auf technischer Seite sind Risikoanalyse und Risikomanagement als zentrale Bestandteile von Forschung und Entwicklung fest verankert. Impulse aus dem ingenieurwissenschaftlichen Diskurs, der Risikoaspekte wie technikgenerierte Unsicherheiten bzw. Unreife der Technik bereits intensiv erforscht (vgl. Dietmayer, 2015; Hubmann et al., 2017), bereichern zunehmend verwandte Debatten in anderen Disziplinen.

So ist auch im ethischen Diskurs der letzten drei bis vier Jahre eine stetig steigende Anzahl an akademischen Beiträgen zu verzeichnen, die die Problemstellung unter Bezugnahme auf Komponenten aus Risikomanagement, Risikoanalyse und Risikoethik diskutieren. Grundlage dieser Entwicklung im deutschen Forschungsdiskurs sind nicht zuletzt auch die Richtlinien der Ethik-Kommission (Di Fabio et al., 2017), die eine Entscheidungsstrategie fordern, welche Risikoaspekte in systematischer Weise berücksichtigt. Der kürzlich in Kraft getretene und daher verstärkt in die öffentliche Wahrnehmung gerückte *AI Act* trägt ebenfalls erheblich dazu bei, dass risikobasierte Konzepte in Zukunft alternativlos sein werden, und daher verstärkt an Relevanz gewinnen. Das neue europäische KI-Gesetz

245 Benannt wird dieser Aspekt beispielsweise bei Bhargava und Kim (2017), Birnbacher und Birnbacher (2016), Davnall (2020), Goodall (2016a, 2016b, 2017), Grunwald (2015), Hübner und White (2018), Keeling (2020) sowie Nyholm und Smids (2016).

legt Unternehmen zukünftig die Verpflichtung auf, die mit den von ihnen entwickelten KI-Anwendungen in Zusammenhang stehenden Risiken mithilfe leistungsfähiger Test- und Validierungsverfahren umfassend zu identifizieren, zu analysieren, zu bewerten und entsprechend zu handeln. Die Durchführung von Folgenabschätzungen in Bezug auf die Grundrechte wird fortan von Betreibern von Hochrisiko-KI-Systemen, wie sie auch selbstfahrende Fahrzeuge darstellen, eingefordert (vgl. Future of Life Institute, 2024).

Ein Schwerpunkt der bisherigen Literatur zu Unfallalgorithmen, die sich risikoethischer Konzepte im weiteren Sinne bedient, liegt auf der Betrachtung von Fragen sozialer Gerechtigkeit.[246] Lebens (2017) prominente Anwendung von Rawls' *Differenzprinzip*, welches das Wohlergehen der am schlechtesten gestellten Personen in einer Gesellschaft fördert, bildet einen zentralen Ausgangspunkt kontroverser Diskussionen hinsichtlich der Analyse distributiver Effekte in Dilemma-Szenarien. In ähnlicher Weise verteidigt auch Smith (2022) eine konstruktivistische Rawls'sche Methode bei der Herleitung gerechtigkeitsethischer Prinzipien für Unfallalgorithmen. Im Zuge seiner Kritik an der Verwendung des Trolley-Problems fordert er eine institutionalistische Orientierung hin zu einem »distributiven Paradigma« in der normativen Analyse autonomer Fahrsysteme, in dessen Rahmen Fragen einer gerechten Verteilung ausgehend von gesellschaftlichen Kernstrukturen erörtert werden.

> […] we have considerable reason to adopt institutionalism about our moral obligations in the context of AVs. Rather than try to understand our obligations in the domain of AV by using trolley problems to interrogate how an ideal agent would behave and then trying to replicate that behavior for AV, we should determine the set of principles or rules that govern the fair distribution of benefits and burdens for the deployment of AVs. Our personal obligations, then, would be to attempt to instantiate those principles, but the primary normative principles would apply to institutions and practices. (Ebd., S. 286)

246 Dieser Fokus entspricht der gerechtigkeitsethischen Agenda, die unter dem Terminus ›Verkehrsgerechtigkeit‹ wissenschaftlich erforscht wird: »Transport justice describes a normative condition in which no person or group is disadvantaged by a lack of access to the opportunities they need to lead a meaningful and dignified life […] but also emphasises a more equal distribution of transport benefits and burdens in society […].« (Martínez-Buelvas et al., 2022, S. 3) Für einführende Beiträge in dieses Forschungsfeld siehe z. B. Karner et al. (2020), Martens (2016) sowie Pereira et al. (2017).

Smith unterstreicht die Bedeutung einer ganzheitlichen Sichtweise, die über die situativ beschränkten Konsequenzen einzelner Dilemma-Szenarien hinausgeht und die Wirkungen auf Strukturen sozialer Gerechtigkeit mitveranschlagt. Mladenovic und McPherson (2016) hingegen gehen im Rahmen ihrer Untersuchung gerechtigkeitsethischer Fragen über den Anwendungsfokus unvermeidbarer Unfallsituationen hinaus, indem sie die Problematik aus der Makroperspektive eines Systems vernetzter Straßenverkehrsüberwachung beurteilen.

Auch wenn diese Forschungsbeiträge nicht explizit risikoethisch ausgerichtet sind, können sie doch in gewissem Sinne als Grundlage für weiterführende risikoethische Forschung dienen. So kombinieren Dietrich und Weisswange (2019) Aspekte beider Ansätze, indem sie sowohl eine theoretische Begründung als auch eine praktische Implementierung eines Planungsmoduls für einen Prototypen vorlegen, welcher der explizit implementierten Regel des *Differenzprinzips* (»Minimiere die maximalen Kosten aller Beteiligten«)[247] folgt. Dabei legen sie den Schwerpunkt nicht auf Dilemma-Szenarien, sondern auf Routinesituationen mit geringen oder moderaten Risiken. In einem weiteren Beitrag entwirft Dietrich (2021) eine Architektur eines automatisierten Fahrzeugbetriebssystems, das ein spezielles Modul für die Risikoabschätzung und das Risikomanagement umfasst, sodass anfängliche Ungleichheiten bei der Risikoexposition ausgeglichen werden. Auch hierbei rekurriert er auf das Rawls'sche *Differenzprinzip* in seiner ursprünglichen Interpretation als institutionelle Zuweisungspolitik. In diesem Sinne plädiert er dafür, das automatisierte Fahren als institutionelle Aktivität zu gestalten (vgl. Dietrich, 2020).[248]

Weitere einschlägige Artikel fokussieren sich auf die Identifizierung und Analyse ethischer Entscheidungsmetriken und die Konzeption kontinuierlicher Entscheidungsprozesse für die Auswahl geeigneter Trajektorien, bei denen risikoethische Aspekte sowohl implizit als auch explizit berücksichtigt werden. Poszler et al. (2023, S. 15–16) geben eine Übersicht über entsprechende Ansätze, zu de-

247 Auf Rawls' *Differenzprinzip* wird in Kap. 7.3.3 vertiefend eingegangen.

248 Für eine Perspektive, die Risiko im Kontext (bio-)technologischer Innovationen und vor dem Hintergrund eines nicht-welfaristischen Verständnisses als eine Last der Kooperation auf den Schultern von Regulierungsinstitutionen interpretiert, siehe Ferretti (2010).

nen die *Data Theories Method* (vgl. Robinson et al., 2021), die *Ethical Valence Theory* (vgl. Evans et al., 2020), der *Expected Moral Value Approach* (vgl. Bhargava & Kim, 2017) sowie diverse Konzepte expliziter ethischer Algorithmen zur Trajektorienplanung zählen. Zu Letzteren gehört auch die von Németh (2022) vorgeschlagene Methode zur Routenauswahl, welche quantitative und qualitative Entscheidungsebenen integriert: Zunächst wird auf der Basis einer quantitativen Wahrscheinlichkeitsauswertung die unkritischste Route ermittelt, bevor bestehende Konflikte mittels zusätzlicher qualitativer ethischer Prinzipien entschieden werden. Auf ähnliche Weise operiert auch die Optimierungsfunktion im Rahmen des Steuerungskonzepts von Thornton et al. (2017).

Geisslinger et al. (2021) nehmen eine praxisorientierte Perspektive ein, indem sie (rationale) Entscheidungsprinzipien der Risikoethik auf die Trajektorienplanung autonomer Fahrsysteme anwenden. Dabei benennen sie zunächst interdisziplinäre Anforderungen, die Unfallalgorithmen im Hinblick auf eine mögliche Implementierung erfüllen müssen.[249] Von diesen Kriterien ausgehend, erarbeiten sie ein Framework, das die Spezifikationen der zentralen rationalen Entscheidungsprinzipien (Bayes-Regel, *Maximin*-Prinzip und Vorsorgeprinzip) in mathematische Gleichungen übersetzt und mittels einer komplexen, situativ gewichteten Kostenfunktion integriert. Dieser Ansatz ist aufgrund seiner pluralistischen Struktur flexibel auf eine Vielzahl möglicher Szenarien anwendbar, kann jedoch interne Konflikte und Widersprüche, die zwischen den einzelnen Prinzipien bestehen, nicht aufheben (vgl. Fossa, 2023, S. 76). In einem neueren Ansatz beschäftigen sich Geisslinger et al. (2023a) mit der Auswertung der Risiken möglicher Trajektorien für alle Verkehrsteilnehmer. Anhand einer Kategorisierung in vier Gültigkeitsstufen definieren sie gültige Trajektorien und identifizieren mittels einer ethischen Kostenfunktion diejenigen, die im Hinblick auf verschiedene konse-

249 Diese lauten: 1. Repräsentation der Realität, d. h. Berücksichtigung von Kontextfaktoren bei zweckmäßiger Vereinfachung komplexer Sachverhalte, 2. Technische Realisierbar- bzw. Formalisierbarkeit, 3. Allgemeingültigkeit bzw. Universalität für eine Vielzahl denkbarer Szenarien, 4. Soziale Akzeptanz als Voraussetzung für erfolgreiche Nutzeradoption, 5. Transparenz als Absicherung bezüglich rechtlicher Anforderungen (vgl. Geisslinger et al., 2021, S. 1037–1038).

quentialistische und deontologische Kriterien die niedrigsten Kosten verursachen.

Einer der zentralen Kritikpunkte an den bisher vorgelegten risikoethischen Ansätzen betrifft die primäre Folgenorientierung bei der Bewertung risikobehafteter Aktionen des Fahrsystems, infolge derer ethisch relevante Aspekte wie Handlungsmotivationen oder spezifische Pflichten vernachlässigt werden, ebenso wie potenzielle Vorteile aus Risikoübertragungen. Ein möglicher Grund für die Priorisierung konsequentialistischer Kriterien ist, dass bisherige Forschung sich der Problematik aus der Perspektive möglicher Implementierungen annähert. Wie bereits in Kap. 4.4.2 und Kap. 4.4.3 dargelegt, lassen sich insbesondere utilitaristische Kriterien einfacher formalisieren als z. B. deontologische. Es stehen Aspekte einer fairen Verteilung von Risiken einerseits (vgl. Dietrich, 2021; Dietrich & Weisswange, 2019) sowie der Risikobewertung und Risikoabschätzung mittels mathematischer Kostenfunktionen andererseits (vgl. Geisslinger et al., 2021, 2023a; Németh, 2022; Thornton et al., 2017) im Fokus. Indem sie sich ausschließlich auf Kriterien rationaler Risikopraxis stützen, gelingt es den bisherigen Ansätzen jedoch nur begrenzt, zentrale ethische Probleme zu klären.

Unter Bezugnahme auf die fünf Problemfelder der Risikoethik, die in Kap. 6.2.1 beschrieben wurden, bleiben im Hinblick auf die bisherige risikoethische Forschung wesentliche Fragen offen. Beispielsweise sind Summenbildungen, wie sie in Kostenfunktionen meist erfolgen, immer in der einen oder anderen Form aggregativ, wohingegen aus risikoethischer Sicht stets die Perspektive der Einzelnen einzunehmen ist. Die Beurteilung komplexer Beziehungsgefüge zwischen Akteuren steht im Zentrum der Risikoethik, wird in gegenwärtigen Implementierungsversuchen jedoch kaum berücksichtigt. So versäumen es die Ansätze bislang, verantwortungsethische Aspekte umfassend in Betracht zu ziehen.[250] Geisslinger et al. (2021) unterstreichen die Notwendigkeit, ihren Entwurf durch die Integration spezieller Gewichtungsfaktoren weiterzuentwickeln, um Verantwortlichkeiten formal zu repräsentieren. Im Rahmen eines neueren Ansatzes versuchen Geisslinger et al. (2023a), konkrete

250 Auch in dieser Forschungsarbeit spielt der Aspekt der Verantwortung keine explizite Rolle in der inhaltlichen Gestaltung des risikoethischen Entwurfs, ist jedoch in Form der Prämissen der deontologischen Grenzkriterien, die in Kap. 7.3.2 und Kap. 7.3.3 begründet werden, implizit enthalten.

Ausprägungen von Verantwortung über die Reduzierung der Risikokosten innerhalb des bayesianischen Funktionsterms auszudrücken. Ihre Auffassung ist dabei einerseits sehr spezifisch; andererseits ist generell fraglich, inwiefern sich (moralische) Verantwortung in quantifizierbaren Faktoren überhaupt adäquat abbilden lässt. Formale Modellierungen, die für Implementierungen unverzichtbar sind, stellen letztlich immer Vereinfachungen dar; komplexe Konzepte können bei einer Darstellung in für Maschinen verständlichen Formeln meist nur verkürzt wiedergegeben werden, wobei u. U. essenzielle Aspekte verloren gehen oder in ihrem ethischen Kern nicht adäquat abgebildet werden. So versuchen Geisslinger et al. (ebd.), eine auf den Anwendungskontext reduzierte Auffassung von Verantwortung in ihre Kostenfunktion zu integrieren, die der Vielschichtigkeit des ethischen Konzepts an solches nicht gerecht werden kann.

Dietrich (2021) hingegen argumentiert zwar explizit verantwortungsethisch, um zu begründen, weshalb mögliche Ungerechtigkeiten bei der Programmierung von Systemalgorithmen aktiv durch die Integration von Gerechtigkeitszielen thematisiert werden müssen. Für deren konkrete Implementierung greift er jedoch wiederum auf die nach Rawls'schem Verständnis zwar kontraktualistisch begründete, aber konsequentialistisch operierende *Maximin*-Regel zurück, welche ihrerseits ethische Schwierigkeiten aufweist (siehe Kap. 6.3.2.2).[251]

Sowohl Geisslinger et al. (2023a) als auch Dietrich (2021) verzichten darauf, die Rechtfertigbarkeit der verwendeten rationalen Prinzipien kritisch zu prüfen. Einen entsprechenden argumentativen Versuch legt Berkey (2022, S. 217–222) vor, indem er eine faire Zuweisung von Risiken im Kontext automatisierter Mobilität fordert:

> […] firms ought to program autonomous vehicles in ways that aim to ensure that the safety-related benefits of those vehicles are distributed as fairly as possible among all of those who could potentially be harmed as a result of the use of motor vehicles. (Ebd., S. 216)

251 Als Grundlage einer verantwortungsethischen Diskussion technologiebezogener Risiken siehe Van de Poel und Nihlén Fahlquist (2013), die das Verhältnis von Risiko und Verantwortung sowie Verantwortungskonzeptionen, insbesondere das *Problem of Many Hands*, thematisieren.

Er diskutiert mögliche Gründe, mit denen ein Abweichen von anfänglich gleichverteilten Risiken gerechtfertigt werden könnte. Ein solcher besteht zum einen in der Vorrangstellung, die er der Reduzierung des absoluten Risikolevels bzw. des Gesamtschadens gegenüber dem Postulat gleichverteilter Risiken einräumt: »The aim of reducing injuries and deaths as much as possible should, on my view, take priority over distributing risks equally when there is a conflict.« (Ebd., S. 225) So würde die Annahme, dass autonome Fahrzeuge das Sicherheitsniveau erhöhen, implizieren, dass im Mischverkehr diejenigen weniger Risiko tragen sollten, die anstelle von konventionellen selbstfahrende Fahrzeuge nutzen.[252] Zum anderen ist eine Differenzierung zwischen Insassen motorisierter Verkehrsmittel einerseits und Fußgängern bzw. Radfahrern andererseits ethisch relevant, da Letztere signifikant niedrigere Risiken für andere heraufbeschwören. In diesem Kontext sieht Berkey vor allem eine prinzipielle Priorisierung von Fahrzeuginsassen kritisch.

Auch Abney (2022) deutet mögliche Grundpfeiler einer dezidiert risikoethischen Perspektive an. Im Kontext von Missbrauch und Manipulation von Systemalgorithmen beschreibt er verschiedene Faktoren, die die Akzeptabilität eines Risikos konstituieren. Zu diesen zählen eine (vollumfängliche) Zustimmung bzw. Freiwilligkeit hinsichtlich eines eingegangenen Risikos, der Umfang der betroffenen Gruppen (Beteiligte vs. Unbeteiligte), die Natur relevanter Risiken (»state risk« vs. »step risk«), Schwellenwerte bei Ausmaß und Wahrscheinlichkeiten sowie die Legitimation von Entscheidungsinstanzen. Er erörtert diese Aspekte jedoch nicht näher, sondern formuliert sie lediglich als Forschungsimpulse, die in Richtung einer fundierten risikoethischen Analyse weisen. Eine solche zu skizzieren ist das Anliegen der vorliegenden Forschungsarbeit. Nachfolgend werden zunächst deren Gegenstand und Ziele näher spezifiziert.

7.1.3 Gegenstand und Ziele eines alternativen risikoethischen Entwurfs

Zentrales Ziel der vorliegenden Arbeit ist es, den Forschungsdiskurs moralischer Unfalldilemmata weiterzuentwickeln, indem untersucht

252 Plausibel ist dies jedoch nur unter der Voraussetzung, dass alle Individuen potenziell gleiche Zugangschancen zu automatisierter Mobilität haben.

wird, welchen Beitrag eine risikoethische Perspektive zur Klärung offener Fragen leisten kann. In Anbetracht der im vorhergehenden Unterkapitel aufgezeigten Forschungslücke ist festzuhalten, dass dem Forschungsdiskurs eine systematisch entwickelte, dezidiert risikoethische Grundlage fehlt, auf der sich sowohl Dilemma-Szenarien als auch Routinefahrsituationen ganzheitlich ethisch wie auch gesellschaftlich-sozial interpretieren lassen. Der hier vorgelegte Entwurf knüpft sowohl an diejenigen Beiträge an, welche die Problematik aus einem implementierungsnahen Blickwinkel betrachten, als auch an diejenigen, die sich auf distributive Aspekte fokussieren. Im Gegensatz zu Ersteren wird in dieser Arbeit unabhängig von (technischen) Erfordernissen möglicher Implementierungsansätze eine ethische Perspektive entworfen, ohne jedoch technisch relevante Spezifika des Anwendungsproblems gänzlich aus den Augen zu verlieren. Es werden keine konkreten Implementierungsprinzipien erarbeitet, sondern grundlegende ethische Fragen diskutiert, aus denen sich normative Implikationen für Implementierungsversuche ableiten lassen. Diese Herangehensweise trägt dem interdisziplinären Charakter der Fragestellung insofern Rechnung, als sie betont, dass auch anwendungsnahe Forschung nicht ohne fundierte, theoretische (ethische) Grundlagen auskommt, die zunächst unabhängig von (formalen) Realisierungsfragen zu klären sind. Auf diese Weise soll vermieden werden, dass es durch eine zu starke Fokussierung auf die technische Realisierung bzw. Realisierbarkeit zu einer Vernachlässigung der Plausibilität und Überzeugungskraft – und damit der Qualität – der zugrundegelegten ethischen Argumente kommt.

Wie in Kap. 4.2 gezeigt, zeichnen sich Dilemma-Szenarien im Kontext automatisierter Mobilität durch eine gesellschaftlich-soziale Dimension aus, der im Zuge des moralphilosophisch dominierten Zugangs kaum Beachtung geschenkt wurde. Der in diesem dritten Teil entwickelte Entwurf kann als Versuch verstanden werden, der Forderung nach einer institutionalistischen Problemperspektive zu entsprechen, wie sie beispielsweise von Smith (2022) formuliert wird. Unter Bezugnahme auf das »distributive Paradigma« wird an die bisherigen risikoethischen Beiträge angeknüpft und zugleich über diese hinausgegangen. Dabei werden u. a. Rawls'sche Konzepte vor dem Hintergrund einer risikoethischen Gesamtinterpretation von Unfallszenarien und als Teil eines deontologischen Entwurfs in-

tegriert, während bisherige Forschungsarbeiten in einer ausschließlich distributiven Perspektive verbleiben.

In der nachfolgend dargestellten risikoethischen Auseinandersetzung werden ausgehend von den in Kap. 6 vorgestellten theoretischen Grundlagen zentrale risikoethische Grundfragen in Bezug auf das Anwendungsproblem auf systematische Weise thematisiert. Einzelne risikoethische Argumente werden verknüpft und schließlich im Rahmen einer Skizze der Grundzüge deontologischer Risikoethik zusammengeführt. Der auf diese Weise erarbeitete risikoethische Entwurf kann als Alternative zu den in Teil II rekonstruierten Diskursen begriffen werden und bietet neue Impulse für zukünftige Auseinandersetzungen mit Unfallalgorithmen. So werden mögliche Ansätze der Rechtfertigung der moralischen Zulässigkeit von Risikoübertragungen anhand von Dilemma-Szenarien diskutiert, welche im Kontext der automatisierten Mobilität auftreten können. Diese dürfen jedoch nicht so strikt sein, dass jegliche praktische Relevanz verloren geht; dies wird vor allem hinsichtlich des in Kap. 5 erarbeiteten metaethischen Postulats der Nicht-Verrechenbarkeit inkommensurabler Werte problematisiert. Zudem werden Entscheidungsstrategien bezüglich der Problematik einer gerechtfertigten Verteilung individueller Vor- und Nachteile im spezifischen Problemkontext kritisch erörtert, wobei deontologische Prinzipien – anders als in den bisherigen implementierungsnahen Ansätzen – als *hard constraints* aufgefasst werden.

Den formalen Anforderungen hinsichtlich des begrenzten Umfangs dieser Arbeit ist es geschuldet, dass dabei nicht alle Themenkomplexe[253] der Risikoethik gleichermaßen intensiv bearbeitet werden können. Einige werden nur angedeutet und bilden weiterführende Forschungsdesiderate (siehe Kap. 8.2). Die risikoethisch relevanten Aspekte der (numerischen) Risikobewertung und Risikoabschätzung sowie verantwortungsethische Fragen werden im Folgenden daher nur gestreift, jedoch nicht umfassend behandelt. Im nächsten Unterkapitel wird die risikoethische Analyse anhand von zentralen Konzepten und Kriterien systematisch entfaltet.

253 Siehe die fünf Problemfelder der Risikoethik aus Kap. 6.2.3.

7.2 Analyse der Risikokonstellationen in Dilemma-Szenarien entlang von Kriterien der Risikoakzeptabilität

7.2.1 Akteure, Beziehungsnetzwerke und private Risiken

Wie in Kap. 4 ausgeführt wurde, lässt der bisher dominante Forschungszugang zentrale ethische Fragen im Kontext der Gestaltung von Unfallalgorithmen ungeklärt; der Fokus auf die (moralphilosophische) Frage nach der ethisch ›richtigen‹ Entscheidung ergibt keine final rechtfertigbaren Strategien. In diesem Kapitel wird eine risikoethische Auseinandersetzung vorgelegt, welche es erlaubt, einen alternativen Betrachtungswinkel auf das Anwendungsproblem einzunehmen, der Impulse für neue Entscheidungsperspektiven bereitstellt. In einem ersten Schritt wird zunächst eine risikoethische Interpretation des zugrundeliegenden Entscheidungsproblems skizziert, das durch seine dilemmatische Struktur eine besondere Risikokonstellation begründet.[254]

Risikoethische Überlegungen kommen immer dann zum Zuge, wenn ethisch relevante Unsicherheiten im Hinblick auf Handlungsfolgen bestehen. Grundlage jeder risikoethischen Betrachtung ist die Annahme, dass, auch wenn es moralisch *unzulässig* ist, einem Individuum eine bestimmte Konsequenz *direkt* zuzufügen, dennoch Umstände vorliegen können, die es *erlauben*, das Individuum einem entsprechenden *Risiko* auszusetzen, infolgedessen die Konsequenz möglicherweise eintreten kann. Hinsichtlich der Gestaltung von Unfallalgorithmen kann zum Zeitpunkt der Entscheidungsfindung, und damit auch der ethischen Bewertung entsprechender Algorithmen, nur von Risiken, nicht aber von tatsächlichen Schädigungen ausgegangen werden:

> The harm that interests us in the case of traffic is not one of intentionally harming third parties. Our main focus lies on actions that might cause harm to others but usually do not, meaning actions that put other members of society at risk, even if the risk of getting harmed for each individual is relatively low. (Müller & Gogoll, 2020, S. 1552)

254 Die Grundstruktur einer risikoethischen Perspektive auf Unfalldilemmata hat die Autorin bereits an anderer Stelle in groben Zügen entwickelt (vgl. Schäffner, 2020b).

Gegenstand des ethischen Entscheidungsproblems sind also nicht tatsächliche Schäden, sondern vielmehr Schadens*risiken*. Weiterhin folgt bereits aus der Definition von Dilemmata, dass Schaden in diesen spezifischen Situationen unausweichlich ist, d. h. für mindestens ein beteiligtes Subjekt muss plausiblerweise das Eintreten einer negativen Konsequenz angenommen werden.

Diese Betrachtungsweise wirft die Frage auf, wie mit drohenden Schäden (= Risiken) in Dilemma-Szenarien umgegangen werden soll. Die Programmierung von Unfallalgorithmen lässt sich als risikoethisches Verteilungsproblem interpretieren; die Frage nach der moralphilosophischen Begründung von Entscheidungsprinzipien für Unfalldilemmata, die im Zentrum des dominanten Forschungszugangs steht, wird dabei transformiert in die risikoethische Frage der Rechtfertigbarkeit von Dilemma-Risiken sowie der fairen Verteilung daraus resultierender Vor- und Nachteile. Im Rahmen der nachfolgenden Auseinandersetzung wird das Anwendungsproblem vor dem Hintergrund zweier miteinander zusammenhängender risikoethischer Grundfragen neu entworfen: Einerseits stellt sich die Frage, wie sich entstehende Risikoübertragungen moralisch rechtfertigen lassen; andererseits geht es darum, unweigerlich bestehende Risiken in einer Weise zuzuweisen, die ethisch plausiblen und validen Kriterien genügt. Beide Problemstellungen sind eng miteinander verwoben und lassen sich nicht vollständig unabhängig voneinander untersuchen. Aus diesem Grund werden beide Aspekte im Folgenden stets im Zusammenhang betrachtet.

Als sozio-technische Systeme bewegen sich autonome Fahrzeuge in einem komplexen Verkehrsgeschehen, das durch das Zusammenwirken miteinander interagierender und aufeinander reagierender Subjekte charakterisiert ist. Dabei sind am Straßenverkehr teilnehmende Personen zum einen jenen Risiken ausgesetzt, die durch andere Verkehrsbeteiligte entstehen, und zum anderen verursachen sie selbst Risiken für andere; diese Praxis wechselseitiger Risikoübertragungen gilt sowohl für die Insassen autonomer Fahrzeuge als auch für andere Verkehrsteilnehmer (vgl. Grunwald, 2015, S. 667). Das zentrale Element, das die Risikokonstellationen im Kontext von Unfalldilemmata kennzeichnet, besteht in der Reziprozität von Risikoübertragungen. Jedes aktive Eingreifen in das Verkehrsgeschehen – ob Ausweichmanöver oder Neupositionierung innerhalb einer Fahrspur, Überqueren einer Straße oder Überholmanöver – stellt eine

dynamische Umverteilung von Risiken dar. Sicherheit und Risiko der Einzelnen hängen daher entscheidend von den Handlungen anderer ab:

> The choice situation—in terms of levels of autonomy and risk to driver—is almost entirely determined by others. Drivers may select certain ›categories‹ of autonomy or risk, but they have very little control over the delineation of the categories themselves. The expected consequences of these decisions are deeply entangled with the choices of others on the road, from pedestrians to animals to cyclists to other vehicles (some of whom may not have drivers at all)—and attempts to manage these risks are dependent on the other actors. (Smith, 2022, S. 284)

Auf diese Weise übertragene Risiken existieren nicht einfach; sie sind vielmehr Bestandteil eines Beziehungsnetzwerks, welches die Akteure durch ihre wechselseitige Risikourheberschaft generieren: »Risks do not just ‚exist' as free-floating entities; they are taken, run or imposed. Risk-taking and risk imposition involve problems of agency and interpersonal relationships [...].« (Hansson, 2007b, S. 27) Risiken sind stets an Handlungen gebunden und etablieren ein spezifisches Beziehungsgefüge zwischen Entscheidenden (*decision-makers*), Betroffenen (*risk-exposed*) und potenziellen Begünstigten (*beneficiaries*). Für die risikoethische Beurteilung konkreter Risiken ist es essenziell, wie sich diese drei Rollen zueinander verhalten: Wer trifft die Entscheidung über eine Risikoübertragung und mit welcher Absicht? Profitieren Entscheidende oder andere Personen davon, Dritte einem Risiko auszusetzen? Findet eine Kompensation von Betroffenen durch Begünstigte statt? (Vgl. ebd., S. 27–28)

Im Zuge der Konkretisierung derartiger Fragestellungen auf das Anwendungsproblem von Unfallalgorithmen sind zunächst die risikoethisch relevanten Akteure zu spezifizieren. Bezogen auf Szenarien der Risikoübertragung können unter ›Entscheidenden‹ diejenigen verstanden werden, welche eine risikobehaftete Handlung initiieren, in deren Folge Risiken für andere entstehen. In der Regel sind in diesem spezifischen Kontext Entscheider und Begünstigte gleichzusetzen, denn diese sind durch ihre (risikoverursachenden) Aktionen in der Lage, ihre eigenen Projekte zu verfolgen, und profitieren daher. Dies könnte beispielsweise auf die Insassen eines autonomen Fahrzeugs zutreffen, die durch die automatisierte Beförderung Zeit für andere Tätigkeiten wie Lesen oder Arbeiten gewinnen.

Da sie sich jedoch durch ihre Teilnahme am Straßenverkehr immer zugleich auch selbst einem Risiko aussetzen, können sie – u. U. und in Abhängigkeit von der spezifischen Konstellation – auch zu den Betroffenen zählen, ebenso wie jede andere beteiligte Person. Weiterhin ergeben sich aus der Analyse der Beziehungsverhältnisse zwischen Risikoakteuren auch Implikationen für die Identifikation relevanter Risikokonstellationen. Diese werden zunächst nur neutral dargestellt und erst in einem zweiten Schritt ethisch bewertet bzw. hinsichtlich ihrer Akzeptabilität diskutiert.

Was die Beziehung zwischen einem Risiko und den von diesem Betroffenen angeht. erscheint es strittig, inwiefern individuelle Risiken, in denen sich Personen selbst einem (privaten) Risiko aussetzen, ohne dass Auswirkungen für andere bestehen, im Kontext automatisierter Mobilität als relevant anzusehen sind. Wie Nida-Rümelin et al. (2012, S. 154) anmerken, stellt das Konzept der allgemeinen Gefährdungshaftung klar, dass allein durch die »Beteiligung am Straßenverkehr auch bei korrektem Verhalten ein gewisses Risiko für andere Personen« besteht. Eine Teilnahme am Straßenverkehr ist daher nicht widerspruchsfrei ohne die Verursachung von Externalitäten denkbar. Dennoch kann eine individuelle Risikosituation vorliegen, denn eine solche kann unter bestimmten Voraussetzungen auch durch ein Kollektiv verursacht werden:

> Alle Mitglieder müssen willentlich diesem Kollektiv beigetreten und eine mögliche ungleiche Verteilung von Konsequenzen, resultierend aus einer Risikosituation, muss von allen akzeptiert sein. Sind diese Bedingungen erfüllt, dann kann von einer individuellen Risikosituation gesprochen werden, dessen Urheber ein Kollektiv ist. (Ebd., S. 29)

Es ist grundsätzlich denkbar, ein solches Kollektiv dynamisch als die Gesamtheit aller Verkehrsbeteiligten aufzufassen, die sich zu einem gewissen Zeitpunkt quasi ›im System‹ befinden und deren jeweilige Risiken innerhalb der Systemgrenzen bleiben; die Teilnahme am Verkehr kann dann als willentliche Akzeptanz der ›Spielregeln‹ ausgelegt werden. Fraglich ist allerdings, inwiefern diese Zustimmung mangels Alternativen freiwillig erfolgt. Ebendies greifen Nida-Rümelin et al. in ihrer Darstellung eines weiteren Typus von Risikosituationen auf, »in denen zwar ein Risiko übertragen wird, dies aber durch die Zustimmung eines ausreichend informierten Risiko-Betroffenen in Abwesenheit jeglichen Zwangs legitimiert wird.« (Ebd., S. 29) Ist die Zustimmung freiwillig und auf Basis ausreichender

Informationen erfolgt, so gilt das Risiko als kollektiv eingegangen, wobei die potenziellen Konsequenzen nicht über die beteiligten Individuen hinausgehen dürfen.

Dennoch ist es im Hinblick auf das Anwendungsproblem nicht sinnvoll, Dilemma-Szenarien primär als individuelle Risikosituationen zu interpretieren. Erstens sind individuelle Risikosituationen mit Konsequenzen katastrophalen Ausmaßes definitionsgemäß unvereinbar, was den praktischen Gegebenheiten jedoch widerspricht, wie in den nachfolgenden Unterkapiteln gezeigt wird. Zweitens ist anzumerken, dass – vorausgesetzt, die von Nida-Rümelin et al. beschriebenen Sonderfälle individueller Risikosituationen würden hier greifen – sich die risikoethischen Fragen nach der Zulässigkeit von Risikoübertragungen bzw. den Kriterien einer fairen Risikoverteilung gar nicht oder nur in sehr begrenztem Maße stellen. Aufgrund der unterschiedlichen Voraussetzungen hinsichtlich individueller Risikoverursachung und Risikotoleranz sind erwarteter Nutzen und erwartete Risiken zwischen Anspruchsgruppen in der Regel derart asymmetrisch verteilt bzw. umstritten, dass eine freiwillige Akzeptanz derselben nicht zweifelsfrei angenommen werden kann. Daher erscheint es sowohl praktisch als auch ethisch notwendig, diese Fragen zu klären. Aus diesem Grund richtet sich der Fokus der vorgelegten risikoethischen Auseinandersetzung auf übertragene Risiken, d. h. Situationen, in denen ein Individuum oder eine Gruppe von einem Risiko betroffen ist, welches nicht selbstverursacht ist; das Risiko geht von den verursachenden Personen auf den bzw. die Betroffenen über (vgl. Rath, 2011, S. 30). Die real-lebensweltliche Manifestation dieser Konstellation ist im Kontext des autonomen Fahrens auf vielfältige Weise denkbar. Im Folgenden werden zur Veranschaulichung der Problematik zunächst verschiedene relevante Szenarien kurz umrissen.

7.2.2 Szenarien der Risikoübertragung

Aufgrund der Vielfalt möglicher Szenarien, in denen moralische Dilemmata mit Beteiligung autonomer Fahrzeuge auftreten können, sind Risikoübertragungen in verschiedenen Beziehungskonstellationen denkbar. Diese lassen sich hinsichtlich ihrer risikoethischen Problematik kategorisieren. Im einfachen Unfallszenario gehen von einem selbstfahrenden Fahrzeug Risiken für andere Verkehrsbetei-

ligte aus, wie in den Beispielszenarien 1 ›Großmutter versus Kind‹ und 2 ›Einzelperson versus Gruppe‹. Umgekehrt verursachen all diese Akteure ebenfalls Risiken für im konkreten Szenario involvierte Parteien, i. e. die Insassen des autonomen Fahrzeugs sowie die jeweils anderen Fußgänger(gruppen). Diese klassische Dilemma-Konstellation kann aus risikoethischer Sicht als Standardfall gelten, der sich durch komplexere ethisch relevante Beziehungsverhältnisse erweitern und ausdifferenzieren lässt.

Bevor diese Varianten im Folgenden erläutert werden, ist zunächst eine Anmerkung zur Situationsstruktur nötig. Da Schaden in moralischen Dilemmata definitionsgemäß unvermeidbar ist, stellen diese Nullsummenspiele dar. Dieses ursprünglich aus der Spieltheorie stammende Konzept impliziert im ökonomischen Sinne, dass die Summe aller Gewinne auf dem Markt stets null ist, weshalb der Erfolg eines Marktteilnehmers immer zulasten des Verlustes eines anderen geht. Übertragen auf den Kontext von Dilemma-Szenarien bedeutet dies, dass die Summe aller Risikowahrscheinlichkeiten stets ›1‹ ergibt, d. h. wenn sich das Risiko einer Partei reduziert, erhöht sich dasjenige der anderen. Daraus folgt, dass risikorelevante Handlungen in geschlossenen Systemen stets mit der Übertragung von Risiken verbunden sind. Voraussetzung dafür ist die Annahme, dass Risikokonstellationen dynamisch sind; jede noch so geringe Aktion eines jeden Beteiligten hat das Potenzial, die Risikokonstellation grundlegend zu verändern.

Hinsichtlich der Schadenskomponente des Risikobegriffs besteht ein solcher interpersoneller Zusammenhang zwischen den Risiken der Beteiligten nur indirekt bzw. in Fällen, in denen risikorelevante Aktionen direkte Externalitäten aufweisen. So senkt die Wahl schwerer Fahrzeugtypen auf der einen Seite tendenziell das zu erwartende Schadensausmaß für die Insassen, erhöht dieses aber zugleich für Radfahrer oder Fußgänger. Auf der anderen Seite wirken sich erhöhte individuelle Sicherheitsmaßnahmen wie das Tragen spezieller Schutzkleidung durch Motorradfahrer nicht direkt auf die zu erwartenden Schadenshöhen anderer aus – wohl aber indirekt auf deren Risiken über die Wahrscheinlichkeitskomponente, wie es in Beispielszenario 4 ›Motorradfahrer mit/ohne Helm‹ unter der gleichzeitigen Annahme eines utilitaristischen Schadenskalküls der Fall ist.

Vor diesem Hintergrund konstituieren Handlungen, die sowohl in Bezug auf Eintrittswahrscheinlichkeit als auch auf Schadenshöhe

direkt oder indirekt Einfluss auf Risikoübertragungen nehmen, verschiedene vom Standardfall abweichende Dilemma-Varianten. Eine solche ist beispielsweise auf verkehrswidriges Verhalten zurückzuführen. So kommt es in Beispielszenario 3 ›Rote Ampel‹ aufgrund des regelwidrigen Überquerens einer Fußgängerampel zu einer Risikoverschiebung zu Ungunsten des nachfolgenden Verkehrs, in Beispielszenario 6 ›Tunnel‹ ist das abrupte Abbremsen die Ursache für eine erhöhte Risikoexposition desselben. Einen ethisch komplexen Spezialfall innerhalb dieser Kategorie stellen Szenarien dar, in denen Risikoübertragungen durch bewusstes Fehlverhalten bzw. manipulatives Verhalten provoziert werden (vgl. Abney, 2022, S. 259–260; Hevelke & Nida-Rümelin, 2015c, S. 19–20). Schützen Algorithmen z. B. im Zweifelsfall stets Fußgänger, so könnten sich Letztere durch rücksichtsloses Handeln einen Sicherheitsvorteil verschaffen. Auch dem ohne Helm fahrenden Motorradfahrer in Beispielszenario 4 ›Motorradfahrer mit/ohne Helm‹ bieten sich im Falle eines auf die Realisierung der geringsten Schadenssumme ausgerichteten Algorithmus Fehlanreize, aufgrund derer die betreffende Person Teile ihres eigenen Risikos auf andere übertragen kann.[255]

Eine weitere Kategorie von Risikoübertragungen manifestiert sich in Szenarien, in denen neben beteiligten Parteien auch Unbeteiligte gewissen Risiken ausgesetzt werden, wie die Fußgängerin in Beispielszenario 5 ›Unbeteiligte auf Bürgersteig‹ oder die Kinder in Beispielszenario 8 ›Herannahender LKW‹.[256] Eine ethisch besonders kontroverse Variante sind sogenannte Selbstopferungsszenarien, in denen das Steuerungssystem des autonomen Fahrzeugs, wie in den Beispielszenarien 6 ›Tunnel‹ und 7 ›Klippe‹, eine Kollision mit unbelebter Infrastruktur bevorzugt, in deren Folge die Insassen zu Schaden kommen, um andere Verkehrsteilnehmer zu schützen. In

255 Auch wenn diese Szenarien meist mit einer an spezifischen ethischen – meist utilitaristischen – Prinzipien ausgerichteten Programmierung der Algorithmen assoziiert sind, ist ein Ausnutzen von deren Funktionsweisen grundsätzlich unabhängig davon möglich, welche konkreten Prinzipien zugrunde gelegt werden, sofern diese transparent und allgemein bekannt sind.

256 Zur ethischen Relevanz der Differenzierung zwischen Beteiligten und Unbeteiligten wurden bereits in Kap. 4.4.1 einige Argumente aus der philosophischen Ethik erläutert. Auch in Kap. 7.3.3 wird diese Problematik nochmals risikoethisch aufgegriffen; an dieser Stelle genügt vorerst die Bemerkung, dass sich die Gesamtkonstellationen in relevanten Szenarien durch Risikoübertragungen auf Unbeteiligte verändern.

diesen Fällen wirken die Risiken, die von den Passagieren durch die Nutzung des autonomen Fahrzeugs auf andere externe Parteien übertragen wurden, gewissermaßen auf ihre Initiatoren zurück.

Im Hinblick auf die Tragweite bzw. das Ausmaß verursachter Risiken lassen sich die in Unfallszenarien relevanten Risikosituationen weiter spezifizieren. Da Dilemma-Situationen selten auftreten, sind die Eintrittswahrscheinlichkeiten entsprechend zu erwartender Schäden sehr gering. Dies wäre im Sinne spezifischer Positionen der Risikoethik als triviales Risiko zu werten. Für Thomson (1985a, S. 128–137) liegt ein solches Risiko genau dann vor, wenn die Eintrittswahrscheinlichkeit hinreichend klein ist, was sie anhand ihrer prominenten Beispiele des Gasherds einerseits und des Russischen Roulettes andererseits demonstriert. Mögliche Konsequenzen sind dabei jeweils irrelevant. Hinsichtlich der spezifischen situativen Charakteristika von Unfallszenarien erscheint es jedoch implausibel, diese als triviale Risikosituationen zu kategorisieren. Wie die in Kap. 3.2.3 aufgeführten Argumente belegen, sind Individuen im Hinblick auf mögliche Risiken bzw. Erwartungswerte in Dilemma-Situationen keineswegs indifferent, was nach Nida-Rümelin et al. (2012, S. 47) jedoch eine Voraussetzung für das Vorliegen eines trivialen Risikos sein müsste.

Andere hingegen definieren triviale Risiken vorrangig über die Tragweite eines Risikos; so betont Posner (2004, S. 141), dass ein triviales Risiko nur dann vorliegt, wenn die Konsequenzen zumindest nicht katastrophal sind, ungeachtet ihrer Eintrittswahrscheinlichkeit. Dies lässt sich auf zweierlei Weisen auf Dilemma-Szenarien übertragen: Zum einen sorgen die systematischen Effekte der in einer großen Zahl von Fahrzeugen wirksamen, identischen Steuerungssoftware dafür, dass diese nicht nur in einer konkreten Situation Anwendung findet, sondern potenziell viele Fälle betrifft, von denen nicht bekannt ist, wann, wo und wie sie auftreten werden. Die entsprechenden Risiken werden also quasi auf die Gesamtheit aller Verkehrsbeteiligten übertragen. Zum anderen sind Szenarien denkbar, in denen unmittelbare Risiken katastrophalen Ausmaßes verursacht werden, z. B. wenn es im Zuge einer Kollision zum Austreten von chemischen Schadstoffen oder größeren Mengen Treibstoff kommt, was wiederum eine große Zahl von Menschen auf katastrophenähnliche Weise tangiert. Auch wenn diese Szenarien berechtigterweise als unwahrscheinlich gelten können, sind Risiken

in dilemmatischen Fahrsituationen auf dem Kontinuum möglicher Risikotragweiten in jedem Fall näher am Extrempunkt der katastrophalen als an jenem der trivialen Risiken anzusiedeln; eine exakte Verortung ist im Hinblick auf pragmatische Aspekte nicht zwingend notwendig. Im nächsten Unterkapitel erfolgt nun die risikoethische Evaluation der hier vorgestellten Szenarien unter konsequentialistischen und vertragstheoretischen Gesichtspunkten.

7.2.3 Diskussion aus Sicht konsequentialistischer und kontraktualistischer Kriterien

Nachdem zuvor veranschaulicht wurde, welche Konstellationen in Unfallszenarien denkbar sind, werden diese in einem zweiten Schritt aus risikoethischer Perspektive analysiert. Besonderes Augenmerk wird dabei auf die spezifische ethische Komplexität gelegt, die durch die zugrundeliegende Dilemma-Struktur begründet wird. Während es offensichtlich moralisch falsch ist, einen Fußgänger zu überfahren, sind Handlungen, die mit sehr geringer Wahrscheinlichkeit zur selben Konsequenz führen, nicht grundsätzlich verboten. Die entscheidende Frage ist, wie schwach die kausale Verknüpfung zwischen Ursache und Wirkung sein muss, um eine entsprechende Handlung zu legitimieren. Hansson (2003, S. 292–301) bezeichnet dies als das »mixture appraisal problem«, das ethische Prinzipien bzw. Theorien auf unterschiedliche Weise beantworten. Die Risikoethik stellt im Wesentlichen drei Begründungsansätze bereit, die jeweils spezifische Kriterien hinsichtlich der moralischen Zulässigkeit von Risikoübertragungen beinhalten: Der erste Ansatz umfasst quantitative Optimierungen bzw. Kosten-Nutzen-Abwägungen auf der Basis konsequentialistischer Argumente, während der zweite kontraktualistische Elemente wie die Zustimmung Betroffener in den Fokus rückt. Der dritte Ansatz schließlich argumentiert deontologisch vor dem Hintergrund individueller Rechte und allgemein anerkannter Gerechtigkeitsprinzipien. Welcher Begründungsversuch im konkreten Fall anwendbar ist, hängt von den jeweiligen situativen Umständen ab. Die nachfolgende risikoethische Analyse erfolgt daher entlang etablierter Kriterien der Risikoakzeptabilität, anhand derer sowohl die Frage nach der Zulässigkeit von Risikoübertragungen als auch distributive Aspekte erörtert werden.

7.2.3.1 Konsequentialistische Kriterien: Grenzen quantitativer Optimierung

Konsequentialistische risikoethische Kriterien bilden den Kern rationaler Risikopraxis. Sie beantworten die Frage nach der Zulässigkeit von Risikoübertragungen mithilfe der Quantifizierung von Eintrittswahrscheinlichkeiten und Schadenshöhen in Form eines numerischen Erwartungswerts, der über alle möglichen Folgen einer Handlungsoption aggregiert (vgl. Nida-Rümelin et al., 2012, S. 36–37). Bezogen auf Unfalldilemmata des autonomen Fahrens würde dies implizieren, dass diejenige Handlungsalternative mit dem geringsten zu erwartenden Gesamtschaden gewählt werden sollte. Allerdings ist die Fokussierung auf eine rein quantitative Bestimmung von Risikowerten als Basis ethischer Bewertungen fragwürdig. An dieser Stelle sei auf die entsprechenden Ausführungen in Kap. 6.3.3 verwiesen, deren Ergebnisse hier nur in zusammengefasster Form wiedergegeben werden. Hauptkritikpunkte an konsequentialistischen Ansätzen sind, dass zum einen durch die Aggregation im Sinne der Optimierung des gesamten Risikoerwartungswerts individuelle Rechte und Risikopräferenzen systematisch missachtet werden. Zum anderen bleiben dabei auch Verteilungsaspekte und Gerechtigkeitsabwägungen unberücksichtigt; ethisch problematisch ist dabei insbesondere, dass unter dem konsequentialistischen Ansatz der Nachteil des einen stets durch einen genügend großen Vorteil eines anderen gerechtfertigt werden kann (vgl. Hansson, 2003, S. 295).

7.2.3.2 Kontraktualistische Kriterien: Formen der Zustimmung und ihre Relevanz für Unfallalgorithmen

Kontraktualistische Ansätze der Risikoethik legitimieren die normative Gültigkeit von Prinzipien und entsprechenden risikobehafteten Handlungen unter Bezugnahme auf das zentrale Kriterium der Zustimmung. In der Risikoethik herrscht weitgehend Einigkeit darüber, dass triviale Risiken zustimmungslos übertragen werden dürfen, da

andernfalls das soziale Leben zu stark erschwert würde.[257] So bewertet Thomson (1985a, S. 134–137) einen grundsätzlichen Ausschluss von Handlungen mit inakzeptablen Konsequenzen, wie ihn beispielsweise das *Maximin*-Prinzip impliziert (vgl. Rath, 2011, Fußn. 37), ungeachtet ihrer Wahrscheinlichkeit als kontraintuitiv. Auch Leonard und Zeckhauser (1986, S. 45) lehnen es ab, triviale Risiken zu regulieren, weil es ineffizient und impraktikabel ist, von jedem potenziell Betroffenen eine Zustimmung einzuholen.

Die Übertragung von Risiken, deren Ausmaß nicht als trivial gelten kann, ist dagegen kontrovers. Dies hängt teilweise damit zusammen, dass, obwohl Zustimmung in allen kontraktualistischen Entwürfen als bedeutsames Konzept erachtet wird, unterschiedliche Ansichten darüber existieren, warum dies so ist. Scheffler (1985, S. 76–83) sieht z. B. rechtebasierte Begründungen als problematisch an, in deren Rahmen Zustimmung lediglich als Mittel interpretiert wird, um praktisch notwendige Verletzungen individueller Rechte mit ihrer Geltung zu versöhnen.[258] Es wäre stattdessen notwendig einzuschränken, wann etwas als Rechtsverletzung gilt, und in diesem Zuge z. B. triviale Risiken zu erlauben, denn über Zustimmung allein lassen sich nicht alle denkbaren Fälle klären: »[…] consent cannot serve as the sole vehicle for the justification of social institutions and patterns of social coordination, for its own value depends in part on those very institutions and patterns.« (Ebd., S. 85) In diesem Sinne spricht er sich gegen ein absolutes Verbot von Risikoübertragungen aus.

Auch weitere Lesarten hinsichtlich der Rolle von Zustimmung in Risikofragen gemäß Scheffler legen nahe, dass es unzulässig ist, Personen ohne deren Zustimmung Risiken auszusetzen. Während die eine die instrumentelle Bedeutung von Zustimmung (»its effectiveness as an instrument for promoting good states of affairs and avoiding bad ones« (ebd., S. 77)) akzentuiert, begreift die andere Zustimmung als intrinsischen Wert bzw. Komponente eines selbstbestimmten, guten Lebens. Alle drei Interpretationsweisen begründen jedoch kein absolutes Verbot von Risikoübertragungen. Ein

257 Siehe die Diskussion des *Problem of Paralysis* in Kap. 7.1.1.

258 In diesem Zusammenhang sind Rechtsbegründungen konsequentialistisch, es fehlt eine plausible nicht-konsequentialistische Begründung (vgl. Scheffler, 1985, S. 82).

solches wäre kontraintuitiv, wenn durch Risikoübertragung ein ›besserer Zustand‹ im Sinne der jeweiligen Ansätze herbeigeführt werden könnte. Letztlich implizieren sie eine Abwägung der Vor- und Nachteile von Handlungsalternativen, die weder zwangsläufig ökonomisch noch aggregativ sein muss, sondern durchaus distributive Effekte berücksichtigen kann (vgl. ebd., S. 78–79).[259]

In der Risikoethik werden drei Arten der Zustimmung diskutiert, die sich hinsichtlich wesentlicher qualitativer Kriterien unterscheiden: »Es können […] drei Bedingungen unterschieden werden, die einer faktischen Zustimmung Gültigkeit verleihen: (1) Abwesenheit von Zwang, (2) Informiertheit und (3) Kompetenz« (Rehmann-Sutter, 1998, S. 51). Für die zustimmungsbasierte Legitimation von Risiken kann grundsätzlich postuliert werden: Je höher der Partizipationsgrad betroffener Individuen an einer Entscheidung, desto höher die Legitimität der entsprechenden Risiken (vgl. Gibson, 1985, S. 153; Rath, 2011, S. 33–34). Als robusteste Form gilt die *explizite Zustimmung,* die fordert, dass dem betreffenden Individuum alle verfügbaren Informationen vorliegen und es frei von äußeren Zwängen entscheidet (vgl. Thomson, 1985a, S. 137). Das zustimmende Individuum wird in diesem Zuge selbst zum Risikourheber, die Risikosituation damit von einer sozialen zu einer individuellen (vgl. Rippe, 2013, S. 524). Im Hinblick auf die Klärung konkreter Risikoübertragungen ist diese Form jedoch in den wenigsten risikoethisch relevanten Fällen gegeben, denn in der Regel ist weder die erforderliche Informationsbasis vorhanden noch sind die praktischen Rahmenbedingungen entsprechend effizient:

259 Es gibt noch einen anderen Spezialfall, unter dem Zustimmung kein notwendiges Kriterium darstellt. McCarthy (1997, S. 210–217) diskutiert in diesem Kontext einen Vorschlag von Thomson, demzufolge die Zulässigkeit einer Rechtsverletzung durch eine Risikoübertragung eine Frage des Abwägens ist. Dabei ist sowohl relevant, was sich daraus für den Verursacher ergibt, als auch das Maß, in dem der Geschädigte dadurch schlechter gestellt wird: »At least roughly, a rights infringement is permissible if the good that would come of the infringement sufficiently outweighs the burden of the infringement to the bearer of the right.« (Ebd., S. 210) Dies impliziert auch, dass eine Rechtsverletzung dann unzulässig bleibt, wenn sie dem Betroffenen unverhältnismäßig schadet. Auf einen Zustimmungsvorbehalt kann demnach verzichtet werden, wenn eine Rechtsverletzung im Sinne eines Abwägens der Vorteile des Verursachers und der Nachteile des Betroffenen zulässig wäre.

> Actual consent is not a realistic criterion in a complex society in which everyone performs actions with marginal but additive effects on many people's lives. According to the criterion of actual consent, you have a veto against me or anyone else who wants to drive a car in the town where you live. Similarly, I have a veto against your use of coal to heat your house, since the emissions contribute to health risks that affect me. In this way each of us can block activities by many other persons. (Hansson, 2003, S. 300)

Schwächere Formen der Zustimmung sind im Hinblick auf praktische risikoethische Probleme relevanter, jedoch hinsichtlich ihrer Legitimität diskussionswürdig. Bei der *indirekten Zustimmung* resultiert eine risikobehaftete Handlung aus einem Entscheidungsverfahren, dem explizit zugestimmt wurde; das Individuum ist hierbei von konkreten Einzelfallentscheidungen ausgeschlossen (vgl. Rath, 2011, S. 34–35). Gemäß Thomson (1985a, S. 137–139) muss in diesem Zusammenhang genau spezifiziert werden, welchen konkreten Inhalt die Zustimmung umfasst. So hat man durch die Teilnahme an einer Lotterie noch nicht einem potenziellen finanziellen Verlust zugestimmt; ebenso wenig hat man durch die Entscheidung, durch ein verrufenes Stadtviertel zu spazieren, seine Zustimmung zu einem möglichen Überfall gegeben. Prinzipiell kann eine indirekte Zustimmung auch *implizit* erfolgen. In diesem Fall lässt sich die Zustimmung eines Individuums aus »ihrer institutionellen Eingebundenheit oder ihrem beobachteten Verhalten in anderen Kontexten« (Nida-Rümelin et al., 2012, S. 196) ableiten. Es liegt nahe anzunehmen, dass diese Form der Zustimmung in vielen Fällen zu zweifelhaften Ergebnissen führt.

Die bezogen auf ihre Legitimationskraft schwächste Form stellt die *hypothetische Zustimmung* dar, die davon ausgeht, dass einem Risiko explizit zugestimmt würde, sofern alle notwendigen Informationen vorlägen (vgl. Rath, 2011, S. 35). Da hier der Risikourheber quasi stellvertretend für den Risikobetroffenen entscheidet, ist diese Form der Zustimmung in ihrer Begründung prinzipiell fragwürdig, weil sie die individuelle Autonomie und Verantwortungsfähigkeit untergräbt.[260]

260 Um dieses Problem zu vermeiden, schlägt Gibson (1985, S. 151–152) vor, die Zulässigkeit hypothetischer Zustimmung auf Fälle zu beschränken, in denen es den Betroffenen, z. B. aufgrund ihres gesundheitlichen oder kognitiven Zu-

Welche Relevanz entfalten konkrete Formen der Zustimmung nun für den in dieser Forschungsarbeit untersuchten Anwendungsfall? Wie in den meisten real-lebensweltlichen Fragen der Gestaltung sozialen Miteinanders ist auch im Hinblick auf Unfallalgorithmen eine Annahme expliziter Zustimmung zu Risikoübertragungen nicht überzeugend. Aufgrund ihres restriktiven Charakters würden Mobilitätssysteme gänzlich zum Erliegen kommen, wenn für jedes denkbare Risikoszenario eine explizite Zustimmung eingeholt werden müsste:

> Würde man versuchen, das explizite Zustimmungskriterium als allgemeines Kriterium für die Risikoexposition einzuführen, würde moralisch richtiges Handeln in einer Vielzahl von Lebenssituationen nahezu unmöglich. Selbst wenn man zu Fuß durch die Stadt geht, setzt dies andere Personen Risiken aus, wie zum Beispiel andere Fußgänger oder Fahrradfahrer, denen man in den Weg laufen könnte. Jeden um Erlaubnis zu fragen, verunmöglichte die Durchführung der Handlung. (Rippe, 2013, S. 528)

Eine auf indirekter Zustimmung basierende Entscheidungsstrategie für das Anwendungsproblem erscheint auf den ersten Blick naheliegender, erweist sich bei näherer Betrachtung aber ebenfalls als diskussionswürdig. Inwiefern kann die Beteiligung am Straßenverkehr bereits als Zustimmung zu sämtlichen kontextrelevanten Risikoübertragungen gewertet werden? Rippe (2013, S. 524) schreibt dazu: »Implizite Zustimmung liegt vor, wenn eine urteilsfähige Person, obwohl sie Informationen über ein Risiko hat, der Situation der Risikoexposition nicht ausweicht oder sich bewusst in sie hineinbegibt.« Gemäß dieser Definition nähmen die Insassen eines autonomen Fahrzeugs also durch ihre Wahl des Beförderungsmittels bewusst Risiken

---

stands, tatsächlich nicht (mehr) möglich ist, selbst zu entscheiden. Auch diese Fälle sind allerdings nicht von einer kritischen Beurteilung ausgenommen. Hansson (2003, S. 300–301) bemängelt, dass Vertreter der hypothetischen Vertragstheorie bisher den Nachweis schuldig sind, wie ein hypothetischer Vertrag – also einer, der nicht tatsächlich eingegangen wird – bindend sein kann. Speziell für Fälle mit involvierten Risiken bzw. Unsicherheiten gilt zudem, dass keiner der existierenden Ansätze zeigen konnte, wieso überhaupt ein solcher Zustand angenommen werden sollte, denn auch in einem solchen kann keine für alle Beteiligten zustimmungsfähige Strategie erreicht werden: »In particular, the debate following Rawls's Theory of Justice has shown that there is no single decision-rule for risk and uncertainty that all participants in Rawls's hypothetical initial situation can be supposed to adhere to.« (Ebd., S. 301).

in eigener Verantwortung in Kauf. Doch um *welche* Risiken handelt es sich dabei? Gemäß Thomson (1985a, S. 137–139) ist es entscheidend zu spezifizieren, *wozu* inhaltlich konkret zugestimmt wurde. Es ist davon auszugehen, dass Verkehrsteilnehmer Risikoübertragungen unter der Voraussetzung zustimmen, dass sich »alle Handelnde[n] an gewisse Regeln halten, denen man wiederum zustimmt oder zumindest vernünftigerweise zustimmen kann.« (Rippe, 2013, S. 529) Risikoübertragungen, die sich aus nicht-regelkonformem Verhalten ergeben, sind demnach nicht Gegenstand der Zustimmung und somit auch nicht legitimiert; so etwa in Beispielszenario 4 ›Motorradfahrer mit/ohne Helm‹, wo für die Risikoübertragung, die daraus resultiert, dass der andere keinen Helm trägt, keine Zustimmung angenommen werden kann.

Im Kontext automatisierter Mobilität umfasst die Forderung regelkonformen Verhaltens sowohl menschliche Akteure als auch autonome Fahrzeuge. Die Zusicherung zustimmungsfähigen Verhaltens kann nur über ein ethisch rechtfertigbares Design von Steuerungsalgorithmen erfolgen. Eine entscheidende Rolle kommt hierbei den Verflechtungen zwischen ethischer Risikoakzeptabilität und soziologischer Risikoakzeptanz zu (siehe Kap. 7.1.1). Mit der Beteiligung am Straßenverkehr nehmen Individuen gewisse gesellschaftlich akzeptierte Risiken in Kauf. Im Zuge der Automatisierung des Verkehrs verändert sich jedoch die Risikostruktur relevanter Situationszusammenhänge grundlegend. Zugleich ist Mobilität weiterhin ein Grundbedürfnis, auf das nicht einfach verzichtet werden kann, auch wenn neu generierte Risikokonstellationen individuell inakzeptabel sein mögen. Im Grunde hat jemand, der den neuartigen Risiken eigentlich nicht zustimmt, keine Alternative; er entscheidet nicht frei von äußeren Zwängen, was die Plausibilität einer in diesem Kontext implizit erfolgten Zustimmung in Frage stellt. Ferner ist die Annahme, Nutzer stimmten durch ihre Nutzung eines selbstfahrenden Fahrzeugs dem implementierten ethischen Entscheidungsalgorithmus vollumfänglich zu, nur so lange plausibel, wie die Akteure auch die Möglichkeit haben, sich potenziell anders zu entscheiden. Sollten autonome Fahrzeuge eines Tages verpflichtend werden, wäre die Voraussetzung der Zwangsfreiheit nicht länger erfüllt und die Legitimierung von Risikoübertragungen müsste neu verhandelt werden.

Dennoch kann auch für den Fall, dass keine Einschränkungen bzw. äußeren Umstände gegeben sind, die als hinreichend zwingend angesehen werden können, die Beteiligung am Verkehr noch nicht automatisch als willentlicher Akt der Zustimmung zur automatisierten Mobilität als Ganze verstanden werden. Eine indirekte Zustimmung setzt weiterhin voraus, dass die Individuen den Inhalt bzw. Gegenstandsbereich ihrer Zustimmung kennen. Hierzu ist es notwendig, dass sie über ausreichende Informationen über mögliche Auswirkungen ihrer eigenen Handlungen und der anderer Akteure verfügen. Zu diesem Zweck ist es unumgänglich, dass algorithmische Handlungsvorschriften transparent und nachvollziehbar sind. Ebenso wie Risiken menschlichen Fehlverhaltens prinzipiell erwartbar sind, muss auch die nicht vollständig eliminierbare Fehleranfälligkeit technischer Systeme in gewissem Maße akzeptiert werden, um das *Problem of Paralysis* zu vermeiden. Da jedoch der Raum möglicher Szenarien unendlich und die Informationsbasis naturgemäß unvollständig ist, kann es immer Situationen geben, über die man nicht nachgedacht und denen man folglich auch nicht zugestimmt hat.[261] Die Forderung einer vollumfänglichen und eindeutigen Informationslage erscheint daher vor dem Hintergrund pragmatischer Überlegungen zu stark:

> Even if AVs only have a ›slave morality‹ in which they always follow orders, and citizens implicitly consent to their use (through, say, political means), that still leaves unanswered whether the risk (of malfunction, unintended consequences, or other error) to *unintended* parties is morally permissible. After all, even if widespread consent is in some sense possible, it is completely unrealistic to believe that all humans affected by AI in AVs could give *informed* consent to their use. So does the morality of consent require adequate knowledge of what is being consented to? (Abney, 2022, S. 267, Hervorh. i. Orig.)

Ebenso wenig ist es jedoch plausibel, aus der willentlichen Teilnahme am Straßenverkehr zwingend zu folgern, dass die Individuen jeder prinzipiell denkbaren Risikoübertragung in diesem Kontext zustimmen. Gänzlich klären lässt sich die Problematik der inhaltlichen Unterbestimmtheit angenommener Zustimmungshandlungen letztlich nicht.

261 Dies lässt sich prinzipiell auf real-lebensweltliche Zusammenhänge verallgemeinern, denn die Zukunft ist immer unsicher.

Auch für das Konstrukt der hypothetischen Zustimmung wird eine allgemeine Zustimmungsfähigkeit der relevanten Risikoübertragungen vorausgesetzt. Die Programmierung von Unfallalgorithmen ist Teil der Regulierungsaufgabe automatisierter Mobilität. Einzelne Risikobetroffene werden zwar nicht explizit gefragt, jedoch genießt die Wahrung ihrer Autonomie hohe Priorität. Die Herausforderung besteht einerseits in der inhaltlichen Ausgestaltung eines Algorithmus, dem alle zustimmen würden, und andererseits in der ethischen Legitimierung stellvertretender Zustimmung, die unabhängig von inhaltlichen Aspekten zu begründen ist. Bisher fehlen überzeugende Ansätze, um eine hypothetische Zustimmung für praktische Probleme ernsthaft in Betracht zu ziehen.

In Anbetracht dessen lassen sich Dilemma-Szenarien nur begrenzt mithilfe der Konstrukte indirekter bzw. impliziter Zustimmung erklären. Vielmehr kann die Schaffung hinreichender Voraussetzungen für eine indirekte Zustimmung als erklärtes Ziel der Programmierung von Unfallalgorithmen verstanden werden; deren ethische Problematik lässt sich nur klären, wenn auch im Einzelfall (normativ) zustimmungsfähige Handlungsvorschriften entwickelt werden.[262]

Ein weiterer Typ kontraktualistischer Ansätze greift auf das Konstrukt der Risikokompensation als Zustimmungsalternative zurück. Um Fälle zu identifizieren, in denen Zustimmung nötig ist, wird zwischen kompensierbaren und nicht-kompensierbaren Risiken differenziert. Liegen Erstere vor, so kann eine Zustimmung entfallen;

262 Bis dato existieren in Bezug auf die ethische Problematik von Unfallalgorithmen keine systematisch ausgearbeiteten risikoethischen Entwürfe, die sich auf vertragstheoretische Argumente stützen. Der Ansatz von Derek Leben ist zwar prominent, aber nicht explizit risikoethisch. Dennoch lassen sich einige seiner Gedankengänge und Implikationen in risikoethische Argumente übertragen. Leben entwirft einen von Rawls inspirierten Algorithmus für autonome Fahrzeuge, dessen zentrale Idee auf der maschinellen Schätzung der Überlebenswahrscheinlichkeit jeder betroffenen Person bei jeder möglichen Handlung der Maschine basiert (vgl. Leben, 2017, S. 110–114). Davon ausgehend wird mittels eines iterierten *Maximin*-Prinzips berechnet, welche Option jede einzelne Person wählen würde, wenn sie sich in einer »original bargaining position of fairness« (ebd., S. 108) befände, bis hin zu einem Pareto-optimalen Zustand. Obwohl Lebens Argumentation auf Rawls' vertragstheoretischer Interpretation des *Maximin*-Prinzips beruht, betont sie damit letztlich auch die Bedeutung von Verteilungseffekten.

die Kompensation ersetzt sozusagen eine formale Zustimmung und legitimiert auf diese Weise die Risikoübertragung. Als kompensierbar gelten Risiken immer dann, wenn es sich um materielle Schäden handelt oder um solche immateriellen, die materiale Auswirkungen, z. B. in Form finanzieller Verluste, haben (vgl. Rath, 2011, S. 38; Thomson, 1986b, S. 157). Kompensation kann auf zwei Arten erfolgen: Entweder wird die betroffene Person ex ante auf Basis des errechneten Erwartungswerts für das Tragen des Risikos entschädigt oder die Entschädigung erfolgt ex post und entspricht dem tatsächlich manifestierten Schadenswert. Während eine Risikoübertragung mit Ex-Ante-Kompensation im Kern kontraktualistische Züge trägt, da sie auch abgelehnt werden kann, weist eine Kompensation ex post den Charakter einer Verbindlichkeit auf, die im Nachhinein beglichen wird (vgl. Rath, 2011, S. 39–40), und hat daher praktische Grenzen. So merken Leonhard und Zeckhauser (1986, S. 32–36) hier kritisch an, dass Fragen der Risikourheberschaft nicht immer eindeutig zu klären sind, woraus sich Konflikte bezüglich der Zuschreibung von Haftbarkeit und Verantwortung ergeben können. Sie plädieren daher für eine formalisierte Ex-Ante-Kompensation in Form eines Vertrags, der eine grundsätzliche Zustimmungsbereitschaft voraussetzt. Dies bedeutet, dass Kompensation nur dann eine Alternative zur Zustimmung sein kann, wenn sie ex ante erfolgt. Aus Sicht von Thomson (1986a, S. 66–71) hingegen kann die Ex-Post-Kompensation auch dann als gleichwertig angesehen werden, wenn sie Handlungen den Weg bereitet, bei denen ex ante keine Verhandlungen möglich sind. In jedem Fall lässt sich festhalten, dass Kompensation eine fehlende Zustimmung nur in sehr begrenztem Maße plausibel ersetzen kann.

Zustimmungsbedürftig sind soziale Risiken insbesondere dann, wenn sie nicht kompensierbar sind (vgl. Rath, 2011, S. 63). Dies trifft primär auf physische Schädigungen bis hin zu tödlichem Ausgang, katastrophale Schadensausmaße oder Beeinträchtigungen immaterieller Werte zu. Für diese Fälle, zu denen auch die Risiken im Kontext von Unfallalgorithmen zählen, müssen alternative Strategien entwickelt werden;[263] in rein kontraktualistischer Fassung sind

263 Der zentrale Beitrag von Raths (2011) Abhandlung besteht in der Entwicklung einer alternativen Entscheidungstheorie, die sich von den etablierten Kriterien (Bayes, *Maximin*, *Prinzip der Vorsicht*) unterscheidet, indem sie besonderen

Konzepte der Kompensation hier nicht zu begründen. Inwiefern sie dennoch im Rahmen deontologischer Ansätze integrierbar sind, wird im Folgenden gezeigt.

7.2.4 Deontologische Risikoethik: Begründung, Ansätze und Konzeptionen

Wie in Kap. 5 erörtert, sind Dilemma-Szenarien im Kontext des autonomen Fahrens durch eine Nicht-Verrechenbarkeit fundamentaler moralischer Gebote charakterisiert. Diese folgt aus der inkompatiblen Verschränkung legitimer Interessen verschiedener Personen, die sich auf inkommensurable Werte zurückführen lassen. Die Ergebnisse der metaethischen Analyse implizieren, dass es durch Risikoübertragungen unvermeidbar zu moralisch unzulässigen Beeinträchtigungen von individuellen Rechten, Autonomie und Verantwortungsfähigkeit kommt. Wie zuvor dargelegt, erscheint ein radikales Verbot risikobehafteter Handlungen im Hinblick auf praktische Aspekte wenig plausibel. Es stellt sich daher die Frage, unter welchen Umständen sich spezifische, im deontologischen Sinne dilemmatische Risikoübertragungen sowohl allgemein als auch im Anwendungskontext ethisch rechtfertigen lassen. In der Grundordnung einer freiheitlichen Demokratie gelten bestimmte individuelle Rechte und Freiheiten kategorisch, d. h. sie unterliegen einem Abwägungsverbot und dürfen keinem Optimierungskalkül unterworfen werden. Diesem Umstand tragen deontologische Ansätze der Risikoethik Rechnung, die auf dem Grundgedanken basieren, dass nicht allein Schadensausmaß und Eintrittswahrscheinlichkeit für die moralische Bewertung eines Risikos relevant sind, sondern auch andere moralische Handlungsgründe Beachtung finden müssen.

Die Notwendigkeit einer deontologischen Risikoethik ergibt sich aus der Unfähigkeit konsequentialistischer Ansätze, bestimmte ethische Problematiken, die sich beispielsweise in Bezug auf Abwägungsverbote stellen, zu erfassen und normativ zu berücksichtigen. Ni-

---

Fokus auf die Entscheidungsinstanz legt. In Abgrenzung zu der Annahme traditioneller Ansätze, dass Risikoentscheidungen von einem neutralen Individuum getroffen werden, skizziert Rath im Rahmen seines sogenannten risikoethischen Kontraktualismus die Grundzüge einer prozeduralen Entscheidungsfindung durch Risikobetroffene.

da-Rümelin et al. (2012, S. 154–155) verdeutlichen dies anhand der »Tragödie der großen Zahl«: Wird eine hinreichend große Anzahl von Personen einem minimalen Risiko ausgesetzt, so führt dies trotz der geringen Eintrittswahrscheinlichkeit im Durchschnitt dennoch zu einer gewissen Anzahl an Toten. Aus konsequentialistischer Sicht wäre diese Risikoübertragung ebenso zu bewerten wie eine alternative Handlung, infolge derer die gleiche Anzahl an Personen gezielt getötet würde:

> Eine konkrete Person zu töten oder ihre Tötung zuzulassen, obwohl sie vermeidbar ist, kann durch ökonomische Vorteile, Annehmlichkeiten des Lebens, den technologischen Fortschritt etc. nicht aufgewogen werden. Im Rahmen deontologischer Risikoethik können unterschiedliche Kriterien der moralischen Beurteilung eine Rolle spielen, die sich nicht auf ein fundamentales Kriterium wie das der Maximierung des Nutzenerwartungswertes reduzieren lassen. Das genannte Abwägungsverbot, das uns moralisch unverzichtbar erscheint, kann nur von deontologischen Risikoethiken integriert werden. (Ebd., S. 155)

Innerhalb der deontologischen Risikoethik existieren verschiedene Abstufungen. Eine strikt deontologische Risikoethik weist jegliche Relevanz von Handlungsfolgen für die moralische Bewertung einer Handlung zurück. Im Hinblick auf lebenspraktische moralische Dilemma-Strukturen sprechen Nida-Rümelin et al. (ebd., S. 159) von einer aporetischen Situation, die durch das absolute Abwägungsverbot im Rahmen einer strikt deontologischen Ethik zweiter Ordnung hervorgerufen wird. Um eine solche Situation vollständig aufzulösen, wäre es nötig, den Geltungsanspruch individueller Rechte zu relativieren, indem dieser an das Vorliegen gewisser Bedingungen geknüpft wird: Individuelle Rechte könnten nur dann geltend gemacht werden, wenn eine Handlungsoption existiert, bei der keine anderen deontologisch begründeten Pflichten verletzt würden. Besteht keine solche Option, werden die entsprechenden Rechtsansprüche ungültig und damit das Dilemma aufgehoben. Ist eine solche Perspektive im Fall des autonomen Fahrens denkbar? Aufgrund der Nicht-Verrechenbarkeit deontologisch begründeter Pflichten gibt es keinen moralischen Grund, der es rechtfertigen würde, das Leben einer bestimmten Person zu bevorzugen. Daher liegt es nahe anzunehmen, dass, wenn Individualrechte nicht-eliminierbare ethische Gründe darstellen, das entsprechende Dilemma bestehen bleibt und

Entscheidende zwangsläufig moralisch scheitern müssen (vgl. ebd., S. 160).[264]

Im Gegensatz dazu schließt eine nicht-strikt bzw. schwach deontologische Risikoethik eine konsequentialistische Optimierung nicht kategorisch aus, sondern lässt diese im Rahmen deontologischer Grenzen zu, die sich aus den normativen Merkmalen der spezifischen Entscheidungssituation ergeben. Übertragen auf den Anwendungskontext des autonomen Fahrens würde das z. B. bedeuten, dass Opferungen von Personen zum Zweck der Schadensoptimierung grundsätzlich unzulässig bleiben, weil dabei individuelle Le-

264 Die rechtebasierte Risikoethik diskutiert die Problematik eines kategorischen Verbots risikobehafteter Handlungen als sogenannte Risiko-These im Sinne der Prämisse eines moralisch begründeten Nicht-Schadens-Rechts. Dieses wird verletzt, wenn eine Person einem Risiko ausgesetzt wird, da dadurch möglicherweise individuelle Interessen tangiert werden, die zur Lebensführung relevant sind. Diese These wird oftmals als zu starke Einschränkung der menschlichen Handlungsfähigkeit angesehen (vgl. Hansson, 2003, S. 297–299). Darüber hinaus wird angezweifelt, ob eine auf Rechten basierende Ethik für risikoethische Fragestellungen grundsätzlich geeignet ist (vgl. Nozick, 1974, S. 73–76; Thomson, 1990, S. 243–246). Befürworter wiederum argumentieren, dass die Risiko-These nur dann ceteris paribus zum Problem der Handlungsunfähigkeit führen würde, wenn sie (nahezu) alle Handlungen verböte (vgl. Steigleder, 2016a, S. 254–261, 2016b, S. 440–441). Ein solches Verbot wäre allerdings nur dann notwendig, wenn Risikoübertragungen in jedem Fall eine unzulässige Rechtsverletzung darstellen würden, was nicht der Fall ist, z. B. wenn ex ante eine Zustimmung vorliegt, in Notwehrsituationen oder bei drohenden Schäden für Dritte (vgl. McCarthy, 1997, S. 217). Risikopraktische Fragen stehen in einem Spannungsfeld gleicher Geltungsansprüche der Rechte auf Nicht-Schaden einerseits und Verwirklichung maximaler Freiheit andererseits, das ein Abwägen von Rechtsverletzungen verschiedener Personen notwendig macht, wobei weder die eine noch die andere Seite unangemessen privilegiert werden darf (vgl. Steigleder, 2016b, S. 441). Die Bewertung der Angemessenheit orientiert sich am Kriterium der Unentbehrlichkeit der Rechte als »Voraussetzungen handelnder Selbstverwirklichung« (ebd., S. 441). Eine gewisse Risikotoleranz, die eine situativ einseitige Bevorzugung von Rechten zulässig macht, ist in diesem Sinne nicht nur pragmatisch, sondern auch normativ notwendig – sowohl auf Akteur- als auch Rezipientenseite. Straßenverkehrsrisiken stellen hierbei ein klassisches Beispiel dar, weil sie sich trotz umfänglicher Vorsichtsmaßnahmen nicht vollständig eliminieren lassen. Diese Risiken können als gering und damit kontrollierbar gelten; ein generelles Verbot würde die Rechte der Betroffenen unangemessen privilegieren und dabei die Freiheit der Akteure zu stark einschränken (vgl. Steigleder, 2016a, S. 262).

bensrechte verletzt werden:[265] »Risiko kann lediglich innerhalb der Grenzen optimiert werden, in denen sichergestellt ist, dass die Optimierung nicht durch die Verletzung von Rechten und anderen deontologischen Kriterien erkauft wird.« (Nida-Rümelin et al., 2012, S. 159) Mit dem Abwägungsverbot, das aus der Nicht-Verrechenbarkeit begründeter Pflichten folgt, gehen deontologische Einschränkungen einher, die bei jeglicher Risikooptimierung berücksichtigt werden müssen.[266] Auf diese Weise verfolgt eine deontologische Risikoethik das Ziel, einerseits Werte wie Freiheit, Autonomie und individuelle Rechte zu schützen, die durch Risikoübertragungen bedroht sind, andererseits aber auch nicht zu restriktiv zu sein, um eine völlige Paralyse sozialen Miteinanders zu vermeiden. Wie kann eine solche Risikoethik konkret aussehen? Im nachfolgenden Unterkapitel wird schrittweise ein entsprechender deontologischer Entwurf für den spezifischen Kontext von Unfallalgorithmen erarbeitet.

7.3 Grundzüge einer deontologischen Risikoethik für Unfallalgorithmen

7.3.1 Kohärente Risikopraxis nach Julian Nida-Rümelin: Grundlinien, Ziele und Anwendung

Der Philosoph, Autor und ehemalige Politiker Julian Nida-Rümelin, der von 2020 bis 2024 Mitglied im Deutschen Ethikrat war, formuliert im Rahmen eines Kapitelbeitrags für das Handbuch *Angewandte Ethik* (1996) die Grundlagen einer deontologischen, in ihrer politischen Dimension kontraktualistisch verfassten Risikoethik. Ge-

265 Dies gilt, entgegen dem gleichlautenden Einwand, auch dann, wenn die Betroffenen noch gar nicht feststehen. Letztlich stehen diese in einer in gewisser Weise deterministischen Welt immer schon fest, wir können sie aufgrund unseres begrenzten epistemischen Vermögens lediglich nicht erkennen. In moralischer Hinsicht ist diese Situation demnach genauso zu bewerten wie die gezielte Schädigung von Personen (vgl. Nida-Rümelin et al., 2012, S. 152–153).

266 Das bedeutet ausdrücklich nicht, dass im Rahmen einer schwach deontologischen Risikoethik keine Abwägungen stattfinden, beispielsweise bei vorliegenden Zielkonflikten zwischen Komfort und Sicherheit; diese lassen sich allerdings kontraktualistisch plausibel begründen (vgl. Nida-Rümelin et al., 2012, S. 155).

meinsam mit Johann Schulenburg und Benjamin Rath legte er 2012 einen Band vor, der die sieben Jahre zuvor skizzierten Eckpfeiler aufgreift und weiterentwickelt. Den Ausgangspunkt dieser risikoethischen Abhandlung bildet die Feststellung, dass die zeitgenössische Risikopraxis Inkohärenzen aufweist, die sich vor allem hinsichtlich des Umgangs mit Hochtechnologien manifestieren. Zurückzuführen sind sie auf das Spannungsverhältnis zwischen konsequentialistischen und deontologischen Intuitionen, das sich in risikopraktischer Perspektive in den normativen Grundausrichtungen widerspiegelt, die Ökonomie einerseits und juridischen Normen des Rechtssystems andererseits zugrunde liegen. Diese Unterschiede sind, trotz ihrer auf den ersten Blick unvereinbar erscheinenden Grundsätze, in ein kohärentes Normensystem überführbar, welches konsequentialistische und deontologische Logik zusammenbringt (vgl. Nida-Rümelin et al., 2012, S. 161–175). Den Weg ebnet hierbei das Nutzentheorem, dessen Postulate nicht zwingend konsequentialistisch, sondern vielmehr inhaltlich neutral sind, und in das daher auch nicht-konsequentialistische Handlungsgründe wie gerechtigkeitsethische Überlegungen einfließen können.[267] Es »wird nicht das individuelle Wohlergehen zum Maßstab der Rationalität gemacht, sondern die Kohärenz der individuellen Präferenzen.« (Ebd., S. 165–166) Deontologische Normen sind also als Handlungsgründe zu verstehen, die sich in Form handlungsleitender Präferenzen formalisieren lassen, um im Rahmen einer quantitativen Bewertungsfunktion einbezogen zu werden.

Ziel dieser deontologischen Konzeption ist eine kohärente Risikopraxis, die konsequentialistische und deontologische Elemente entsprechend ihrer jeweiligen Rechtfertigungsgrundlage integriert. Konkret bedeutet das: Eine (konsequentialistische) Risikooptimierung wird nicht kategorisch zurückgewiesen, sondern durch deontologische Grenzen eingeschränkt. Diese Konzeption scheint in gewissem Sinne an die in Kap. 4.4.4.5 formulierte Forderung eines pluralistischen Ansatzes anzuknüpfen, wie sie unter dominanten Forschungszugängen häufig vertreten wird. Eine kohärente Risikopraxis berücksichtigt zwar verschiedene ethische Handlungsgründe, recht-

267 Die vier Kohärenzpostulate lauten: Transitivität, Vollständigkeit, Monotonie und Stetigkeit der Präferenzen über Wahrscheinlichkeitsverteilungen (vgl. Nida-Rümelin et al., 2012, S. 173).

fertigt diese aber unabhängig voneinander. Die Idee, Unfallszenarien als primär utilitaristisches Optimierungsproblem zu interpretieren, das deontologische Prinzipien in Form von Nebenbedingungen (*soft constraints*) einbezieht, liegt den meisten bisher vorgeschlagenen, implementierungsnahen Ansätzen zugrunde (vgl. Geisslinger et al., 2021, 2023a; Gerdes & Thornton, 2015; Németh, 2022; Thornton et al., 2017). Im Gegensatz zu diesen und ähnlichen Ansätzen betont eine deontologische Risikoethik jedoch, dass deontologische Kriterien *harte* Grenzen (*hard constraints*) darstellen sollten, die Vorrang vor quantitativen Optimierungszielen haben: »Optimierung ist nur in den Grenzen zulässig, in denen vorrangige deontologische Kriterien nicht verletzt werden.« (Nida-Rümelin et al., 2012, S. 175)

Der Ansatz von Nida-Rümelin et al. nimmt Bezug auf die kontraktualistische Rekonstruktion einer deontologischen Ethiktheorie, die auf der Kompatibilität des Vertragsarguments von Immanuel Kant[268] (vgl. 1900ff., TP, AA 08) einerseits und den Grundzügen der Wesensbestimmung der Moral nach Thomas Scanlon (1982) andererseits beruht (vgl. Nida-Rümelin et al., 2012, S. 181–187).

268 Das weniger prominente kontraktualistische Argument von Kant lautet in konziser Fassung: »Hier ist nun ein ursprünglicher Contract, auf den allein eine bürgerliche, mithin durchgängig rechtliche Verfassung unter Menschen begründet und ein gemeines Wesen errichtet werden kann. – Allein dieser Vertrag (*contractus originarius* oder *pactum sociale* genannt), als Coalition jedes besondern und Privatwillens in einem Volk zu einem gemeinschaftlichen und öffentlichen Willen (zum Behuf einer bloß rechtlichen Gesetzgebung), ist keineswegs als ein Factum vorauszusetzen nöthig (ja als ein solches gar nicht möglich); gleichsam als ob allererst aus der Geschichte vorher bewiesen werden müßte, daß ein Volk, in dessen Rechte und Verbindlichkeiten wir als Nachkommen eingetreten sind, einmal wirklich einen solchen Actus verrichtet und eine sichere Nachricht oder ein Instrument davon uns mündlich oder schriftlich hinterlassen haben müsse, um sich an eine schon bestehende bürgerliche Verfassung für gebunden zu achten. Sondern es ist eine bloße Idee der Vernunft, die aber ihre unbezweifelte (praktische) Realität hat: nämlich jeden Gesetzgeber zu verbinden, daß er seine Gesetze so gebe, als sie aus dem vereinigten Willen eines ganzen Volks haben entspringen können, und jeden Unterthan, so fern er Bürger sein will, so anzusehen, als ob er zu einem solchen Willen mit zusammen gestimmt habe. Denn das ist der Probirstein der Rechtmäßigkeit eines jeden öffentlichen Gesetzes.« (Kant, 1900ff., TP, AA 08: 02-21, Hervorh. i. Orig.) Kants Werk *Über den Gemeinspruch: Das mag in der Theorie richtig sein, taugt aber nicht für die Praxis* wurde erstmals 1793 veröffentlicht.

Grundlage ist dabei folgender Gedankengang: Das Fundament einer deontologischen Ethik besteht in der kantianisch geprägten Konzeption individueller Autonomie, die mit Ansprüchen auf Freiheit und Gleichheit verbunden ist. Wie Scanlon im Rahmen seines ethischen Kontraktualismus erläutert, lässt sich eine auf diese Weise verstandene deontologische Ethik mit einem kontraktualistischen Verständnis dessen, was das Wesen der Moral ist, zusammenbringen. Auf dieser Basis stellt sich die Konzeption von Nida-Rümelin et al. letztlich als kontraktualistisch verfasste Risikopraxis auf der Grundlage deontologischer Begründungen dar, die ihre inhaltliche Ausgestaltung aus dem Wesen individueller Autonomie speist. Die zentralen Grenzkriterien einer solchen Risikopraxis für demokratisch und freiheitlich verfasste Gesellschaften sind daher durch individuelle Autonomie in Form von Individualrechten[269] einerseits und etablierte distributive

269 Die beiden wichtigsten Typen deontologischer Theorien sind die würdebasierte Moraltheorie, die auf Kant zurückgeht, einerseits und rechtebasierte Moraltheorien andererseits. Sie unterscheiden sich hinsichtlich dessen, was sie als normative Grundkategorie akzeptieren: Während die würdebasierte Moraltheorie moralische Pflichten aus der menschlichen Würde ableitet, legen rechtebasierte Moraltheorien moralische (oder juridische) Anspruchsrechte zugrunde. Im Allgemeinen werden rechtebasierte Ansätze kontrovers wahrgenommen; häufig werden sie als fundamentale Verkürzung des Phänomens des Moralischen betrachtet. Dies erscheint auch für das Themenfeld dieser Arbeit einschlägig. Eine Verletzung der relevanten Rechte ist nicht zuletzt deshalb moralisch unzulässig, weil Personen individuell autonom sind und daher Träger eigenständiger Entscheidungskompetenz, die ein Paternalismusverbot in Kraft setzt (vgl. Nida-Rümelin, 2005, S. 875–876; Nida-Rümelin et al., 2012, S. 57). Eine Argumentation allein auf der Basis von Anspruchsrechten würde den dahinter liegenden intrinsischen und inkommensurablen Werten der individuellen Autonomie und Freiheit nicht gerecht. Eine kritische Haltung gegenüber einem rechtebasierten Ansatz in der Risikoethik nimmt Hansson (2003, S. 289–299) ein. Er argumentiert, dass ein Nullrisiko auch bei absoluten Rechten impraktikabel ist: »Unfortunately, such a strict extension of rights and prohibitions is socially untenable. Your right not to be killed by me is certainly accompanied by a prohibition for me to perform many types of acts that involve a risk of killing you, but it does not involve a prohibition of all such acts. Such a strict interpretation would make human society impossible. I am allowed to drive a car in the town where you live, although this increases the risk of being killed by me.« (Ebd., S. 298) Zudem erscheint es implausibel, die Akzeptabilität von Risiken nur anhand ihrer Wahrscheinlichkeit zu beurteilen und mögliche durch sie hervorgerufene Vorteile außer Acht zu lassen. Aus

Gerechtigkeitsvorstellungen andererseits gegeben (vgl. Nida-Rümelin, 2005, S. 874).

Während der Entwurf von Nida-Rümelin et al. (2012) auf rein konzeptioneller Ebene verbleibt, wird die ihm zugrundeliegende Argumentation in diesem letzten Teil des Buches auf den Anwendungskontext von Unfallalgorithmen übertragen. Dabei wird der Ansatz vor dem Hintergrund der Problemspezifika inhaltlich konkretisiert und weiterentwickelt. Es werden zwei zentrale Prinzipien herausgearbeitet, die den beiden deontologischen Grenzkriterien im Hinblick auf die Realisierung einer kohärenten Risikopraxis für das Anwendungsproblem entsprechen und dessen Grundlage bilden. Diese beiden Grenzkriterien stellen nicht-verhandelbare Rahmenbedingungen dar, innerhalb derer sich die Verteilung von Vor- und Nachteilen aus Risikoübertragungen als ›neues‹ Optimierungsproblem bewegt.

Wie im zweiten Teil des Buches ausgeführt wurde, zeichnen sich Unfalldilemmata durch nicht-eliminierbare Individualrechte aus. Aus pragmatischer Sicht erscheint es jedoch plausibel zu fordern, dass gewisse Risiken in Kauf genommen werden sollten, ohne dass dies aus ethischer Sicht als Rechtsverletzung zu beurteilen wäre. Die Frage ist also, in welchem Rahmen Verletzungen von Individualrechten als zulässig gewertet werden können und wie diese zu begründen sind. Risikoethische Argumentationen sind dabei nicht aggregativ, sondern nehmen stets den Einzelnen in den Blick. Um die Wahrung individueller Autonomie trotz der Aufweichung des absoluten Geltungsanspruchs der entsprechenden Individualrechte zu gewährleisten, kann ein etabliertes risikoethisches Konzept zu Rate gezogen werden, das sowohl individuelle Interessen als auch soziale Erfordernisse in Betracht zieht: Risikoübertragungen sind genau dann zu akzeptieren, wenn sie den Einzelnen *zugemutet* werden dürfen. Neben ihrer Orientierung an den Interessen der Einzelnen sind Risiken zudem auch interpersonell relevant und damit beziehungskonstituierend. Für anwendungspraktische Fragestellungen ergibt sich aus ethischer Warte das Postulat einer fairen Verteilung von resultierenden Vor- und Nachteilen, die gerechtigkeitsethischen Anforderungen genügt. Wie diese beiden Forderungen vor dem Anwendungskontext automatisierter Mobilität konkret begründet und

---

den hier erörterten Gründen wird in diesem Buch auf eine tiefergehende und explizite Darstellung einer rechtebasierten Sichtweise verzichtet.

in Prinzipien umgesetzt werden können, die eine rechtfertigbare Risikopraxis konstituieren, wird im Folgenden diskutiert.

7.3.2 Die (absolute) Frage der Zumutbarkeit: Eine moralische Gratwanderung entlang von Risikoschwellen

7.3.2.1 Individualorientierung als Referenzpunkt

Interaktionen im verkehrlichen Umfeld sind durch wechselseitige Risikoübertragungen charakterisiert, welche die individuelle Autonomie der Einzelnen betreffen. Aus praktischen Gründen sind gewisse Risiken zu akzeptieren, auch wenn dabei individuelle Rechte potenziell verletzt werden. Doch wo genau verläuft die Grenze zwischen dem, was akzeptiert werden muss, und dem, was inakzeptabel ist? Offensichtlich geht es darum, das rechte Maß zwischen dem (utopischen) Ideal eines Nullrisikos und einem eindeutig untragbaren Risiko zu finden. Diese Problemstellung wird in der Risikoethik unter dem Konzept der Zumutbarkeit gefasst, die das Maß bezeichnet, in dem ein Risiko aus ethischer Sicht als akzeptabel gelten kann.[270]

Entscheidend für die ethische Legitimierung zumutbarer Risiken ist, dass sie sich gegen eine aggregative Perspektive wendet. Zumutbarkeit nimmt als risikoethisches Kriterium stets das Individuum, seine Interessen und Ziele in den Blick. Das lässt sich zum einen auf das Bestehen individueller Abwehrrechte zurückführen (vgl. Rippe, 2013, S. 532); zum anderen ist es für eine ethische Beurteilung von hoher Relevanz, ob diejenigen, die von einem Gut oder Zustand profitieren, auch die entsprechenden Risiken tragen. Das Konzept der Zumutbarkeit tritt als Vermittler einer Verhältnismäßigkeit zwischen den individuellen Vor- und Nachteilen auf, die eine Risikoübertragung hervorbringt. Dabei werden zumeist subjektive Sicherheitsbedürfnisse tangiert, die abhängig sind von individuellen Risikopräferenzen. Jedoch sind objektive Risikoeinschätzungen vonnöten, um normative, interpersonell verbindliche Implikationen ab-

270 In der TFA spielt das Konzept der Zumutbarkeit ebenfalls eine zentrale Rolle, indem es eine Brücke zwischen Akzeptanz und Akzeptabilität von Technik schlägt (vgl. Grunwald, 2005).

zuleiten; sich an subjektiven Präferenzen zu orientieren ist weder praktikabel noch ethisch zu rechtfertigen. Vielmehr stellt Rippe (2013) fest, dass das ethische Problem der Zumutbarkeit von Risiken »nur dann zu klaren moralischen Antworten führen [kann], wenn es um sogenannte objektive, nicht um subjektive Risikoeinschätzungen geht.« (Ebd., S. 530)[271] Aus der Individualorientierung bei der Bewertung zumutbarer Risiken folgt also ausdrücklich nicht, dass sich die Zumutbarkeit eines Risikos an subjektiver Wahrnehmung bemisst. Implizit akzeptierte soziale Praktiken sind ebenfalls irrelevant. Zumutbarkeit ist kein Synonym für Sozialakzeptanz; es geht nicht darum, was *tatsächlich* sozial akzeptiert ist, sondern was *rational betrachtet* zu akzeptieren ist. Sie ist kein subjektives Empfinden, sondern nimmt individuelle Rechte bzw. Interessen ernst, indem sie mögliche Einschränkungen derselben thematisiert. Die entscheidenden Aspekte, die subjektive Risikopräferenzen in objektivierte Risikobewertungen transferieren, sind die Reziprozität in der Übertragung zumutbarer Risiken und der daraus resultierende Vorteil, keine absoluten Einschränkungen hinsichtlich der eigenen individuellen Handlungsfreiheit und der Verfolgung persönlicher Ziele hinnehmen zu müssen:

> Wird von Zumutbarkeit gesprochen, ist darin enthalten, dass der Person etwas auferlegt wird, was sie vermeiden will, es aber von ihr zu teilende rationale Gründe gibt, dies dennoch anzunehmen. […] Die zu teilenden rationalen Gründe sind hier letztlich, dass gewisse Risiken hinzunehmen sind, da man ansonsten auch selbst andere keinerlei Risiko aussetzen darf. (Ebd., S. 529)

In Kontexten mit wechselseitigen Risikoübertragungen stehen zumutbare Risiken in einem Spannungsverhältnis zwischen persönlichen Präferenzen und denen anderer. Eine interpersonelle Vergleichbarkeit zumutbarer Risiken ist jedoch nicht ohne Weiteres möglich. Grundsätzlich ist es zwar plausibel zu fordern, dass Personen diejenigen Risiken, die sie anderen zumuten, auch selbst hinnehmen müssen. Der Umkehrschluss gilt allerdings nicht: Ein Risiko, das man bereit ist für sich selbst zu akzeptieren, darf deshalb noch nicht anderen zugemutet werden (vgl. ebd., S. 531). Luhmann

271 Die Verwendung objektiver Risiken bezieht sich nur darauf, wie die Risiken in ihrer Höhe zu bestimmen sind, und nicht auf die Einstellungen, die einzelne Individuen diesen gegenüber haben.

(1997, S. 330) führt als Begründung hierfür mögliche Unterschiede in der persönlichen Risikopräferenz an. Bei eigenen Entscheidungen ist man tendenziell risikobereiter, darf aber anderen gleichzeitig nicht das zumuten, was man für sich selbst akzeptiert. Analog lässt sich im Sinne eines stärker rezipientenorientierten Ansatzes argumentieren, dass ausgehend von der eigenen Risikoeinstellung keine Rückschlüsse dahingehend gezogen werden dürfen, ob eine Risikoübertragung für ein anderes Individuum zumutbar ist (vgl. Birnbacher, 1996, S. 204–205).

Die grundrechtlich garantierte individuelle Autonomie gerät in praktischen Situationen regelmäßig mit risikobehaftetem Handeln in Konflikt. Basierend auf den bisherigen Ausführungen lässt sich ein (absolutes) Leitprinzip formulieren, mit dessen Hilfe unzulässige Rechtsverletzungen abgewendet werden können. Dieses stellt sicher, dass im Rahmen einer Risikooptimierung keine unzumutbaren Risiken für Einzelne entstehen, welche die individuelle Autonomie in unzulässiger Weise beeinträchtigen:

(1) *Rechtsverletzungen, die durch Risikoübertragungen im Kontext von Unfallalgorithmen verursacht werden, können genau dann als moralisch zulässig gelten, wenn die auf die Einzelnen übertragenen Risiken unter Berücksichtigung aller ethisch relevanten Aspekte jeweils individuell zumutbar sind.*

Konkrete inhaltliche Kriterien für die Bewertung der Zumutbarkeit von Risiken im gegebenen Anwendungskontext ergeben sich im Zuge einer Auseinandersetzung mit den spezifischen Umständen relevanter Risikoübertragungen, wie nachfolgend dargelegt wird.

7.3.2.2 Risikoschwellen: Ansätze und Schwierigkeiten

Bereits unsere moralische Intuition legt nahe, dass gewisse Risiken zulässig sein müssen und es zu deren definitorischer Abgrenzung eines Grenzwerts bedarf. Zu den Prinzipien, die im Kontext der ethischen Zumutbarkeit von Risiken am häufigsten rezipiert werden, gehören Schwellenwert-Konzeptionen, deren Ziel es ist, eine Risikoschwelle zu bestimmen, unterhalb derer Risikoübertragungen als unproblematisch eingeschätzt werden können. In diesem Sinne liegt ein zumutbares Risiko dann vor, wenn ein nicht mehr weiter

minimierbares Restrisiko erreicht wurde, das »nur mit Maßnahmen weiter zu verringern wäre, die als unzumutbar gelten, […] obwohl der unwahrscheinliche Schadensfall nach wie vor seinem Umfang nach missbilligt wird.« (Ropohl, 2017, S. 903–904) Eine Handlung kann somit genau dann als zumutbar gelten, wenn der Risikoerwartungswert aller Betroffenen einen bestimmten Grenzwert nicht überschreitet und dabei unabhängig von jeglicher Form der Zustimmung ist. So spricht Hansson (2003, S. 298–299) im Kontext deontologischer Theorien von einem »probability limit«; Rechte und Verbote dürfen verletzt werden, wenn die Eintrittswahrscheinlichkeit eines Schadens gering genug ist. Dies entspricht in etwa dem, was Ryazanov et al. (2023, S. 12) als »a more nuanced understanding of deontology« bezeichnen und das zwischen probabilistischen und sicheren Schäden zu differenzieren vermag: »While deontologists may not be willing to kill one to save five, they may deem it acceptable to risk a 1 % chance of harm to one to save five.« (ebd.)

Ein praktisches Beispiel für einen solchen Schwellenwert stellen Dosisgrenzwerte im Zusammenhang mit Fragen des Strahlenschutzes dar, welche die gerade noch zulässige Dosis festlegen (vgl. Hansson, 2007a, S. 152–154). Auch die von Thomson (1985a, S. 125) diskutierte Subklasse risikoethisch relevanter Handlungen, die sich durch »threshold effects« auszeichnen, ist hier einschlägig. Dabei werden zunächst triviale oder auch gar keine Risiken übertragen, jedoch entsteht infolge der wiederholten Ausführung einer Handlung durch dieselbe oder eine andere Person letztlich ein nicht-triviales Risiko, das nicht ohne Weiteres übertragen werden darf. Andere Ansätze verwenden das maximal akzeptable Risiko als zentralen Parameter zur Trajektorienplanung (vgl. Geisslinger et al., 2023a, 2023b). Sütfeld et al. (2019) plädieren ebenfalls für die Installation eines Risikoschwellenwerts, oberhalb dessen eine Risikoübertragung als unzulässiges Opfer angesehen wird, da ein absoluter Schutz von Unbeteiligten an einem gewissen Punkt nicht mehr vernünftig begründet werden kann:

> The exact point at which it becomes unreasonable not to save the person at risk would need to be debated, but most will agree that at some point the absolute protection of bystanders from any risk will become unreasonable. […] Introducing thresholds that define at which level of risk a decision is considered a sacrifice could recover some notion of

> reasonableness, but only at the cost of robustness and transparency of the decisions. (Ebd., S. 10–11)

Eine der wesentlichen Anforderungen an Schwellenwert-Konzeptionen ist deren Konsistenz. Wurde ein konkretes Risiko als zumutbar eingestuft, muss dies für ein anderes, das bezüglich Schadensausmaß und Wahrscheinlichkeit gleich ist, ebenfalls gelten. Dabei muss jedoch zwischen Risikoübertragung und privatem Risiko unterschieden werden: Was eine Person sich selbst zumutet, darf ihr deshalb noch nicht durch andere zugefügt werden. So setzt sich eine Person selbst einem gewissen Risiko aus, wenn sie sich in einem autonomen Fahrzeug befördern lässt. Daraus folgt jedoch noch keine Legitimation für die Handlungen anderer, die ein Risiko mit demselben Erwartungswert generieren.

Als entscheidende Voraussetzung für die ethische Legimitation spezifischer Schwellenwerte gilt, dass alle ethisch relevanten Aspekte des Kontextes einbezogen werden. Eines der gewichtigsten Probleme von Schwellenwert-Konzeptionen besteht in der numerischen Repräsentation von Risikoerwartungswerten, die mit spezifischen Problemen behaftet und deshalb nur begrenzt für risikoethische Fragen geeignet ist (siehe Kap. 6.3.3). Dies mag in Anwendungskontexten weniger problematisch sein, in denen die direkten Folgen relativ zuverlässig quantifizierbar und entsprechende Verantwortlichkeiten eindeutig sind, beispielsweise im Kontext der Emission radioaktiver Strahlung. Weitaus anspruchsvoller gestaltet es sich jedoch in Fällen, in denen nur rudimentäre Informationen über mögliche technisch induzierte Handlungsfolgen vorliegen.

Als Annäherung an die Problematik ist ein technischer Blickwinkel hilfreich. Zumutbarkeit nach technischem Verständnis beruht auf den Festschreibungen von Sicherheits- und Risikokonzepten gemäß den Vorgaben einschlägiger ISO-Normen. Die Vision des autonomen Fahrens ist eng mit dem Versprechen verbunden, die Zahl schwerer Unfälle signifikant zu reduzieren und unsere Straßen auf diese Weise sicherer zu machen. Doch wie sicher ist ›sicher genug‹? Papadimitriou et al. (2022, S. 9–11) sprechen in diesem Zusammenhang von einem Siloeffekt, da das Konzept der Sicherheit bisher aus unterschiedlichen wissenschaftlichen Perspektiven unabhängig voneinander untersucht wurde, die von ethischen bis zu rein technischen Aspekten reichen. In Forschungs- und Entwicklungsprojekten zu autonomen Fahrsystemen wird Sicherheit daher kaum als eigen-

ständiger Designwert anerkannt. Gemäß Birnbacher (1996, S. 197) ist Sicherheit »eher eine Zuschreibung als eine Beschreibung. Was sicher in diesem Sinne ist, gilt als sicher – und deswegen als akzeptabel und zumutbar.« Von Sicherheit kann im Hinblick auf technische Systeme gesprochen werden, wenn das Risiko unter dem Grenzrisiko liegt; dieses ist quasi das technische Äquivalent der ethischen Zumutbarkeitsschwelle. Gemäß ISO 26262 besteht ein zumutbares Risiko dann, wenn es unterhalb einer Schwelle liegt, die als »nicht akzeptabler Wert in einem spezifischen Kontext gemäß gesellschaftlicher, moralischer und ethischer Auffassungen« zu interpretieren ist:

> Beim Betrieb eines automatisierten Fahrzeugs hängt das zumutbare Risiko von der aktuellen Situation, in der sich das Fahrzeug befindet, ab. Zur Situation gehören hier […] alle für eine Fahrentscheidung relevanten stationären und dynamischen Objekte, die Intention der dynamischen Objekte einschließlich des autonomen Fahrzeugs, die geltenden rechtlichen Bedingungen, die Mission des autonomen Fahrzeugs, die aktuelle Leistungsfähigkeit des autonomen Fahrzeugs. (Reschka, 2015, S. 491)

Wie ist dieser Wert nun aber zu bestimmen? Geisslinger et al. (2023a, S. 140–141) empfehlen, als Initialwert vom derzeitigen Sicherheitsniveau des herkömmlichen Verkehrsgeschehens auszugehen, wobei das anvisierte Pendant für die automatisierte Mobilität zukünftig deutlich höher anzusetzen sein wird.

Schwellenwert-Konzeptionen erfordern, dass kontinuierlich in jeder Situation das Risiko für jeden beteiligten Verkehrsteilnehmer ermittelt wird und diejenigen Optionen identifiziert werden, bei denen das aktuelle und zukünftige Risiko unterhalb des Schwellenwertes bleibt. Erschwerend kommt hinzu, dass Risiken nicht nur von den Aktionen des eigenen Fahrzeugs, sondern auch vom Verhalten anderer abhängen. Dies ist technisch herausfordernd und bisher ungelöst. Potenziale bietet hier das Prinzip der funktionalen Degradation, um bei Eintreten sicherheitsrelevanter Ereignisse einen Zustand zumutbaren Risikos wiederherzustellen (vgl. Reschka, 2015, S. 505–506).

Risikoethische Ansätze greifen diese Aspekte auf und interpretieren sie unter Bezugnahme auf ethisch relevante Gesichtspunkte. Rippe (2013, S. 532) beschreibt zwei grundsätzliche Vorgehensweisen zur Bestimmung zumutbarer Risiken. Die eine stützt sich ausschließlich auf die Eintrittswahrscheinlichkeit als relevantes Kriterium und

definiert damit Ereignisse, die mit hoher Wahrscheinlichkeit eintreten, als unzulässig. Hierbei würden allerdings Risiken durch das Raster fallen, die zwar unwahrscheinlich sind, aber hohe Schäden verursachen. Um derartige Katastrophenrisiken auszuschließen, können zusätzlich die erwarteten Schadenshöhen veranschlagt werden; bei höheren Schäden wären nur sehr geringe Wahrscheinlichkeiten zugelassen. Dies entspricht z. B. dem Konzept des moralischen Erwartungswerts (*expected moral value*), den Bhargava und Kim (2017, S. 9) vorschlagen: »One cannot simply compare 0.2 and 0.8. One must consider the value of the outcomes. Of course, car designs cannot be perfect, but a 20 % probability of a life-threatening malfunction is obviously too high.« Um zwischen schweren und verhältnismäßig leichten Schäden zu differenzieren, wären hierbei je nach Schadensart unterschiedliche Schwellenwerte anzusetzen. Da eine kardinale Anordnung von Interessen hingegen kaum realisierbar erscheint, schlägt Rippe (2013) vor, Abwehrrechte gegen spezifische Risiken als Referenzpunkt zu nehmen, die sich hierarchisch anordnen lassen. Diese würden jeweils missachtet, wenn ein Risiko die für ein konkretes Recht als kritisch befundene Eintrittswahrscheinlichkeit übersteigt. Zumutbare Risiken lassen sich dann entlang von ordinalen Risikoschwellen bestimmen: »Die Zulässigkeit des Handelns lässt sich damit stets mit Blick auf jene Betroffenen prüfen, welche mit höchster Wahrscheinlichkeit in ihrem Recht gefährdet werden.« (Ebd., S. 534)

Dass individuelle Risikopräferenzen entscheidend für die Frage nach der Zumutbarkeit von Risiken sind, ist neben Eintrittswahrscheinlichkeit und Schadenshöhe ein zentrales Element des rezipientenorientierten Ansatzes, den Birnbacher (1996) vertritt. Diesem zufolge liegt der entscheidende Unterschied hinsichtlich der ethischen Frage, unter welchen Umständen wir anderen Personen Risiken anstelle sicherer Schäden zumuten dürfen, gerade darin, dass bei Ersterem die individuellen Risikoeinstellungen der potenziell Betroffenen berücksichtigt werden. Birnbachers Position stehen Ansätze gegenüber, »die von einem situations- und einstellungsinvarianten Zumutbarkeitsmaß ausgehen und entweder eine durchgängig risikoscheue oder durchgängig risikoneutrale Risikostrategie postulieren.« (Ebd., S. 205)

Daraus folgt, dass die Zulässigkeit einer Handlung stets von demjenigen Individuum abhängt, welches durch die Risikoübertragung

am schlechtesten gestellt wird. Dies gilt unabhängig davon, ob für die Bewertung der Risikoübertragung Risikopräferenzen, Eintrittswahrscheinlichkeit und Schadenshöhe oder nur die beiden Letzteren als relevant erachtet werden. In kritischen Situationen müsste für jede beteiligte Person das individuell zumutbare Risiko entsprechend dem jeweils involvierten Abwehrrecht bestimmt werden. Bei Unfalldilemmata hilft eine Hierarchisierung von Abwehrrechten hingegen nicht weiter. Es stehen sich im Hinblick auf den jeweils zu schützenden Wert inkommensurable, individuelle Abwehrrechte gegenüber. Je nach gewählter Trajektorie unterscheiden sich die übertragenen Risiken aber hinsichtlich ihrer Eintrittswahrscheinlichkeiten. So ist in Beispielszenario 2 ›Einzelperson versus Gruppe‹ das Risiko der einzelnen Person viel größer als das jedes Einzelnen in der Gruppe, wenn das Fahrzeug nach links ausweicht. Die im Verkehrskontext produzierten Risiken für Individuen sind sehr unterschiedlich und komplex, was u. a. auf bestehende Unterschiede in der individuellen Vulnerabilität[272] zurückzuführen ist. Da diese, neben der persönlichen physischen und psychischen Konstitution, vom gewählten Verkehrsmittel abhängt, lassen sich entsprechende Risikoübertragungen nicht allein auf Basis der Reziprozitätsprämisse legitimieren; diese ist nur dann anwendbar, wenn alle Beteiligten prinzipiell ähnliche Risiken generieren. Es kann jedoch nicht davon ausgegangen werden, dass alle Verkehrsmittel von allen Personen genutzt werden.

7.3.2.3 Kernkriterium moralische Sorgfaltspflichten

In bestimmten Fällen, die im Kontext autonomer Fahrsysteme nicht selten sind, bedarf es einer von der Reziprozitätsthese unabhängigen Begründung.[273] Besonders deutlich wird dies, wenn sich Personen bewusst manipulativ oder wenigstens unvorsichtig verhalten, indem sie Verkehrsregeln missachten. Reziprozität kann als quasi-kontraktualistisches Element nur funktionieren, wenn sich alle an gewisse Regeln halten. Unsere moralische Intuition sagt uns, dass bei bewusstem Fehlverhalten wie im Fall des Fußgängers, der in Bei-

272 Der Aspekt der Vulnerabilität wird in Kap. 7.3.3 genauer spezifiziert.

273 Versuche, Zumutbarkeit über (kontraktualistische) Verfahrenslösungen zu begründen, scheitern, wie die Diskussion möglicher Zustimmungsformen in Kap. 7.2.3.2 zeigt.

spielszenario 3 ›Rote Ampel‹ die Straße verkehrswidrig überquert, die Schwelle des zumutbaren Risikos für die entsprechende Person deutlich höher anzusetzen ist. Bewusste Regelmissachtungen schaffen Risiken, die anderen nicht zugemutet werden dürfen; derartige Verhaltensweisen fallen nicht mehr unter die individuelle Autonomie, denn diese endet dort, wo diejenige der anderen beginnt. Begründen lässt sich dies mithilfe eines weiteren risikoethischen Konzepts, dem Prinzip der Sorgfaltspflichten. Dieses stellt spezifische Anforderungen an sorgfältiges Handeln im Zusammenhang mit Risikoübertragungen, die sich sowohl auf die Minimierung der Eintrittswahrscheinlichkeit eines Schadens als auch des erwarteten Schadensausmaßes beziehen. Bei risikobehafteten Handlungen müssen angemessene Präventionsmaßnahmen getroffen werden, sowohl auf Seiten des Risikoverursachers als auch des Betroffenen (vgl. Rippe, 2013, S. 531).

Da Risikoübertragungen im Verkehrskontext zumindest der Art, nicht aber der Höhe nach reziprok sind, gilt diese Forderung prinzipiell für alle Verkehrsbeteiligten und äußert sich in vielfältigen Verhaltensweisen. So sollten Fußgänger sich aufmerksam und vorsichtig im Verkehr bewegen, Radfahrer auf ausreichende Abstände achten und Handzeichen für Richtungsänderungen einsetzen. In Bezug auf autonome Fahrsysteme sind hier vor allem Anforderungen an eine defensive und vorausschauende Fahrweise zu nennen:

> Insofar as AVs are morally permitted to impose risks of harm to road users given their prudential goal of time efficiency, the AV must drive with the required level of caution so that it can come to a stop in what if cases that could easily occur, given the AV's evidence about relevant features of the driving environment. (Keeling, 2022, S. 44–45)

Die Passagiere eines autonomen (Privat-)Fahrzeugs müssen ihrerseits sicherstellen, dass die Sicherheitsanforderungen an die Fahrtüchtigkeit, technische Funktionsfähigkeit etc. erfüllt sind. In diesem Sinne nimmt die Person, die in Beispielszenario 5 ›Unbeteiligte auf Bürgersteig‹ die Straße vorschriftsmäßig überquert, ihre Sorgfaltspflichten ernst. Ob das auch auf die Insassen des selbstfahrenden Fahrzeugs zutrifft, dessen Bremsen unvorhergesehen versagen, kann hingegen nicht final geklärt werden. Der Motorradfahrer in Beispielszenario 4 ›Motorradfahrer mit/ohne Helm‹ generiert durch seinen wissentlichen Verzicht auf Schutzkleidung ein Risiko, das dem anderen, der seinerseits seine Sorgfaltspflichten diesbezüglich

erfüllt hat, nicht zugemutet werden darf. Auch das ungebremste Auffahren des LKW in Beispielszenario 8 ›Herannahender LKW‹ ist sicherlich unzumutbar, ebenso das verkehrswidrige Schneiden der Kurve durch den Schulbus in Beispielszenario 7 ›Klippe‹. In diesem letzten Fall wird zudem deutlich, dass auch andere ethische Probleme für die risikoethische Bewertung relevant sein können, denn Sorgfaltspflichten können nicht immer eindeutig geklärt werden. In diesem konkreten Szenario ist das Fehlverhalten dem Fahrer des Busses zuzuordnen, nicht aber den Insassen, die ebenfalls zu den Betroffenen zählen. Besonders in Szenarien, in denen Personen stellvertretend für andere risikobehaftete Entscheidungen treffen, dürfen Fragen der Verantwortlichkeit nicht außer Acht gelassen werden.

Diese beispielhaften Szenarien verdeutlichen, dass gemäß unserer ethischen Intuition gewisse Risiken bei der Bewertung regelwidrigen Verhaltens als unzumutbar einzustufen sind. Welche praktischen Implikationen sich daraus ergeben, bleibt allerdings zumindest teilweise unklar. Sind Sorgfaltspflichten erfüllt, gelten Risiken als zumutbar. Somit begründet das Grenzkriterium der Zumutbarkeit im Kern die moralische Pflicht, Sorgfaltspflichten zu ergreifen, um unzumutbare Risiken zu vermeiden und dadurch die individuelle Autonomie jedes Einzelnen zu wahren. Die entscheidende Frage ist nun, in welchem Umfang Sorgfaltspflichten im Kontext von Dilemma-Situationen gefordert werden sollten und wann diese als erfüllt anzusehen sind.

Der moralische Charakter von Entscheidungen zur Gestaltung von Unfallalgorithmen wurde bereits vielfach betont. Agieren Verkehrsbeteiligte in dem Bewusstsein, dass sie Risiken auf andere übertragen, so stellen sich sämtliche Handlungsentscheidungen im Verkehrsgeschehen als moralische dar. Es ist also zu fragen, welches Risiko angesichts der eigenen Sorgfaltspflichten, die man realisiert oder unterlassen hat, zumutbar ist. Aufgrund der moralischen Dimension von Unfalldilemmata liegt es nahe, an dieser Stelle einen moralischen Sorgfaltsbegriff zugrunde zu legen: Sorgfalt im moralischen Sinne bedeutet, unter Berücksichtigung spezifischer situativer Merkmale zu begründeten, rechtfertigbaren Entscheidungen zu gelangen. Dies umfasst eine moralische Verantwortung, der man sich nur stellen kann, indem man sich der Pflicht zur sorgfältigen Einbeziehung sämtlicher moralisch relevanter Gründe annimmt. Als moralfähige Wesen schulden wir einander eine gewisse Sorgfalt in der Begründung unserer moralischen Entscheidungen, die sich in einen

gesellschaftlichen Kontext stellen lassen (vgl. Nyholm & Smids, 2016, S. 1278–1279). Dieser Umstand begründet die Notwendigkeit, ethische Entscheidungsprozesse gewissenhaft zu durchdenken:

> [...] 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. In ethics, the process of thinking through a problem is as important as the result. (Lin, 2014a, o. S.)

Das zentrale moralische Argument, das im Rahmen dieses Sorgfaltsprozesses zur Anwendung kommt, besteht in der Reziprozität, durch die soziale Gefüge – oder im Fall automatisierter Mobilität: soziotechnische Gefüge – erst tragfähig werden. Die Gegenseitigkeit von Risikoübertragungen bildet die Grundlage des zweiten Grenzkriteriums, indem sie es einerseits erlaubt, die eigene Freiheit durch risikobehaftetes Handeln zu verwirklichen, andererseits aber zugleich an die Pflicht zur Wahrung der Autonomie der anderen appelliert, denen die gleiche Freiheit zusteht. Dabei bereitet das Kriterium der Zumutbarkeit in gewisser Weise die Basis für die Diskussion gerechtigkeitsbezogener Fragen, wie sie im folgenden Unterkapitel entwickelt wird. Gerechtigkeitsethische Fragen sind nicht vollständig unabhängig von jenen der Zumutbarkeit, weshalb an geeigneten Stellen Rückbezüge stattfinden bzw. das Kriterium der Zumutbarkeit nochmals aufgegriffen und erweitert wird:

> Es geht daher im ›harten‹ Kern der Debatte zu Technikakzeptanz einerseits um die *unfreiwillig einzugehenden Zumutungen und ihre gesellschaftliche Verteilung, die der – im Prinzip nicht in Frage gestellte – technische Fortschritt mit sich bringt* […]. (Grunwald, 2005, S. 58, Hervorh. i. Orig.)

7.3.3 Die (relative) Frage der Gerechtigkeit: Zwischen Reziprozität und Vorteilsausgleich

7.3.3.1 Unfallalgorithmen vor dem Hintergrund egalitärer Gerechtigkeitskonzeptionen

Risikobehaftete Entscheidungen als eine Frage fairer Risikoverteilungen anzusehen, ist in anderen Anwendungskontexten bereits etabliert. Goodall (2017, S. 496) nennt als beispielhafte Problemstellun-

gen die Zuweisung von Organspenden im Gesundheitswesen, die Zulässigkeit von Strahlungsbelastungen, die Beibehaltung der Wehrpflicht oder Vorgaben für industrielle Sicherheitsstandards. Wie in Kap. 3.1.1 bereits gezeigt wurde, sind weder Algorithmen an sich noch ihre Effekte wertneutral. Neben der Forderung eines individuell zumutbaren Risikos wirft die Unvermeidbarkeit von Schäden, die charakteristisch für ethische Dilemmata ist, auch distributive Fragen zur moralischen Zulässigkeit von Risikoübertragungen auf; »a choice is required about how to distribute harms or risks of harm between multiple people whose interests are in conflict.« (Keeling, 2022, S. 50) Welche gerechtigkeitsethischen Kriterien sollen bei der Verteilung von Vor- und Nachteilen aus spezifischen Risikokonstellationen Anwendung finden? Da der thematische Rahmen dieses Buches keine umfassende philosophische Auseinandersetzung mit komplexen gerechtigkeitsethischen Fragen vorsieht, werden im Folgenden lediglich einige grundlegende Überlegungen zu problematischen Aspekten skizziert.

Die prominente Gerechtigkeitskonzeption von John Rawls (1971), welche in der gerechtigkeitsethischen Debatte als Dreh- und Angelpunkt gilt, fußt auf einem kantianisch geprägten Verständnis individueller Autonomie und Freiheit, die in einem Spannungsverhältnis von Individual- und Gesamtnutzen stehen. Letzterer ist dabei nicht aggregativ, sondern als Menge der von allen Individuen geteilten Vorteile zu verstehen. Durch diese wird schließlich eine Fairness konstituiert, die sich dadurch auszeichnet, dass diejenigen, die profitieren, auch die Nachteile tragen. Das Postulat einer gerechten Verteilung beinhaltet also Fairnessansprüche (*fairness claims*). Der relative Charakter des Fairnessgedankens impliziert, dass etwas immer in Bezug auf etwas anderes fair ist. Die Idee der Fairness vollzieht sich quasi erst im interpersonellen Vergleich (vgl. Broome, 1984, S. 43). Bezogen auf Unfallalgorithmen bedeutet das: Es ist nicht nur relevant im Sinne des ersten Grenzkriteriums, dass die Risiken für die Einzelnen zumutbar sind, sondern auch, wie diese im Vergleich zu denen anderer Personen ausfallen. Ergänzend zu der Forderung zumutbarer Risiken lässt sich daher im Hinblick auf eine deontologische Risikoethik für Unfallalgorithmen folgendes (relatives) Prinzip formulieren:

(2) *Risikoübertragungen im Kontext von Unfallalgorithmen können genau dann als moralisch zulässig gelten, wenn die aus der jeweiligen Risikokonstellation resultierenden Vor- und Nachteile unter Berücksichtigung aller ethisch relevanten Aspekte auf die Einzelnen fair verteilt sind.*

Traditionell beziehen sich gerechtigkeitsethische Fragen – und damit auch die Idee der Fairness – auf die Verteilung von Vorteilen und Lasten, die aus sozialer Kooperation resultieren. Die Verkehrsinfrastruktur ist ein öffentliches Gut, durch das Bürger ihr Recht auf Mobilität verwirklichen können (vgl. Dietrich, 2021, S. 728–729). In Dilemma-Szenarien bestehen relevante Vor- bzw. Nachteile in Bezug auf Sicherheit bzw. Nicht-Schädigung: Einen Vorteil erlangt ein Individuum dann, wenn eine Trajektorie gewählt wird, die es einem geringeren Risiko aussetzt als die anderen Beteiligten des Dilemmas. Risiken für Einzelne sind demnach so zu gestalten, dass sich daraus diejenige Verteilung von Vor- und Nachteilen ergibt, die als fair angesehen werden kann.

Unter welchen Voraussetzungen können risikoinduzierte Vor- bzw. Nachteile als fair gelten? Hinsichtlich philosophischer Konzeptionen der Gerechtigkeit verläuft eine scharfe Trennlinie zwischen egalitaristischen und non-egalitaristischen Ansätzen,[274] die jeweils unterschiedliche Aspekte der Gerechtigkeit akzentuieren. Gerechtigkeit definiert einen Zustand, in dem jedes Individuum das hat oder erhält, worauf es einen gültigen Anspruch hat; ein solcher kann beispielsweise über ein Recht begründet sein. Egalitaristische und non-egalitaristische Konzeptionen unterscheiden sich in der Begründung von Ansprüchen und Rechten. Egalitaristische Ansätze bestimmen Gerechtigkeit über eine relationale Interpretation von Gleichheit, die sich daran bemisst, wie andere im Vergleich zu dem betrachteten Individuum dastehen. Es sind zwei Arten egalitaristischer Gerechtigkeit zu unterscheiden. Gemäß der einen wird Gerechtigkeit mit

274 Zu den prominentesten Vertretern egalitaristischer Gerechtigkeitskonzeptionen gehören John Rawls (1971), Amartya Sen (1980), Ronald Dworkin (1981a, 1981b, 1987a, 1987b, 2000), Thomas Nagel (1979a, insb. Kap. 8; 1991) und Richard Arneson (1989). Non-egalitaristische Positionen, die sich zumeist als kritische Antwort auf egalitaristische Grundideen verstehen, werden hingegen von Harry Frankfurt (1997), Joseph Raz (1986, insb. Kap. 9) und Derek Parfit (2003) vertreten.

einem intrinsischen Wert der Gleichheit identifiziert, der um seiner selbst willen zu beachten ist. Daraus folgt, dass ungleiche Zustände mit normativen Implikationen verbunden sind (vgl. Anderson, 2000, S. 121). Während kontingente Ungleichheiten ausgeglichen werden müssen, sind selbstverschuldete bzw. leistungsabhängige hingegen akzeptabel.[275]

In diesem Zusammenhang wirken egalitäre Rechte, die als ›Rechte auf das Gleiche‹, z. B. gleiche Freiheit, gleiche Partizipation, zu verstehen sind; diese Form wird daher auch als egalitäre Gerechtigkeit bezeichnet. Die entsprechenden Rechte werden dabei non-egalitaristisch begründet, denn sie haben nur in Bezug auf ihren Inhalt bzw. die Größe des geforderten Gutes relationalen Charakter. Wichtig ist nicht nur, *dass* andere etwas bekommen, sondern *wie viel*. In einer anderen Konzeption egalitaristischer Gerechtigkeit wird Gleichheit über Ansprüche definiert, die aus zwei Komponenten bestehen: zum einen deshalb, weil andere etwas auch haben; zum anderen, weil ein Recht auf Gleichbehandlung und eine damit korrespondierende Pflicht besteht. Der Gleichheitsanspruch gründet sich auf die bestehende Gleichheit von Personen; diese wiederum folgt aus der Zugehörigkeit zu einem sozialen System, aus der sich normative Implikationen ergeben – verstanden als das Recht auf Gleichbehandlung in Bezug auf alle Ansprüche, die mit der Mitgliedschaft in diesem System in Zusammenhang stehen.

Im Gegensatz dazu geht der Non-Egalitarismus davon aus, dass Menschen von Natur aus ungleich sind. Da Gleichheit hier keinen intrinsischen Wert an sich darstellt, beinhalten bestehende Ungleichheiten in diesem Fall keine normativen Implikationen (vgl. Nida-Rümelin, 2006, S. 11–13). Gerechtigkeitsansprüche sind unabhängig davon, ob andere etwas haben; sie sind suffizienzorientiert, d. h. darauf gerichtet von etwas *genug* zu haben. Ziel ist es daher, für jeden die (Minimal-)Bedingungen zu schaffen, die ein menschenwürdiges Leben erlauben (vgl. Margalit, 1996, 1997), wobei nicht notwendigerweise die Beseitigung von Ungleichheiten impliziert ist (vgl. Nida-Rümelin, 2006, S. 9–10).

Welche Gerechtigkeitskonzeptionen erscheinen für den Anwendungskontext von Unfallalgorithmen adäquat? Plausible Gründe

275 Diese Unterscheidung ist beispielsweise Grundlage der Konzeption von Dworkin (2000).

sprechen dafür, dass ein egalitaristischer Ansatz dem Anwendungsproblem am besten gerecht wird. Zum einen ist eine konzeptionelle Suffizienzorientierung, wie sie non-egalitaristische Konzeptionen vorsehen, für die spezifische Problematik von Dilemmata nicht zweckmäßig. Da es sich bei Dilemmata um Situationen mit sozialer Dynamik handelt, reicht es nicht aus, jedes Individuum nur isoliert zu sehen, denn Risikoübertragungen und daraus resultierende Vor- bzw. Nachteile sind entscheidend von den Interaktionen und Handlungen Einzelner abhängig. Weiterhin ist die Risikoethik naturgemäß relational, es sind stets Beziehungen zwischen Exponenten und Exponierten relevant; risikoethische Fragen lassen sich daher unter einer non-egalitaristischen Auffassung nur begrenzt diskutieren.

Dass die Manifestation kontingenter Ungleichheiten durch Praktiken algorithmischer Verteilungsstrategien kaum zu rechtfertigen ist, kann anhand einer einfachen Anwendung des Prinzips der Schadensminimierung veranschaulicht werden. Würden sich Unfallalgorithmen an der Maßgabe orientieren, möglichst geringe persönliche Schäden zu verursachen, bestünde die Gefahr, dass Personen geopfert werden, »um die eigentlichen Verursacher des Unfalles vor den Konsequenzen ihres Fehlverhaltens zu schützen.« (Hevelke & Nida-Rümelin, 2015c, S. 19) Dies wäre im Fall des Motorradfahrers in Beispielszenario 4 ›Motorradfahrer mit/ohne Helm‹ gegeben. Auch hinsichtlich bestimmter Fahrzeugtypen käme es zu diskriminierenden Effekten, wenn leichtere Fahrzeuge mit geringeren Sicherheitsstandards möglichst als Kollisionsobjekte vermieden und auf diese Weise sicherer für ihre Insassen werden. Derartige Mechanismen setzen systematisch Fehlanreize, sich durch nicht-regelkonformes Verhalten einen persönlichen Vorteil zu verschaffen. Dies erscheint uns intuitiv unfair; es ist nicht einsichtig, weshalb andere aufgrund des Fehlverhaltens von Einzelnen Nachteile in Kauf nehmen sollten. Dieses Beispiel pointiert, was für risikoethische Situationsbewertungen generell Gültigkeit besitzt: Verursacher von (höheren) Risiken sollten auch die (größeren) Nachteile tragen.

Zum anderen ist das Bestehen kontingenter Ungleichheiten bei Risikoübertragungen prinzipiell unvereinbar mit der Nicht-Verrechenbarkeit inkommensurabler Werte, wie sie im Rahmen der metaethischen Analyse in Kap. 5 begründet wurde. Die Gleichheit und der sich daraus ableitende Anspruch auf Gleichbehandlung eines jeden Verkehrsbeteiligten beruhen auf der intrinsischen Unvergleich-

barkeit des Individuums und nicht auf der Zugehörigkeit zu einem (sozialen) System. Ein solches lässt sich für den Anwendungskontext Verkehr nicht plausibel denken, denn es existiert keine spezifische Gruppendynamik und keine ethisch relevante Abgrenzung zu Personen außerhalb des Verkehrsgeschehens.

Daraus folgt, dass egalitäre Gerechtigkeitskonzeptionen im Hinblick auf die Problemstellung am besten geeignet sind. Ein anwendungsnahes Argument stützt diese These: Auch wenn Gleichheit im Sinne gleicher Rechte ›auf das Gleiche‹ das erklärte Ziel egalitärer Gerechtigkeit ist, so lässt sich daraus nicht schließen, dass jede Form der Ungleichheit zwangsläufig ungerecht ist. Wie Nida-Rümelin et al. (2012, S. 141) konstatieren, kann auch eine Ungleichverteilung fair sein, nämlich genau dann, wenn Ansprüche auf bestimmte Vorteile anderen gegenüber bestehen.

Die entscheidende Frage ist, wie sich solche Ansprüche begründen lassen. Zuvor wurde anhand des Prinzips der Schadensminimierung erläutert, dass ein spezifisches Design von Unfallalgorithmen genau dann zulässig ist, wenn dadurch das Risiko jedes Einzelnen in gleichem Maße reduziert wird und es daher im Interesse jedes Einzelnen ist. Hevelke und Nida-Rümelin (2015c, S. 17) zeigen, dass Letzteres jedoch nicht den entscheidenden Punkt trifft; für die Zulässigkeit ist vielmehr die Erfüllung individueller, begründeter Ansprüche maßgeblich. Konsequenterweise müssen nicht nur Benachteiligungen ausgeglichen werden, sondern auch ungerechtfertigte Bevorteilungen. Gewisse Gruppen müssten im Zuge eines Benachteiligungsausgleichs ihre Sicherheitsvorteile aufgeben, deren Verlust jedoch genau dann unproblematisch im Sinne einer Instrumentalisierung der entsprechenden Individuen ist, wenn sie auf diese Vorteile gar keinen moralisch begründbaren Anspruch haben. Ein solcher ist insbesondere dann nicht vorhanden, wenn diese Vorteile im Gegenzug mit Nachteilen für andere verbunden sind. Ein Ausgleich in diesem Sinne ungerechtfertigter Vorteile kann daher mindestens als zulässig, möglicherweise sogar als wünschenswert angesehen werden.

Aus dem Postulat egalitärer Gerechtigkeit folgt somit im Hinblick auf das Ziel einer gerechten Ordnung die normative Forderung, bestehende kontingente Ungleichheiten auszugleichen. Dabei gilt, dass sich Vor- und Nachteile ausschließlich über berechtigte Ansprüche legimitieren lassen. Hier schließt sich nun die Frage an, inwiefern

berechtigte Ansprüche im Kontext von Unfalldilemmata bestehen und wie sich diese begründen lassen. Dies wird im Folgenden anhand von zwei ausgewählten Aspekten erörtert.

7.3.3.2 Risikoverursachung und Vulnerabilität als Begründungsversuche berechtigter Vorteilsansprüche

Vor dem Hintergrund des Fairnesspostulats sind Sicherheitsvorteile nur begründbar, wenn berechtigte Ansprüche auf eine Bevorzugung vorliegen. Eines der am häufigsten diskutierten Verteilungsprinzipien in dieser Hinsicht ist jenes, welches die Sicherheit der Insassen autonomer Fahrzeuge im Fall einer Kollision priorisiert. Das zentrale Argument, das in einschlägigen Diskussionen vorgebracht wird, lautet, dass autonome Fahrzeuge im Vergleich zum herkömmlichen motorisierten Verkehr die Verkehrssicherheit erhöhen und daher ihren Insassen ein Vorteil gegenüber den Insassen herkömmlicher Fahrzeuge im Mischverkehr zugesprochen werden sollte. Berkey (2022, S. 217–219) hebt hervor, dass dies allerdings in mancherlei Hinsicht unplausibel ist: Erstens wäre eine solche Schlussfolgerung nur unter der Einschränkung denkbar, dass alle Individuen über (annähernd) gleiche Zugangschancen zur automatisierten Mobilität verfügen, was praktisch jedoch aufgrund unterschiedlicher Wohlstandsniveaus einerseits und der hohen Preisklasse autonomer Fahrzeuge andererseits nicht gegeben ist. Zweitens widerspräche eine generelle Programmierung auf Insassenpriorisierung dem Prinzip der Schadensminimierung. Die Algorithmen würden in einer Weise verzerrt, die Anreize für die Wahl von Trajektorien setzt, welche jegliche noch so geringe Schäden von Insassen autonomer Fahrzeuge abwenden und daher in überproportional hohen Schäden für Insassen herkömmlicher Fahrzeuge resultieren. Drittens würde auf diese Weise bereits die Möglichkeit einer fairen Verteilung untergraben. Da potenzielle Nutzer sich tendenziell eine Priorisierung ihrer eigenen Sicherheit wünschen (vgl. Bonnefon et al., 2015, S. 5–8), existieren hohe Anreize für Hersteller, ihre Algorithmen dementsprechend zu gestalten. Faire Verteilungen von Vor- und Nachteilen sind daher ein Problem kollektiven Handelns: Sobald ein einziger Hersteller Fahrzeuge anbietet, welche die Nutzersicherheit priorisieren, würden aufgrund des Marktdrucks alle anderen zur Anpassung

ihrer Produkte genötigt, wodurch sämtliche Fairnessbestrebungen hinfällig würden (vgl. Berkey, 2022, S. 222–223).

Auch wenn sich ein Anspruch auf Vorteile für die Insassen autonomer Fahrzeuge nicht glaubwürdig rechtfertigen lässt, bleibt der Grundgedanke, dass das Ausmaß einer Risikoverursachung ethisch relevant ist, ein prinzipiell valides Argument. Es ist offensichtlich, dass nicht-motorisierte Verkehrsteilnehmer wie Radfahrer oder Fußgänger verhältnismäßig geringere Risiken für andere herbeiführen, sodass es gute Gründe gibt, von einer ursprünglich angenommenen Gleichverteilung zu ihren Gunsten abzuweichen. Aus der geringeren Risikoübertragung lässt sich durchaus ein legitimer Anspruch auf Sicherheitsvorteile ableiten, was als Umkehrschluss der eingangs in diesem Unterkapitel festgestellten Intuition aufgefasst werden kann: Nicht nur soll gelten, dass, wer höhere Risiken bewirkt, mehr Nachteile in Kauf nehmen muss, sondern auch, dass, wer weniger Risken verursacht, weniger Nachteile tragen sollte:

> […] it seems to me that a potentially significant deviation from an equal distribution of risks between, on the one hand, occupants of vehicles, and on the other, pedestrians and cyclists, might be required. This is because while pedestrians do not expose others to any risks at all just in virtue of using the road (or sidewalk), and cyclists expose others only to relatively small and (on average) less serious risks, those who choose to ride in vehicles expose others to (much more) substantial risks of harm, including risks of serious injury and death, just in virtue of their use of vehicles. It seems plausible that those who choose to introduce risks of this kind in order to enjoy the benefits of the activities that unavoidably involve these risks, should, where possible, at least bear a greater share of the risks than those who are not engaged in the activities that impose them. (Ebd., S. 217)

Die Prämisse, die dieser Argumentation zugrunde liegt, hat allerdings nur so lange Bestand, wie sich die betreffenden Personen an gewisse Regeln halten. Es ist einleuchtend, dass durch regelwidriges Verhalten höhere Risiken produziert werden.[276] Verkehrsteilnehmer müssen plausiblerweise davon ausgehen können, dass sich alle regelkonform verhalten; Hevelke und Nida-Rümelin (2015a, S. 222–224) betrachten die Vorhersehbarkeit des Verhaltens im Straßenverkehr als ausschlaggebend dafür, dass einer bestimmten Person Schutz

276 Siehe hierzu auch die bisher diskutierten Argumente in Kap. 4.4.1.

gewährt werden kann. Abney (2022, S. 264–272) legt dar, dass Forderungen nach einer Gleichverteilung häufig auf der Annahme einer grundsätzlichen moralischen Gutheit (*moral goodness*) aller Personen basieren. Wird diese angesichts praktischer Evidenz aufgegeben, so wird deutlich, dass moralische Urteile und damit auch Argumente in Bezug auf bestimmte risikoethische Verteilungsmechanismen wesentlich von den Motiven und Intentionen von Individuen abhängen:

> If we accept that moral judgments will crucially depend on the intentions of the agents involved, then people with immoral purposes will morally require a different response from those with moral purposes—and this is hardly a novel concept in law and policy; there are commonplace ethical and legal distinctions between mere negligence versus intentional abuse. (Ebd., S. 271–272)

Anstatt über individuelles Fehlverhalten hinwegzusehen, muss dieses zur Verantwortung gezogen werden, wenn man die Idee der Gerechtigkeit ernst nimmt. Hier muss stets im Einzelfall geprüft werden, wie ein Regelbruch konkret zu bewerten ist:

> If we take justice seriously, we cannot assume the moral equality of all persons in calculating consequences for moral decision-making. Some people deserve to be treated worse than others. In other words: some people, in some circumstances, deserve to be hit. In fact, they deserve to die – regardless of utility. (Ebd., S. 265)

Einen weiteren, ebenfalls häufig thematisierten Aspekt in diesem Kontext stellt die Gewährung von Sicherheitsvorteilen auf der Basis einer Kategorisierung von Verkehrsbeteiligten dar. Dieses Vorgehen entspricht der Konzeption von Theorien sozialer Gerechtigkeit, die neben dem zu verteilenden Gut, deren Empfängern und dem formalen Verteilungsprinzip auch die Eigenschaften der Empfänger, welche für die Verteilung relevant sind, als eine Hauptdimension gerechter Verteilungen begreifen (vgl. Dietrich & Weisswange, 2019, S. 229–230). Wie die Auswertung statistischer Daten zeigt, werden bestimmte Gruppen von Verkehrsteilnehmern unverhältnismäßig oft in tödliche Verkehrsunfälle mit Fahrzeugen verwickelt (vgl. Dietrich, 2021, S. 728–729; Mullen et al., 2014, S. 238). Dies ist auf diverse Gründe zurückzuführen, von denen für den Kontext dieser Forschungsarbeit vor allem zwei relevant sind: zum einen auf ein häufig besonders rücksichtsloses Verhalten von Autofahrern gegenüber

Radfahrern (vgl. Europäische Kommission, 2020, S. 31) und zum anderen darauf, dass gewisse Gruppen durch eine erhöhte Schadensneigung charakterisiert sind. So ist sich die Forschungsliteratur dahingehend einig, dass der motorisierte Verkehr gegenüber Fußgängern generell begünstigt ist (vgl. Dietrich, 2021, S. 727–728).[277]

Dies führt dazu, dass der erwartete Schaden für Fußgänger und Radfahrer im Fall einer Kollision mit einem Fahrzeug deutlich höher ausfällt als für die Fahrzeuginsassen. Ursächlich dafür sind primär physikalische Faktoren wie die deutlich geringere Masse, die geringere Bewegungsgeschwindigkeit und nicht vorhandene Schutzvorrichtungen wie Karosserie oder Rückhalteeinrichtungen. Es ist also plausibel anzunehmen, dass zwischen Personengruppen ungleiche Ausgangsbedingungen hinsichtlich der Möglichkeiten herrschen, ihre Ansprüche auf Schutz von Leben und Gesundheit zu verwirklichen (vgl. Luetge, 2017, S. 552–553).[278] Im Kontext automatisierter Mobilität verschärft sich diese Problematik ungleicher Risikoexpositionen zusätzlich:

> The most severe concerns are that AVs will reinforce existing inequalities and introduce new injustice aspects in the context of road traffic. An existing inequality is the unequal exposure to risks of bodily harm induced by a traffic accident between different types of road users. (Dietrich, 2021, S. 727)

Ein weit verbreiteter Terminus des einschlägigen ethischen Diskurses ist der Begriff der *Vulnerable Road Users* (*VRUs*) (= gefährdete Verkehrsteilnehmer)[279], der als Sammelbegriff dient für »non-motorised road users, such as pedestrians and cyclists as well as motorcyclists and persons with disabilities or reduced mobility and orientation.« (European Commission Mobility und Transport, 2023) Ethisch-politische Direktiven und Empfehlungen verweisen auf die

277 Dies führt Dietrich (2021, S. 728) auf den Umstand zurück, dass die Infrastruktur der Verkehrsstraßen in urbanen Gesellschaften motorisierten Verkehr begünstigt; siehe auch Nello-Deakin (2019, S. 699–704).

278 Auch innerhalb der Gruppe der Fahrzeuge bestehen gewichtige Unterschiede hinsichtlich möglicher Risikoexpositionen aufgrund von unterschiedlicher Masse und aktiven sowie passiven Schutzvorrichtungen (vgl. Dietrich, 2021, S. 728).

279 In einschlägigen Publikationen findet sich das Akronym *VRU* sowohl mit als auch ohne Pluralendung ‚-s'. Aus Gründen der Einheitlichkeit wird in diesem Buch die Pluralendung verwendet.

Notwendigkeit einer Kategorisierung von Verkehrsteilnehmern, um das Sicherheitsniveau vulnerabler Gruppen zu heben. Die deutsche Ethik-Kommission beispielsweise versteht eine »Programmierung der Fahrzeuge zu defensivem und vorausschauendem, schwächere Verkehrsteilnehmer […] schonendem Fahren« (Di Fabio et al., 2017, S. 10) als Teil einer übergeordneten Strategie zur Erhöhung der Verkehrssicherheit durch automatisierte Mobilität. Die Expertengruppe der Europäischen Kommission (2020, S. 30–32) betrachtet den Schutz von *VRUs* als Element einer Solidarität, die sich als Ergänzung zu gerechtigkeitsethischen Anforderungen ergibt. In diesem Sinne ist es Teil der Designaufgabe autonomer Fahrzeuge, *VRUs* besonders zu schützen: »CAVs should, among other things, adapt their behaviour around vulnerable road users instead of expecting these users to adapt to the (new) dangers of the road.« (Ebd., S. 7)

Ethische Begründungsgrundlagen werden im Rahmen politischer Richtlinien meist nur angedeutet. Die Expertengruppe der Europäischen Kommission nennt die Verringerung der Geschwindigkeit sobald *VRUs* erkannt werden und das Einhalten größerer Abstände zu diesen als mögliche praktische Maßnahmen, um *VRUs* zu schützen. Sie sollen allerdings nur solange in Kraft bleiben, wie sich der Gesamtschaden für andere Verkehrsteilnehmer nicht erhöht (vgl. ebd., S. 31). Wie in Fällen verfahren werden soll, in denen diese Voraussetzung nicht erfüllt ist, wird hingegen offengelassen.

Auch in der ethischen Forschungsliteratur bezüglich des Designs von Unfallalgorithmen werden Konzepte thematisiert, gefährdete Gruppen im verkehrlichen Umfeld besonders zu schützen. Der Schutz von *VRUs* wird dabei als Postulat egalitärer Gerechtigkeit gedeutet, das den Ausgleich der Ungleichheiten beinhaltet, die durch bestehende Unterschiede bei der Schadensneigung hervorgerufen werden. Wichtig ist zu beachten, dass den Interessen von *VRUs* durch den besonderen Schutz nicht mehr Gewicht gegeben, sondern lediglich ihrem legitimen Anspruch auf Ausgleich entsprochen wird:[280] »This programming would not amount to giving greater value to the safety of cyclists—it would rather be an attempt to correct safety inequalities, which partly result from the current behaviour of human drivers.« (Ebd., S. 31) Martínez-Buelvas et al. (2022) stellen

280 Dieses Argument lässt sich als Antwort auf Keelings (2018a) Kritik verstehen, dass die Ansprüche der Schlechtergestellten übermäßig gewichtet würden.

mögliche Strategien für die Gestaltung einer sicheren und gerechten Interaktion zwischen autonomen Fahrzeugen und *VRUs* vor, aus denen sich eine Forschungsagenda ableiten lässt. Dabei interpretieren sie Verkehrsgerechtigkeit in Bezug auf Gleichheit, Fairness und Zugang.

Auf Basis der jeweiligen Vulnerabilität ergeben sich also berechtige Ansprüche auf Sicherheitsvorteile für bestimmte Gruppen. Nun sind allerdings nicht nur physikalische Faktoren dafür verantwortlich, wie hoch der individuelle Schaden im Fall einer Kollision ausfällt. Auch persönliche Merkmale wie das Alter, die Fitness oder der Gesundheitszustand betreffender Personen haben unter bestimmten Umständen Einfluss darauf, ob eine Person beispielsweise tödlich verletzt wird. Sollen derartige Aspekte im Rahmen von Unfallalgorithmen berücksichtigt werden, führt dies zu ernstzunehmenden ethischen Problemen. Im Kontext von Algorithmen und Künstlicher Intelligenz wird der Begriff der Fairness häufig mit der Freiheit von systematischen Verzerrungs- (*biases*) und Diskriminierungseffekten assoziiert. Der Ausschluss jeglicher ungerechtfertigten Bevorzugung bzw. Benachteiligung bestimmter Personengruppen in automatisierten Entscheidungsprozessen gilt als eine der Grundanforderungen an ein robustes Systemdesign.[281] Inwiefern Unfallalgorithmen – und algorithmische Entscheidungssysteme im Allgemeinen – in konkreten Anwendungsfällen diskriminierend sind, ist eine häufig thematisierte, kontroverse Fragestellung. Dies ist u. a. darauf zurückzuführen, dass einerseits die Definitionen dessen, was unter problematischer Diskriminierung verstanden wird, sehr vage sind.

Andererseits kann Diskriminierung in Entscheidungen maschinell trainierter Algorithmen auf verschiedene, oftmals subtile Weisen erzeugt werden (vgl. Zweig, 2019, Kap. 8). Eine anwendungsorientierte Sichtweise im Kontext autonomer Fahrsysteme legt Leben (2022, S. 132–138) vor. Demnach sind diskriminierende Handlungen nicht per se moralisch schlecht, sondern nur dann, wenn sie Schaden verursachen (vgl. Thomsen, 2023) bzw. wenn sie ungerechtfertigt sind. Letzteres ist genau dann der Fall, wenn einzelne Personen oder Gruppen aufgrund von Faktoren benachteiligt werden, die in Bezug auf die Aufgaben, die der Algorithmus erfüllen soll, unerheb-

281 Für eine Übersicht zum gegenwärtigen Forschungsstand diesbezüglich siehe Hütt und Schubert (2020).

lich sind. In anderen Worten: Ob Algorithmen in ungerechtfertigter Weise diskriminierend sind, hängt davon ab, welchem praktischen Zweck sie dienen. So besteht die primäre Aufgabe von Unfallalgorithmen in der Verteilung von Schäden; Faktoren, die für diese Zielsetzung als irrelevant gelten, sind beispielsweise sozialer Status oder Berufsgruppenzugehörigkeit. Diese sind als Grundlage für Entscheidungsalgorithmen nicht gerechtfertigt und würden daher berechtigterweise als diskriminierend angesehen. Jedoch sind auch Faktoren denkbar, die im Hinblick auf die aus einem Unfall resultierenden Personenschäden durchaus eine Relevanz besitzen. Statistisch gesehen besteht eine Korrelation zwischen bestimmten persönlichen Merkmalen wie dem Alter der Unfallopfer einerseits und deren physischer Vulnerabilität bzw. Schadensneigung andererseits; Alter konstituiert also Ungleichheiten im Sinne einer erhöhten Vulnerabilität. Gemäß Lebens Auffassung wäre demnach die Berücksichtigung des Alters als Entscheidungskriterium durchaus zu rechtfertigen.

Ein anderes hier relevantes Argument liefert Černý (2022, S. 34–38). Ihm zufolge ist ein Algorithmus, der sich in einem Dilemma für diejenige Trajektorie entscheidet, die eine jüngere Person bevorzugt, aus zwei Gründen nicht notwendigerweise diskriminierend. Zum einen ist das Vorliegen diskriminierender Handlungen an deren Unausweichlichkeit geknüpft: Ist es möglich, auch anders zu handeln? In Dilemmata ist dies naturgemäß nicht erfüllt, negative Konsequenzen sind unvermeidbar. Zum anderen ist das Alter nicht als Kriterium zu verstehen, aufgrund dessen entschieden wird, sondern lediglich als Variable einer Theorie, die Černý als »deprivation theory of the badness of death« (ebd., S. 34–35) bezeichnet. Deren Grundlage ist einerseits die These, dass der Tod auch gut für eine Person sein kann, nämlich genau dann, wenn er diese von einem schlechten Leben erlöst, und andererseits die Annahme, dass mit zunehmendem Alter die Lebensqualität abnimmt. Daraus leitet Černý die normative Implikation ab, dass es moralisch schlechter ist, eine jüngere Person zu töten:

> When the AV decides between a and b, it is deciding based on age. But —and this is crucial—the age is not in fact the criterion of the choice, but merely a variable determining the extent of the badness of death. In other words, the AV decides by means of age, but not based on age: the basis of its choice is comparing the badness of death between a and b. (Ebd., S. 37)

An Černýs Ansatz lässt sich kritisieren, dass er Konflikte mit grundlegenden Normen wie der Unvergleichbarkeit menschlichen Lebens aufwirft. Sowohl Lebens als auch Černýs Ansatz liegt der Gedanke zugrunde, dass es bei Entscheidungen, die *mithilfe* des Merkmals ›Alter‹ getroffen werden, nicht um eine Bevorzugung *aufgrund* des Alters an sich geht, sondern um die Beurteilung dessen, was in direktem Zusammenhang damit steht. Werden durch das Alter kontingente Ungleichheiten produziert – wie es im Hinblick auf die individuelle Vulnerabilität der Fall ist –, so entstehen daraus begründete Ansprüche auf Sicherheitsvorteile, die nur indirekt durch das Alter, vielmehr aber durch die daraus resultierende Vulnerabilität verursacht sind.[282]

Zusammenfassend ergeben sich sowohl aus einer Bewertung der Risikoverursachung als auch der individuellen Vulnerabilität bedeutende normative Implikationen, die in einem berechtigten Anspruch auf Sicherheitsvorteile bzw. eine Priorisierung nicht-motorisierter Verkehrsteilnehmer münden. Eine Gleichverteilung von Risiken ist hier nicht plausibel; Ziel ist es vielmehr, bestehende kontingente Ungleichheiten im Sinne einer ergebnisegalitaristischen Position fair zu berücksichtigen, welche die Erreichung eines Pareto-optimalen Zustands anvisiert.[283] Ungeachtet ihrer ethischen Relevanz stellen sich hinsichtlich der praktischen Implementierung Herausforderungen auf technischer Seite. Die Fairness risikoethischer Verteilungsstrategien hängt in letzter Konsequenz auch von der technischen Leistungs- und Perzeptionsfähigkeit der jeweiligen Systeme ab. Eine korrekte und zuverlässige perzeptive Differenzierung zwischen Kategorien von Verkehrsteilnehmern[284] ist Grundvoraussetzung dafür, dass *VRUs* ein besonderer Schutz zuteilwerden kann (vgl. Keeling, 2022, S. 47–53):

> In all these examples, the recommendation to provide greater road safety to a subset of road users must always be premised on evidence

282 Auch in Bezug auf den Faktor ›Geschlecht‹ bestehen möglicherweise derartige Ungleichheiten; dies ist allerdings noch nicht tiefergehend untersucht worden.

283 Der Egalitarismus kann generell auf drei Arten verstanden werden (vgl. Horn, 2003, S. 26): Verteilungsegalitarismus (gleich große Güter), Verfahrensegalitarismus (gleichmäßige Regelanwendung), Ergebnisegalitarismus (gleiche Ergebnisse).

284 Eine noch größere technische Hürde ist es freilich, die Intentionen der Handelnden korrekt zu erkennen.

> that it is technically possible for a CAV to detect and respond to these road users accurately and reliably, that some users' harm-to-exposure ratio is high, that improving road safety for one subset of road users does not raise the total harm inflicted to another category of road users above its current baseline. (Europäische Kommission, 2020, S. 32)

Im Folgenden werden zwei bedeutende Konzeptionen für egalitäre risikoethische Verteilungsstrategien rekonstruiert und kritisch beleuchtet.

7.3.3.3 John Rawls: Eine prioritaristische Antwort auf Ungleichheiten

Philosophische Gerechtigkeitskonzeptionen der neueren Zeit sind vornehmlich kontraktualistisch orientiert. Ankerpunkt ist John Rawls' vertragstheoretischer Ansatz, den er 1971 in seinem epochalen Werk *A Theory of Justice* vorgelegt hat und der dem bisherigen Gerechtigkeitsdiskurs eine neue Wendung fort von vornehmlich utilitaristischen Denkstrukturen gab.[285] Rawls begründet seine liberale Gerechtigkeitskonzeption in der Tradition hypothetischer Vertragstheorie und basierend auf dem kantianischen Verständnis von Freiheit, aus der Gleichheit (in politischer Hinsicht) folgt.[286] Die kantianisch geprägte Konzeption individueller Autonomie ist nicht nur mit einem Freiheits-, sondern auch mit einem Gleichheitsanspruch verbunden:

> Mit dem Freiheitsanspruch wird gefordert, dass jede Person sich idealiter zugleich als Autor der Gesetze bzw. der Maximen anderer Akteure ansehen können sollte, denen sie sich ausgesetzt sieht. Mit dem Gleichheitsanspruch wiederum wird gefordert, dass sich die Handlungsbeschränkungen, die sich aus dem Freiheitsanspruch ergeben, für das Handeln einer jeden Person gegenüber allen anderen Personen gelten; dies bedeutet, dass insbesondere die situative Ungleichbehandlung einzelner Personen die Möglichkeit des Ansehens aller Personen als Gleiche nicht grundlegend erschüttern darf. (Nida-Rümelin et al., 2012, S. 181)

285 Für eine Übersicht zu zeitgenössischen Ansätzen sozialer Gerechtigkeit siehe Pojman (2005).

286 Ebenso wie Rawls verteidigt auch Nida-Rümelin (2006) eine egalitaristische Position, die Gleichheit im Sinne der kantianischen Verknüpfung von Freiheit und Gleichheit versteht; Freiheit ist ohne Gleichheit nicht denkbar.

Gemäß der Rawls'schen Auffassung stellt sich Gerechtigkeit als Fairness dar, die dadurch konstituiert wird, dass Akteure in einer hypothetischen Entscheidungssituation hinter einem *Schleier des Nichtwissens* unvoreingenommen über die Grundprinzipien einer gerechten Ordnung beraten. Unter den Prämissen der Gleichheit aller Personen, knapper Ressourcen und überwiegend egoistischer Motivationen würden sie sich für zwei lexikalisch angeordnete Grundsätze entscheiden: Der erste bezieht sich auf die fundamentale Grundstruktur einer Gesellschaft und postuliert eine gleiche Verteilung individueller Freiheitsrechte. Der zweite widmet sich sozio-ökonomischen und ethischen Fragen, indem er faire Chancen beim Zugang zu öffentlichen Positionen fordert. In vorläufiger Form lauten die beiden Prinzipien folgendermaßen:

> First: each person is to have an equal right to the most extensive basic liberty compatible with a similar liberty for others. Second: social and economic inequalities are to be arranged to that they are both (a) reasonably expected to be to everyone's advantage, and (b) attached to positions and offices open to all. (Rawls, 1971, S. 60)

Im Rahmen seiner Diskussion des zweiten Grundsatzes formuliert Rawls schließlich das berühmte *Differenzprinzip (difference principle)*, das den Schlechtergestellten einer Gesellschaft im Hinblick auf Verteilungsfragen von Primärgütern Priorität einräumt:

> Social and economic inequalities are to be arranged so that they are both (a) to the greatest benefit of the least advantaged and (b) attached to offices and positions open to all under conditions of fair equality of opportunity. (Ebd., S. 83)

Daraus folgt, dass Ungleichheiten hier nicht per se schlecht sind, sondern genau dann akzeptabel, wenn sie den Schlechtergestellten zu einem Zustand verhelfen, der sie im Vergleich zu alternativen Güterverteilungen besserstellt. Vor diesem Hintergrund lassen sich die beiden Grundsätze wie folgt reformulieren:

> *First Principle*[:] Each person is to have an equal right to the most extensive total system of equal basic liberties compatible with a similar system of liberty for all. *Second Principle*[:] Social and economic inequalities are to be arranged so that they are both: (a) to the greatest benefit of the least advantaged, consistent with the just savings principle, and (b) attached to offices and positions open to all under conditions of fair equality of opportunity. (Ebd., S. 302, Hervorh. i. Orig.)

Wichtig ist anzuerkennen, dass das *Differenzprinzip* seinem Wesen nach ein Konzept gegenseitigen Vorteils ist; Ungleichheiten sind genau dann gerechtfertigt, wenn sie zum Vorteil aller dienen. Es begründet damit eine Position, die auf einer egalitären Gerechtigkeitsauffassung[287] fußt (vgl. Nida-Rümelin & Rechenauer, 2009, S. 304) und dabei einem prioritaristischen Grundsatz folgt:

> All social primary goods–liberty and opportunity, income and wealth, and the bases of self-respect–are to be distributed equally unless an unequal distribution of any or all of these goods is to the advantage of the least favored. (Rawls, 1971, S. 303)

Die Mehrheit der Ansätze, die ausgehend von den normativen Implikationen, die sich aus den berechtigen Ansprüchen von *VRUs* ergeben, im Rahmen des Forschungsdiskurses entwickelt wurden, konzipieren risikobezogene Verteilungsfragen für Unfallalgorithmen auf der Basis eines Rawls'schen Prioritarismus. Vor allem in der implementierungsnahen risikoethischen Literatur findet ein von Rawls inspiriertes Design von Unfallalgorithmen Zustimmung. Übertragen auf den Kontext von Unfalldilemmata impliziert das *Differenzprinzip* spezifische Fahrentscheidungen: »[…] it will result in driving decisions in which the worst-off vehicles profit the most and it is prevented that high risk situations for certain vehicles are accepted for the benefit of all other vehicles.« (Dietrich & Weisswange, 2019, S. 233) Zudem wird betont, dass eine Priorisierung von *VRUs* als derjenigen Gruppe von Verkehrsteilnehmern, welche hinsichtlich ihres Schadenserwartungswerts im Rawls'schen Sinne als die am schlechtesten Gestellten gelten können, sich positiv auf das Verhalten von autonomen Fahrzeugen gegenüber *VRUs* auswirkt:[288]

> It seems to be more appropriate, to always give full priority to the entity which has the highest expected harm. To formalize this, the severity of a potential encounter between two or more entities can be calculated by predicting the severity for each of the entities separately and then selecting the maximum value. In this way, the entity with highest risk has a

287 Siehe Arneson (1999) für eine Übersicht zu möglichen Begründungen egalitärer Gerechtigkeitskonzeptionen, die von bessergestellten Individuen verlangen, den schlechtergestellten zu helfen.

288 Dietrich und Weisswange (2019, S. 233) merken dazu an, dass eine Berücksichtigung der Interessen der Schwächsten auch die Wahrscheinlichkeit von Diskriminierung reduziert.

> strong effect on the risk-based behavior planning—the consequence is that cautious behaviors around vulnerable road users will become more likely. (Dietrich, 2021, S. 735)

Auf der formalen Ebene lässt sich das *Differenzprinzip* mithilfe der *Maximin*-Regel ausdrücken; das Ergebnis ist Pareto-optimal, d. h. es existiert keine andere Schadensverteilung, die ein Individuum besserstellt, ohne gleichzeitig mindestens ein anderes schlechterzustellen. Der prominente, von Leben (2017) entwickelte formale Algorithmus basiert auf einer Schätzung der Überlebenswahrscheinlichkeit für jedes involvierte Individuum bei jeder möglichen Trajektorie und der Auswahl derjenigen, bei der die Schwächsten mit der höchsten Wahrscheinlichkeit überleben. Anstelle der einfachen *Maximin*-Regel verwendet er eine iterierte Variante, das sogenannte *Leximin*-Prinzip.[289] Dieses unterscheidet sich vom *Maximin*-Prinzip dahingehend, dass für den Fall, dass zwei Alternativen zum gleichen Ergebnis führen, die Wahrscheinlichkeit desjenigen Individuums den Ausschlag gibt, das am zweitschlechtesten gestellt ist. So soll verhindert werden, dass die alleinige Orientierung an demjenigen Individuum, welches am schlechtesten gestellt ist, zur Wahl einer insgesamt schlechteren Verteilung führt.

Lebens Entwurf ist im Forschungsdiskurs kontrovers rezipiert worden. Eine prominente Kritik stammt von Keeling (2018a, S. 264–270), der einwendet, dass Lebens Ansatz auf der Basis Rawls'scher Gründe nur schwer zu rechtfertigen ist und einer unabhängigen Begründung bedarf. Er stellt u. a. klar, dass die Individuen bei Rawls nicht zwischen konkreten Güterverteilungen anhand des *Maximin*-Prinzips wählen, sondern zwischen Gerechtigkeitsprinzipien, wobei die Wahl aus rationalen Gründen auf die *Maximin*-Regel fällt. In Lebens Konzeption hingegen wählen die Individuen zwischen konkreten Verteilungen mithilfe des bereits akzeptierten *Maximin*-Prinzips. Zudem legt Leben den Individuen in der Ausgangsposition des *Schleiers des Nichtwissens* Restriktionen hinsichtlich der verfügbaren Informationen auf, die aus Rawls'scher Sicht nicht gerechtfertigt sind.

289 Das *Leximin*-Prinzip stammt aus der Wohlfahrtsökonomik und wurde von Amartya Sen (1976, 1980) in Anlehnung an das Rawls'sche *Differenzprinzip* entworfen, um eine etablierte Konvention innerhalb der Sozialwahltheorie zu beschreiben.

Ferner ist die an Rawls angelehnte Konzeption aufgrund der Spezifika des Anwendungskontextes in mancherlei pragmatischer Hinsicht ethisch problematisch. Für das Entscheidungsproblem, welches sich im Kontext von Unfallalgorithmen stellt, gilt, dass die Identität der betroffenen Parteien zum Zeitpunkt der Implementierung unbekannt ist. Gemäß Hevelke und Nida-Rümelin (2015c, S. 11–14) folgt aus dieser fehlenden Determiniertheit der Identität potenzieller Opfer, dass eine auf Schadensminderung ausgerichtete Programmierung nicht grundsätzlich unvereinbar mit einer deontologischen Ethik sein muss. Denn auch wenn Schadensminimierung letztlich immer ein Abwägen erfordert, muss dieses nicht zwangsläufig ethisch problematisch sein. Wenn die konkreten Vor- und Nachteile einer Person nicht bestimmbar sind, weil schon ihre Identität nicht bestimmbar ist, so lässt sich lediglich untersuchen, welche Handlung im Sinne aller wäre. Hevelke und Nida-Rümelin (2015c, S. 11) folgern daher, »dass eine auf die Minimierung der Opfer ausgelegte Programmierung durchaus im Interesse jedes Einzelnen sein kann – nämlich genau dann, wenn diese Programmierung das Risiko eines jeden Einzelnen reduziert bzw. minimiert.« Dadurch, dass es grundsätzlich jeden treffen könnte, würde eine Programmierung auf Schadensminimierung niemanden im kantischen Sinne bloß als Mittel gebrauchen. Analog konstatiert die Ethik-Kommission des BMVI in ihrem Bericht, dass eine solche Programmierung keinen Verstoß gegen Art. 1 Abs. 1 GG darstellen würde, sofern die Voraussetzung erfüllt ist, dass »die Programmierung das Risiko eines jeden einzelnen Verkehrsteilnehmers in gleichem Maße reduziert. Solange nämlich die vorherige Programmierung für alle die Risiken in gleicher Weise minimiert, war sie auch im Interesse der Geopferten, bevor sie situativ als solche identifizierbar waren.« (Di Fabio et al., 2017, S. 18)

Diese Konstellation wird in der Literatur häufig als heuristisch interpretierte Instanz des Rawls'schen *Schleier des Nichtwissens* diskutiert (siehe auch Kap. 4.4.4.2). Dadurch, dass zum Zeitpunkt des Designs von Unfallalgorithmen die Identität späterer Opfer unbekannt ist, werden die Individuen der Möglichkeit beraubt, sich selbst taktische Vorteile zu verschaffen, indem sie sozialschädliche Regeln installieren. So ist es unter der Annahme weitgehend egoistischer Individuen plausibel zu erwarten, dass diese sich für faire Verteilungsprinzipien entscheiden würden, um auf diese Weise den

Nutzen aller – und damit auch ihren eigenen – zu steigern. Wenn jede Person jede mögliche Rolle in einem Dilemma mit gleicher Wahrscheinlichkeit einnimmt, dann verhalten sich rationale Individuen risikoavers (vgl. Rath, 2011, S. 77) und entscheiden sich für eine Gleichverteilung von Risiken; ebendas ist im Kontext autonomen Fahrens jedoch fragwürdig. Auch wenn die Individuen nicht sicher wissen, wie mögliche Dilemmata für sie ausgehen würden, so haben sie doch Anhaltspunkte dafür, mit welcher Wahrscheinlichkeit sie z. B. die Position eines *VRU* oder eines Fahrzeuginsassen innehätten. Wer niemals Rad fährt, für den sind die Interessen von Radfahrern unerheblich, und es fehlen Anreize, eine faire Behandlung dieser Gruppe und damit eine egalitäre Gerechtigkeitskonzeption zu fordern. Sind Risiken dagegen ungleich verteilt, fehlen die Voraussetzungen dafür, dass eine Einigung auf eine entsprechende Konstellation erzielt werden würde; eine Ungleichverteilung wäre genau dann nicht im Interesse aller, wenn dadurch die Sicherheit einer Gruppe von Personen systematisch priorisiert würde (vgl. Hevelke & Nida-Rümelin, 2015c, S. 13–14), unabhängig davon, ob dies moralisch zulässig ist oder nicht. Aus (straf-)rechtlicher Sicht ist die Anonymität möglicher Opfer zum Tatzeitpunkt unerheblich für die juristische Bewertung:

> Moralisch und rechtlich gesehen bedeutet es nach den bisher akzeptierten Standards für die Unrechtsbewertung einer Tat aber keinen wesentlichen Unterschied, ob man das Opfer schon persönlich identifiziert hat oder ob die Identität des Opfers vom Zufall abhängt bzw. von Umständen, die z. Z. der Tötungshandlung noch nicht bekannt waren. […] Das Unrecht der Tat liegt in der Opferung des Menschen (als solchem), auf irgendwelche Identitätsmerkmale kommt es nicht an. (Hilgendorf, 2018a, S. 693)

Mithilfe eines modifizierten *Schleier des Nichtwissens* lassen sich berechtige Ansprüche also nicht begründen. Zudem haften dem Prioritarismus in seinem Wesen als utilitaristische Variante in gewisser Hinsicht auch die Schwächen des Standardutilitarismus an. So ist ein prioritaristischer Ansatz nur anwendbar auf Fälle, in denen sich die ethisch relevanten Merkmale in Form von gerechtigkeitsethischen Abwägungen zwischen Schlechter- und Bessergestellten ausdrücken lassen. Hingegen können Fälle nicht erfasst werden, in denen sich Individuen willentlich selbst einem Risiko aussetzen, dieses deshalb aber noch nicht auf eine andere Person übertragen dürfen; der Prio-

ritarismus geht fälschlicherweise davon aus, dass Risiken interpersonell austauschbar sind:

> Just like utilitarianism, prioritarianism treats risks as entities that can be unproblematically transferred between persons, albeit with a change in magnitude if the persons in question are not equally well-off. Furthermore, like utilitarianism, it always allows us to justify a disadvantage affecting one person by a sufficiently large advantage to some other person. (Hansson, 2003, S. 297)

7.3.3.4 Sven Ove Hansson: Gerechtigkeit durch Reziprozität

Hansson (2003) entwirft im Rahmen seines deontologisch-risikoethischen Ansatzes, der auf der Grundlage individueller Rechte und Pflichten fußt, eine alternative Möglichkeit, berechtigte Vorteile für *VRUs* zu begründen. Er entwickelt sein Kriterium der Risikoakzeptabilität ausgehend von der Prämisse, dass das Recht, keinem Risiko ausgesetzt zu werden, als ein *Prima-Facie*-Recht zu verstehen ist, das absoluten Geltungsansprüchen individueller Rechte eine Absage erteilt. Dieses steht grundsätzlich jedem Individuum zu; jedoch können Risikoübertragungen und damit einhergehende Verletzungen dieses Rechts aus lebensweltlich-pragmatischen Gründen erlaubt sein.[290] Diesen Umstand beschreibt er als das sogenannte »exemption problem«:

> It is a prima facie moral right not to be exposed to risk of negative impact, such as damage to one's health or one's property, through the actions of others. What are the conditions under which this right is overridden, so that someone is allowed to expose other persons to risk? (Ebd., S. 303)

Unter welchen Bedingungen darf ein Recht nun überschrieben und eine Person einem Risiko ausgesetzt werden? Hanssons zentrale These lautet: Das *Prima-Facie*-Recht, keinen Risiken exponiert zu

290 Es ist anzumerken, dass Hanssons Verwendung der Bezeichnung *prima facie* von derjenigen abweicht, wie sie in Kap. 5.3.1 skizziert wurde. Die risikoethische Verwendungsweise, die Hansson impliziert, beruht auf der Einsicht, dass Individualrechte zwar absolut gelten, Risiken, die diese zu verletzen drohen, sich aber nicht zwangsläufig manifestieren müssen. Risikoexpositionen sind damit noch nicht als Rechtsverletzungen zu werten und tangieren den Status von Individualrechten nicht in unzulässiger Weise.

werden, darf genau dann überschrieben werden, wenn die Risikoübertragung Teil eines gerechten Systems von Risiken ist, welches dem Einzelnen Vorteile bringt.

Schrittweise entwickelt Hansson ein soziales System zur reziproken Verteilung von Vor- und Nachteilen aus Risikoübertragungen, das allen zum wechselseitigen Vorteil gereicht: »Exposure of a person to a risk is acceptable if and only if this exposure is part of an equitable social system of risk-taking that works to her advantage.« (Ebd., S. 305) Dabei installiert er die Reziprozität von Risikoübertragungen als zentrales Element distributiver Gerechtigkeit, indem er betont, dass bei der Frage nach der Zulässigkeit einer Risikoübertragung zwei zentrale Komponenten zusammenwirken: auf der einen Seite deontologisch begründete Individualrechte, auf der anderen ein reziprokes soziales System zur Verteilung von risikoinduzierten Vor- und Nachteilen. Letzteres kann beispielsweise dadurch gewährleistet sein, dass Vor- und Nachteile zwischen den Betroffenen ausgetauscht werden, sofern sich diese Verteilung insgesamt betrachtet als für alle vorteilhaft erweist: »Each of us takes risks in order to obtain benefits for ourselves. It is beneficial for all of us to extend this practice to mutual exchanges of risks and benefits.« (2007b, S. 31) Hansson zufolge ist es für die Problemstellung zwar unerheblich, ob ein überschriebenes Recht vollständig eliminiert wird, nicht aber für die praktischen Konsequenzen, die sich daraus ergeben. Bleiben alle Rechte in ihrer Geltungskraft erhalten, so verbleiben wiederum Pflichten für den Handelnden, die z. B. in Form von Kompensationen oder verstärkten Maßnahmen zur Risikosenkung bestehen können (vgl. 2003, S. 303).

Die von Hansson vorgeschlagene soziale Praxis des wechselseitigen Austausches von Vor- und Nachteilen erhält ihre Legitimation dadurch, dass auf diese Weise die Handlungsmöglichkeiten aller ausgeweitet werden. Die Einführung einer Technologie ist also zulässig, wenn insgesamt gesehen jeder von ihr profitiert, auch wenn im konkreten Einzelfall gegebenenfalls Nachteile in Kauf genommen werden müssen. Eine Rechtsverletzung ist nur dann zulässig, wenn den Betroffenen durch Teilhabe am gerechten, sozialen System grundsätzlich dessen Vorteile offenstehen. Genau darin besteht der kritische Punkt von Hanssons Ansatz: Insbesondere im Kontext von Technologien bleibt fraglich, ob wechselseitige Vorteile immer gegeben sind. Innovative, auf Endkunden bezogene Technologien

sind bei der Markteinführung häufig mit hohen Kosten verbunden, die bestimmte Bevölkerungsgruppen zumindest zunächst von deren Nutzung und damit auch Vorteilen ausschließen, nicht aber ihren Risiken. Daraus müsste im Sinne von Hanssons Zulässigkeitskriterium folgen, dass Risikoübertragungen auf diese Personen nicht legitimiert wären, was wiederum die Rechtfertigbarkeit der Technologie an sich in Frage stellen würde.

Als Einwand auf diese Schlussfolgerung lässt sich argumentieren, dass eine Technologie meist verschiedene Vorteile bietet. Auch wenn eine Person von den (direkten) Vorteilen des autonomen Fahrens ausgeschlossen wäre, welche sich u. a. in einem gesteigerten Fahrkomfort äußern, so könnte sie dennoch indirekt davon profitieren, z. B. von einer allgemein erhöhten Straßenverkehrssicherheit und verringerten Schadstoffemissionen. Müller und Gogoll (2020) erweitern den Horizont der Teilhabe an möglichen Vorteilen auf die lokale Güterversorgung entsprechend:

> The reason why we generally accept that other people expose us to risk by driving, we submit, is that there is a wide consensus that our traffic system works to the advantage of everybody. Everybody who uses vehicles as a means of transportation to get from one point to another benefits by it. However, it is not only people who drive that benefit from traffic but also—for instance—people who, even though they never drive, rely on their local supermarket to be fully stocked with all sorts of products that would never be there if it was not for a (international) system of traffic that relied on some mode of transportation that is associated with a tiny risk of hurting people. (Ebd., S. 1554)

Die zentrale Frage ist hierbei, ob das Risiko, dem diese Personen ausgesetzt sind, durch indirekte Vorteile gerechtfertigt werden kann. Zur Feststellung der Gleichwertigkeit notwendige Bewertungsskalen sind angesichts der Tatsache, dass sich der Wert eines menschlichen Lebens und seine Facetten grundsätzlich nicht ethisch vertretbar quantifizieren lassen, problematisch. Darüber hinaus können ideologische Gründe Einzelne dazu bewegen, eine Technologie nicht zu nutzen. Ob ein solcher bewusster Verzicht auf die Vorteile jedoch als Zustimmung zu entsprechenden Risikoübertragungen gewertet werden kann, ist fraglich. Das risikoethische Konzept der Zustimmung ist als Kriterium für zulässige Risikoübertragungen nur begrenzt sinnvoll und wird überdies kontrovers beurteilt, denn Handlungen können auch dann als ethisch fragwürdig gelten, wenn Betroffene

zugestimmt haben. In der Konsequenz erscheint Hanssons Ansatz zwar grundsätzlich plausibel, aber unvollständig. Er greift risikoethisch relevante gerechtigkeitsethische Grundfragen auf, spezifiziert seinen Entwurf aber nicht konkret genug, um für eine direkte Anwendung zu taugen; die Integration anderer risikoethischer Konzepte ist notwendig.

Aufgrund des reziproken Designs kann es dazu kommen, dass Unterschiede hinsichtlich erlangter Vorteile nicht direkt ausgeglichen werden können, sondern beispielsweise über Kompensationen (vgl. McCarthy, 1997, S. 219–220) entschädigt werden müssen. Ein Nachteil in einer Situation wäre dann durch einen Vorteil in einer anderen abgegolten. Was für den Fall nicht-dilemmatischer, alltäglicher Fahrsituationen noch legitim sein mag, ist es für Dilemma-Situationen hingegen nicht. In Letzteren können mögliche Nachteile so gravierend sein, dass spätere Vorteile irreversible Schäden nicht mehr kompensieren können. Zudem sind Konzeptionen einer möglichen Kompensation aufgrund fehlender Zustimmung generell risikoethisch fragwürdig; aus praktischer Sicht kommen auch hohe Implementierungskosten hinzu (vgl. Dietrich & Weisswange, 2019, S. 233). So ist es beispielsweise sinnvoll, ein bestimmtes Maß an grundsätzlich erforderlicher Risikotoleranz zugrunde zu legen, um das System reziproker Vor- und Nachteile dahingehend zu entlasten, bei jedem noch so geringen Risiko für Vorteilsausgleich sorgen zu müssen. Um zu verhindern, dass Einzelnen sehr große Nachteile aufgebürdet werden, bietet es sich an, eine Zumutbarkeitsgrenze festzulegen, wie sie in Kap. 7.3.2 vor dem Hintergrund eines moralischen Sorgfaltsbegriffs diskutiert wurde. Auf diese Weise lassen sich die beiden Grenzkriterien integrieren und als deontologische Rahmenbedingungen eines Optimierungsproblems formulieren, das zumutbare und faire Risiken in den Fokus rückt.

8. Fazit und Ausblick

8.1 Ergebnisse der philosophischen Untersuchung: Zusammenfassung

Das autonome Fahren stellt eines der faszinierendsten technologischen Entwicklungsprojekte unserer Zeit dar. Als disruptive Technologie ist es Schauplatz vielfältiger politischer, rechtlicher und nicht zuletzt ethischer Debatten, die miteinander in Beziehung gesetzt werden müssen, um ein umfängliches Verständnis seines Wesens als gesellschaftliches Phänomen zu erlangen.

In dieser Forschungsarbeit wurde das Problem der ethischen Gestaltung von Unfallalgorithmen als eines der zentralen Puzzleteile auf dem Weg zur anvisierten Mobilitätswende untersucht. Die vorliegende Auseinandersetzung fokussierte spezifische moralische Dilemma-Strukturen, die Entscheidungssituationen angesichts unabwendbarer Unfälle kennzeichnen. Sie lässt sich dabei im weiteren Kontext eines sich neu formierenden, risikofokussierten Forschungszugangs verorten. Die erarbeiteten Ergebnisse wurden entlang von Teilzielen, Arbeitshypothesen und Zwischenergebnissen entfaltet und strukturiert, was die Nachvollziehbarkeit der Argumentation erleichtern soll. An dieser Stelle werden abschließend die gedanklichen Schritte und zentralen Ergebnisse des in diesem Buch entwickelten argumentativen Narrativs in wertender Form zusammengetragen.

Kap. 1 widmete sich einer Einführung in die Untersuchung, in deren Rahmen zunächst das autonome Fahren als sozio-technisches Phänomen skizziert und hinsichtlich seiner ethischen Dimension problematisiert wurde. In diesem Zuge wurde die ethische Gestaltung von Unfallalgorithmen als das zentrale Forschungsanliegen des Buches motiviert, das entsprechende Erkenntnisinteresse prägnant formuliert und dessen Relevanz in theoretischer und praktischer Hinsicht aufgezeigt. Als zentraler Forschungsauftrag wurde die Entwicklung eines alternativen Problemzugangs herausgestellt, der relevante Dilemma-Szenarien als risikoethische Entscheidungs-

probleme interpretiert. Der methodische Ansatz der Forschungsarbeit wurde als interdisziplinär ausgerichtete, angewandte ethische Untersuchung charakterisiert, die Methoden und Konzepte aus der Maschinenethik, Digitalen Ethik und Ethik der Künstlichen Intelligenz einerseits sowie der Risikoethik andererseits integriert. Auf eine präzise Darstellung der forschungsleitenden Thesen und Ziele der philosophischen Untersuchung folgte schließlich eine Skizze des argumentativen Gedankengangs, welcher dem Narrativ dieses Buches zugrunde liegt.

Teil I begann mit einer einführenden Darstellung des Forschungsgegenstands. In Kap. 2 wurde ein Bild des autonomen Fahrens in seinem visionären Charakter und seinen vielfältigen Herausforderungen sowie der Verortung spezifischer ethischer Fragen im Forschungsdiskurs gezeichnet. Hierbei wurden zunächst die Grundlagen zum Verständnis moralischer Dilemma-Strukturen im gewählten real-lebensweltlichen Kontext gelegt. Die Senkung der sozialen Kosten motorisierter Individualmobilität wurde als übergeordnete Zielsetzung in der Entwicklung autonomer Fahrsysteme erläutert. Den primären Motivator stellt ein signifikant erhöhtes Sicherheitspotenzial dar, das im Zuge der Eliminierung menschlicher Fahrfehler zu erwarten ist. Anhand des Stufenmodells des autonomen Fahrens, das als zentrale Richtschnur für Forschung, Entwicklung und Regulierung gilt, konnten die Herausforderungen herausgearbeitet werden, mit denen sich das praktische Projekt der Verkehrsautomatisierung konfrontiert sieht.

Wie eine nähere Analyse zeigte, prägen Wechselwirkungen zwischen dem technischen Reifegrad der Fahrzeugsysteme bzw. deren Komponenten, ihrer Wirtschaftlichkeit und dem jeweils geltenden Rechtsrahmen das Voranschreiten der Mobilitätsrevolution maßgeblich. Einen Meilenstein in der Schaffung eines innovationsfördernden nationalen Rechtsrahmens markiert das 2021 beschlossene »Gesetz zum autonomen Fahren«. Nichtsdestotrotz sehen sich regulierende Institutionen der Notwendigkeit gegenüber, zwischen Sicherheits- und Entwicklungszielen zu vermitteln: Gelockerte Restriktionen dürfen nicht zulasten der Sicherheit gehen. Der kürzlich in Kraft getretene europäische *AI Act* stellt einen bedeutenden Schritt in diese Richtung dar, indem er Automobilunternehmen in die Pflicht nimmt, sich proaktiv mit möglichen negativen Auswirkungen der von ihnen entwickelten Hochrisiko-KI-Systeme auseinanderzu-

setzen. Vor diesem Hintergrund ist das autonome Fahren prinzipiell als Initiator ambivalenter Wirkungen anzuerkennen. Eine kritische Reflexion auf die assoziierten Erwartungen im Sinne eines ›digitalen Heilsversprechens‹ legte offen, dass assoziierte positive Effekte von ernstzunehmenden negativen Externalitäten begleitet werden.

In einem zweiten Schritt wurden in Kap. 3 ethische Problemfelder und die entsprechenden Diskurse überblicksartig skizziert. Dabei rückten angewandte ethische Fragen in den Vordergrund, die im Kontext des Sicherheitsversprechens autonomer Fahrzeuge stehen. Neben kursorischen Einblicken in den relevanten Verantwortungsdiskurs konnte die Problematik unvermeidbarer Unfallsituationen sowohl in ihrer ethischen als auch informationstechnischen Dimension eruiert werden. Die Begrenztheit des Antizipationspotenzials autonomer Systeme, ihre generelle Fehleranfälligkeit sowie die Nicht-Eliminierbarkeit spezifischer Gefahrenpotenziale eines dynamischen Verkehrsgeschehens sind die entscheidenden Faktoren, aufgrund derer sich folgenreiche Unfälle nicht vollständig ausschließen lassen. Als Spezialfall beschränkt antizipierbarer Szenarien wurden Situationen evaluiert, in denen Personenschäden unabhängig von gewählter Trajektorie und Bremsverhalten des betreffenden Fahrzeugs unabwendbar sind. Derartige Szenarien weisen dilemmatische Strukturen auf, weshalb man sie im ethischen Diskurs als Instanzen moralischer Dilemmata auffasst: Ist eine Kollision unausweichlich, gibt es keine eindeutige ›richtige‹ Entscheidung.

Ferner wurde moralischen Dilemma-Szenarien sowohl in theoretischer als auch in praktischer Hinsicht eine essenzielle Rolle für die Realisierung einer automatisierten und vernetzten Mobilität attestiert. Eines der zentralen Argumente lautete dabei, dass die dilemmatischen Strukturen, die dem Entscheidungsproblem zugrunde liegen, sich tatsächlich in vielen alltäglichen Fahrsituationen stellen – oftmals ohne dass es uns bewusst ist. Das vor allem in der Tagespresse gezeichnete Bild tragischer Dilemma-Entscheidungen über Leben und Tod pointiert lediglich ein Problem, das omnipräsent ist: Überholvorgänge, Abstand zum Vordermann, gewählte Fahrgeschwindigkeit – in vielen Situationen sind ethisch relevante Abwägungen zwischen relevanten Handlungsgründen bzw. den Interessen Einzelner implizit enthalten. Die Komplexität, Vielfalt und Dynamik denkbarer Szenarien wurden als Gründe angeführt, weshalb sich Unfalldilemmata nicht anhand von Heuristiken normieren lassen;

sie sind »nicht ethisch zweifelsfrei programmierbar« (Di Fabio et al., 2017, S. 11) und müssen als Einzelfälle ethisch gewürdigt werden. Wie ausgewählte Literatur zeigte, kommt Unfalldilemmata in gesellschaftlicher Hinsicht insofern eine hohe Bedeutung zu, als sie eine große psychologische Rolle für potenzielle Nutzer spielen und auf diese Weise zu den erfolgskritischen Determinanten der Marktdurchdringung autonomer Fahrzeuge zählen. Schlussendlich erfordert die technische Robustheit der Entscheidungslogik von Fahrsystemen, dass für alle denkbaren Fälle jederzeit definierte Zustände und entsprechende Handlungsvorgaben abrufbar sind.

In Teil II wurde das zentrale Argument entwickelt, welches die erste forschungsleitende These begründet: Ein Zugang zur Gestaltung von Unfallalgorithmen, wie ihn der einschlägige Forschungsdiskurs bisher diskutiert, lässt (zu) viele Fragen offen. In Kap. 4 erfolgte eine kritische Auseinandersetzung mit bis dato dominanten Forschungszugängen, unter denen moralische Dilemma-Situationen als Problematik moralischer Designentscheidungen interpretiert werden. Auf der Basis maschinenethischer Grundlagen wurde argumentiert, dass durch Algorithmen getroffene Entscheidungen sich in moralisch relevanter Weise von menschlichen Entscheidungen unterscheiden. Im Zuge ihrer moralphilosophischen Problematisierung konnte eine systematisierte Zusammenstellung repräsentativer Dilemma-Szenarien erarbeitet werden, auf welche die Argumentation im weiteren Verlauf der Forschungsarbeit immer wieder zurückgriff.

Ausgehend von einer kritischen Darstellung bisheriger Forschungsliteratur wurde dem im Diskurs weit verbreiteten Narrativ einer möglichen Analogie zum bekannten Trolley-Problem nachgespürt, wobei essenzielle Diskrepanzen offengelegt werden konnten. Eine eingehende Untersuchung des praktischen Kontextes, in den real-lebensweltliche Unfalldilemmata eingebettet sind, lieferte wesentliche Anhaltspunkte für eine mangelnde Eignung bisher vorgeschlagener Ansätze. Dilemma-Szenarien sind – anders als durch die vermeintliche Trolley-Analogie suggeriert – keine individuellen Entscheidungsprobleme, sondern vielmehr soziale Problemstellungen, die in einem Netzwerk gesellschaftlicher Verflechtungen stehen. Es wurde dargelegt, dass Unfallalgorithmen als Gegenstand politischer Regulierung in einem Spannungsverhältnis zwischen individuellen Präferenzen einerseits und pluralistischen Wertvorstellungen ande-

rerseits verhaftet sind. Aus konzeptionellen und praktischen Gründen ließ sich die Option personalisierter ethischer Einstellungen für Unfallalgorithmen zurückweisen. Weiterhin wurde das Verhältnis von Ethik und Politik im Hinblick auf die Gestaltung von Unfallalgorithmen ausgelotet. Die dargestellte Argumentation bekräftigte, dass ethische Begründbarkeit und gesellschaftliche Akzeptanz im Sinne einer öffentlichen Vernunft nach Rawls'scher Lesart Hand in Hand gehen.

Auf der Grundlage dieser Ergebnisse demonstrierte eine umfassende Rekonstruktion der relevanten Forschungsliteratur, dass bisher dominante Forschungszugänge sowohl in methodischer als auch inhaltlicher und struktureller Hinsicht bedeutende Schwächen offenbaren. Anhand einer systematischen Analyse konnte erörtert werden, dass es weder deskriptiven noch normativen Herangehensweisen bisher gelungen ist, überzeugende Entscheidungsstrategien für Unfalldilemmata bereitzustellen. Utilitaristische, deontologische, tugendethische, vertragstheoretische, rechtsphilosophische, meta-normative und pluralistische Ansätze stoßen ebenso wie Konzepte und Methoden experimenteller Ethik diesbezüglich an ihre Grenzen. Als eine der essenziellsten Fehlinterpretationen, die dominanten Forschungszugängen aufgrund ihrer Fokussierung auf das Trolley-Problem zugrunde liegt, wurde die Annahme sicherer Handlungskonsequenzen bzw. des Eintretens bestimmter Umweltzustände akzentuiert. Es wurde betont, dass Entscheidungen in Bezug auf Unfalldilemmata tatsächlich mit Unsicherheiten hinsichtlich der zu erwartenden Folgen behaftet sind. Die Verfügbarkeit probabilistischer Bewertungsmethoden und stochastischer Entscheidungsmodelle erlaubt es, Eintrittswahrscheinlichkeiten grob zu bestimmen, weshalb man von Entscheidungen unter Risiko ausgeht. Dies motiviert die Entwicklung eines alternativen, spezifisch risikobasierten Zugangs, wie er in Teil III entwickelt wurde.

In Kap. 5 wurde das zentrale Argument, welches die erste These stützt, um eine metaethische Perspektive erweitert. Eine metaethische Rekonstruktion der Wertekonflikte, welche die relevanten moralischen Dilemma-Strukturen begründen, ermöglichte es, eine ganzheitliche Perspektive auf die Problemstellung zu eröffnen. Als zentrales Argument in der Begründung der Existenz unlösbarer Dilemmata wurde die Idee der Inkommensurabilität spezifischer moralischer Werte bzw. Gebote erörtert und anhand zweier Konzeptio-

nen veranschaulicht. Der Ansatz des unausweichlichen moralischen Fehlverhaltens nach Christopher Gowans fußt auf der Inkonvertibilität bzw. Nicht-Einlösbarkeit von Werten, die aufgrund einer einzigartigen Verantwortung gegenüber Personen besteht. Im Fall einer Verletzung derselben kommt es zu unersetzbaren Verlusten, die, so folgert Lisa Tessman im Zuge ihrer Weiterentwicklung des Konzepts von Gowans unter Bezugnahme auf Martha Nussbaums *Capability Approach*, nur dann moralisch relevant sind, wenn sie Kosten auferlegen würden, die nicht kompensiert werden können. Können Werte weder substituiert noch kompensiert werden, muss der Akteur zwangsläufig moralisch scheitern, die entsprechenden moralischen Gebote sind nicht verhandelbar. Es folgte eine Diskussion der Ergebnisse der metaethischen Analyse im Hinblick auf ihre Relevanz für den Anwendungskontext moralischer Unfalldilemmata. Die Argumentation machte deutlich, dass das zentrale Problem, welches die Komplexität moralischer Dilemma-Situationen im Kontext des autonomen Fahrens theoretisch-formal begründet, in der inkompatiblen Verschränkung moralisch gleichrangiger legitimer Interessen von Individuen und der gleichzeitigen Nicht-Verrechenbarkeit deontologischer Pflichten zum Schutz des Lebens besteht. Das entscheidende Argument bildet dabei die Inkommensurabilität unersetzbarer und einzigartiger Werte, auf die das Fehlen systematischer Herangehensweisen an derartige Dilemmata zurückzuführen ist.

Sodann folgte eine Untersuchung konkreter Entscheidungsstrategien, die auf metaethischer Basis argumentieren. Mit dem Verweis auf verantwortungsethische Gründe sowie eine fragwürdige Haltung der Indifferenz gegenüber Leben und Rechten der Betroffenen erfolgte eine Zurückweisung des Vorschlags, Dilemma-Szenarien per Zufallsprinzip zu entscheiden. Im Anschluss an Thomas Nagel wurde ein alternativer Ansatz dargelegt, der einen vielversprechenden Ausweg skizziert und dabei auf der Grundannahme basiert, dass sich, wenn formale Lösungen nicht möglich sind, auf praktischer Ebene dennoch Perspektiven eröffnen können. Daraufhin wurden Entscheidungsstrategien für echte moralische Dilemmata, die sich an der Maßgabe einer pragmatischen Ethik orientieren, zum Desiderat erklärt. Diese abstrahieren von theoretischen Analysen und fokussieren praktisch zielführende Abwägungen auf der Basis legitimierter Werte. Auf diese Weise ließ sich das in Kap. 4 entwickelte Argument dahingehend vervollständigen, dass sich aufgrund der

Konflikte inkommensurabler Werte eine pragmatische Ethik als vielversprechende Bewältigungsstrategie anbietet. Die Sachbezogenheit auf die Spezifika von Dilemma-Szenarien ist der entscheidende Blickwinkel, von dem ausgehend offene Fragen im Sinne eines pragmatischen Ansatzes geklärt werden können. Damit liefern die Ergebnisse dieses Kapitels kein eigenes Argument im engeren Sinne, sondern erlauben als argumentative Stütze einen Brückenschlag zwischen der ersten und zweiten These des Buches: Im Hinblick auf eine explizite ethische Auseinandersetzung mit dem Unsicherheitsaspekt, der der Gestaltung von Unfallalgorithmen anhaftet und in dieser Forschungsarbeit im Fokus steht, erscheint eine risikoethische Perspektive vielversprechend.

In Teil III wurde schließlich eine entsprechende risikoethische Interpretation der Problematik von Unfallalgorithmen präsentiert. Diese orientierte sich an der zweiten These, welche besagt, dass sich unter einem risikoethischen Zugang zentrale Fragen des Anwendungsproblems klären lassen und normative Implikationen freigelegt werden können, die neuen Entscheidungsperspektiven den Weg bereiten. In Kap. 6 war es zunächst das Anliegen, begriffliche und konzeptionelle Grundlagen in Bezug auf wissenschaftliche Verortung, Historie, Gegenstandsbereich und Paradigmen der Risikoethik zu legen. Es folgte eine Veranschaulichung möglicher Begründungsansätze hinsichtlich einer generellen moralischen Zulässigkeit von Risikoübertragungen anhand dreier risikopraktischer Paradigmen (konsequentialistisch-objektivistisch, postmodern-subjektivistisch, partizipatorisch). Im Kontext einer Kritik rationaler Risikopraxis wurden gängige entscheidungstheoretische Kriterien für fragwürdig befunden, da diese aufgrund ihrer konsequentialistischen Grundlage weitgehend unvereinbar sind mit dem Selbstverständnis einer Risikoethik, welche die Legitimität eines risikobehafteten Systems stets an deren Wirkungen auf Einzelinteressen bemisst. Das Kapitel mündete in der Forderung nach einem Paradigmenwechsel in risikopraktischen Fragen, der mit der Abkehr von einer konsequentialistischen Grundorientierung einhergeht. Angesichts der konfligierenden inkommensurablen Werte, die Unfalldilemmata kennzeichnen, lässt sich im Hinblick auf die Gestaltung von Unfallalgorithmen für die Hinwendung zu einer deontologischen Risikoethik werben.

Grundzüge einer solchen konnten in Kap. 7 systematisch entwickelt werden. Übergeordnetes Ziel war es dabei, die spezifischen Umstände zu bestimmen, unter denen Risiken übertragen werden dürfen. In einem ersten Schritt ging es darum, eine Interpretation vorzulegen, welche die Gestaltung von Unfallalgorithmen als risikoethisches Verteilungsproblem betrachtet, das sich im Spannungsfeld zwischen soziologischer Risikoakzeptanz und ethischer Risikoakzeptabilität bewegt. Es fand eine Abgrenzung der Forschungsarbeit von bisheriger risikoethischer Forschung zur Thematik des autonomen Fahrens statt, die sich auf eine Analyse von Effekten distributiver Gerechtigkeit sowie Entwürfe implementierungsnaher, formaler Ansätze konzentriert. Schließlich wurde eine dezidiert ethische Perspektive systematisch mit dem Ziel erarbeitet, den ›blinden Fleck‹ des Diskurses auszumerzen. Um die entsprechende Forschungslücke zu schließen und den Diskurs voranzubringen, fand zunächst eine Neuinterpretation der bisherigen Leitfrage nach der moralphilosophischen Begründung von Entscheidungsprinzipien für Unfallalgorithmen statt. In diesem Zuge rückten die ethische Rechtfertigbarkeit und Verteilung risikoinduzierter Vor- und Nachteile in den Vordergrund. Neben konsequentialistischen wurden auch kontraktualistische Kriterien der Risikoakzeptabilität evaluiert und für lediglich begrenzt adäquat befunden.

Als Antwort auf deren festgestellte Schwächen erfolgte die Skizze einer deontologischen Risikoethik, die das Ziel verfolgt, einerseits Werte wie Freiheit, Autonomie und individuelle Rechte zu schützen, die durch Risikoübertragungen bedroht sind, aber andererseits auch nicht zu restriktiv zu sein, um eine vollständige Paralyse sozialen Miteinanders zu vermeiden. Hier wurde im Anschluss an den schwach deontologischen, vor dem Hintergrund politischer Fragen kontraktualistisch verfassten Ansatz von Julian Nida-Rümelin, Johann Schulenburg und Benjamin Rath die Zielvorgabe einer kohärenten Risikopraxis ausgerufen. Diese integriert konsequentialistische und deontologische Elemente entsprechend ihrer jeweiligen Rechtfertigungsgrundlage und besteht darauf, dass deontologische Kriterien bei der Implementierung von Algorithmen harte Grenzen (*hard constraints*) darstellen, die vor quantitativen Optimierungszielen Vorrang haben müssen.

Schließlich wurde ein Entwurf einer kohärenten Risikopraxis skizziert, der auf zwei zentralen Prinzipien beruht, die als Grenz-

kriterien nicht-verhandelbare Rahmenbedingungen konstituieren, innerhalb derer sich die Verteilung von Vor- und Nachteilen aus Risikoübertragungen als Optimierungsproblem bewegt. Als erstes und absolutes Prinzip wurde die Zumutbarkeit von Risiken begründet. Diese fokussiert das, was rational betrachtet akzeptiert werden sollte, und nicht, was tatsächlich sozial akzeptiert ist. Die entscheidenden Aspekte, aufgrund derer sich subjektive Risikopräferenzen in objektivierte Risikobewertungen transferieren lassen, bilden die Reziprozität in der Übertragung zumutbarer Risiken und der daraus resultierende Vorteil, keine absoluten Einschränkungen individueller Handlungsfreiheit und in der Verfolgung persönlicher Ziele hinnehmen zu müssen. Zur Bestimmung des zumutbaren Risikos in Unfalldilemmata wurden moralische Sorgfaltspflichten als Kernkriterium einer entsprechenden Schwellenwert-Konzeption vorgeschlagen. Das zentrale moralische Argument, das dabei zur Anwendung kommt, besteht wiederum in deren Reziprozität, durch die soziale bzw. sozio-technische Gefüge wie Mobilität erst tragfähig werden.

Als zweites und relatives Prinzip wurde das Postulat einer fairen Verteilung von Vor- und Nachteilen im Kontext von Risikoübertragungen formuliert. Bezugnehmend auf die Rawls'sche Idee der ›Gerechtigkeit als Fairness‹ ließen sich risikopraktische Strategien genau dann als fair bezeichnen, wenn die sich daraus ergebende Verteilung von Vor- und Nachteilen als fair gelten kann. Ferner wurde argumentiert, dass sich die spezifische Problematik von Unfallalgorithmen nur anhand von egalitaristischen Gerechtigkeitskonzeptionen adäquat fassen lässt. Auch eine Ungleichverteilung kann fair sein – nämlich genau dann, wenn berechtigte Ansprüche auf bestimmte Vorteile anderen gegenüber bestehen. Die Überprüfung dieser Hypothese identifizierte das Ausmaß, in dem eine Person Risiken für andere verursacht, als ethisch relevanten Faktor für die Fairness von Risikoübertragungen. Hierbei wurde erläutert, inwiefern sich auf der Basis einer Kategorisierung von Verkehrsbeteiligten entlang ihrer jeweiligen Vulnerabilität die Gewährung von Sicherheitsvorteilen für besonders gefährdete Gruppen rechtfertigen lässt. Erklärtes Ziel ist es dabei, bestehende kontingente Ungleichheiten im Sinne einer ergebnisegalitaristischen Position fair zu berücksichtigen.

Um den Rückgriff auf egalitäre Gerechtigkeitskonzeptionen im Hinblick auf das gewählte Anwendungsproblem zu rechtfertigen, wurden zwei prominente Konzeptionen rekonstruiert und kritisch

beleuchtet. Zum einen wurde unter Anwendung von John Rawls' *Differenzprinzip* ausgeführt, dass die Orientierung an denjenigen Personen, die am schlechtesten gestellt sind, höchste Priorität genießen sollte. Eine kritische Diskussion deutete an, dass eine heuristisch interpretierte Instanz des Rawls'schen *Schleier des Nichtwissens*, wie sie im Diskurs häufig verwendet wird, berechtigte Ansprüche in Bezug auf Unfalldilemmata nicht final begründen kann. Zum anderen wurde ein Ansatz von Sven Ove Hansson vorgestellt, der Risikoübertragungen genau dann als zulässig betrachtet, wenn sie Teil eines gerechten Systems von Risiken sind, welches dem Einzelnen Vorteile bringt. Die Idee einer distributiven Gerechtigkeit wird hier durch die Reziprozität hergestellt, die jedem Einzelnen zum Vorteil gereicht. Abschließend blieb festzuhalten, dass beide Ansätze hinsichtlich ihrer Eignung, finale Begründungen fairer risikoethischer Verteilungsstrategien für Unfallalgorithmen zu liefern, unvollständig sind. Aufgrund ihrer Erweiterungsbedürftigkeit begründen sie weiterführende Forschungsdesiderate.

8.2 Kritische Reflexion und Ausblick: Wissenschaftliche Relevanz, Forschungsdesiderate und Limitationen

Im Rahmen der Forschungsarbeit wurde das anspruchsvolle Vorhaben verfolgt, Perspektiven der Digitalen Ethik, Metaethik und Risikoethik in einer pragmatisch orientierten Interpretation des Anwendungsproblems zusammenzuführen. Ziel war es dabei, den Problemhorizont bisheriger dominanter Forschungszugänge über eine moralphilosophische Sichtweise hinaus zu erweitern und auf diese Weise Raum für die Berücksichtigung theoretisch-formaler Grundlagen einerseits und anwendungsspezifischer Determinanten andererseits zu schaffen. Der Auftrag war es, eine dezidiert ethische Betrachtungsweise zu erarbeiten, die insbesondere dem Risikoaspekt moralischer Unfalldilemmata Rechnung trägt.

Der Beitrag, den dieses Buch zur Weiterentwicklung des Forschungsdiskurses leistet, besteht zu wesentlichen Teilen in der Begründung der Relevanz und Eignung einer risikoethischen Neuinterpretation des Anwendungsproblems. So wurde nicht nur erläutert, über welche Potenziale die Risikoethik bei der Gestaltung von Unfallalgorithmen verfügt, sondern es wurden auch konzep-

tionelle Grundlagen gelegt, auf die weiterführende Forschung aufbauen kann. Dabei ging es ausdrücklich nicht darum, bisherige Forschungsansätze grundsätzlich zurückzuweisen, sondern lediglich zu pointieren, inwiefern diese durch die Integration einer risikoethischen Perspektive profitieren können. Zudem bleiben die erarbeiteten Ergebnisse nicht auf Dilemma-Situationen beschränkt, sondern lassen sich ebenfalls auf Routinefahrsituationen anwenden, in denen – von uns oftmals unbemerkt – implizite Abwägungen zwischen persönlichen Zielen und Werten stattfinden. Wenn etwa Fahrentscheidungen über eingehaltene Abstände bei alltäglichen Überholvorgängen getroffen werden, bilden sich situationsspezifische Risikokonstellationen, die von nicht zu vernachlässigender ethischer Relevanz sind.

Als kritische Reflexion auf die erreichten Ergebnisse bleibt festzuhalten, dass die Limitationen der vorgelegten Untersuchung zugleich als Forschungsdesiderate zu verstehen sind. Diese bestehen zum einen hinsichtlich inhärenter Schwächen traditioneller ethischer Denkschulen, die im Hinblick auf das Anwendungsproblem zutage treten. Es wurde argumentiert, dass ein deontologisches Framework grundsätzlich offen für die Integration diverser ethischer Prinzipien ist. Die Forschungsarbeit verzichtet auf eine konkrete Ausarbeitung dessen, wie sich das neu definierte risikoethische Optimierungsproblem jenseits der skizzierten deontologischen Grenzkriterien von Zumutbarkeit und Fairness inhaltlich gestalten lässt. In welchem Umfang können konsequentialistische Elemente Beachtung finden? Inwieweit ist eine Schadensminimierung aus kontraktualistischer Sicht über Ex-Ante-Kompensation legitimierbar? An relevanten Stellen wurden diese und ähnliche offene Fragen explizit als solche benannt, um zukünftiger Forschung Anknüpfungspunkte zu bieten und kreative Ansätze zu inspirieren. Spezifische Probleme ethischer Theorien existieren grundsätzlich unabhängig von einer risikoethischen Sichtweise; im Zuge ihrer Bewältigung im Kontext des Anwendungsproblems müssen jedoch Risiken und Unsicherheiten als entscheidungstheoretische Komponenten Berücksichtigung finden.

Überdies ist die Sphäre moralischer Verantwortung für Risikoübertragungen im Rahmen dieser Forschungsarbeit weitgehend unbearbeitet geblieben. Eine fundierte und explizite Auseinandersetzung mit diesem Themenfeld bedarf mehr als einer Randnotiz, was dieses Buch aufgrund seiner Fokussierung auf den konkreten

Anwendungsfall von Unfallalgorithmen sowie risikoethische Fragen der Zulässigkeit und Fairness nur unzureichend leisten kann. Hierbei sollten auch Perspektiven Erwähnung finden, die auf technikinduzierte Risiken im Speziellen rekurrieren. So wird es mit zunehmender Automatisierung technischer Innovationen immer schwieriger, mögliche Schadensfolgen zuverlässig abzuschätzen und verantwortlichen Instanzen zuzurechnen. Die Grenzen der Verantwortlichkeit verschwimmen angesichts kollektiver Entscheidungslagen sowohl im vorwärts- als auch im rückwärtsgerichteten Sinne oder müssen hinsichtlich der Verantwortungsunfähigkeit autonomer Maschinen neu ausgelotet werden.

Zum anderen stellt der vorgelegte Entwurf keine abgeschlossene risikoethische Untersuchung dar, sondern ist vielmehr als skizzenhafter Impuls für weitere risikoethische, inter- und transdisziplinäre Forschung zu verstehen. Forschungsdesiderate bestehen z. B. hinsichtlich der Risikobewertung bzw. Risikoabschätzung, die notwendig ist, um zu implementierbaren Konzeptionen zu gelangen. Die Definition konkreter handlungsrelevanter Prinzipien und Algorithmen für spezifische Dilemma-Szenarien liegt ebenfalls jenseits des definierten thematischen Rahmens dieser Forschungsarbeit. Diese schafft jedoch eine argumentative Grundlage, auf der Konzepte einer möglichen Implementierung aus ethischer Sicht gerechtfertigt oder zurückgewiesen werden können. Technische Umsetzbarkeit allein ist nicht ausreichend, um die Wahl bestimmter Algorithmen zu begründen. Es ist zu hoffen, dass die vorgelegte risikoethische Auseinandersetzung als Hilfestellung dienen und auf diese Weise zu einer rechtfertigbaren Entscheidungsfindung beitragen kann.

Literaturverzeichnis

Abdel-Aty, M. & Ding, S. (2024) »A matched case-control analysis of autonomous vs human-driven vehicle accidents«, *Nature communications*, Vol. 15, No. 1, S. 4931.

Abney, K. (2022) »The Ethics of Abuse and Unintended Consequences for Autonomous Vehicles«, in Jenkins, R., Černý, D. & Hríbek, T. (Hg.) *Autonomous Vehicle Ethics: The Trolley Problem and Beyond,* New York, Oxford University Press, S. 257–274.

Adnan, N., Md Nordin, S., bin Bahruddin, M. A. & Ali, M. (2018) »How trust can drive forward the user acceptance to the technology? In-vehicle technology for autonomous vehicle«, *Transportation Research Part A: Policy and Practice*, Vol. 118, S. 819–836.

Aischylos (2018 [Uraufführung ca. 458 v. Chr.]) *Die Orestie: Agamemnon. Choephoren. Eumeniden*, Stuttgart, Reclam.

Allen, C., Smit, I. & Wallach, W. (2005) »Artificial morality: Top-down, Bottom-up, and Hybrid Approaches«, *Ethics and Information Technology*, Vol. 7, No. 3, S. 149–155.

Altay, B. C., Boztas, A. E., Okumuş, A., Gul, M. & Çelik, E. (2023) »How Will Autonomous Vehicles Decide in Case of an Accident? An Interval Type-2 Fuzzy Best–Worst Method for Weighting the Criteria from Moral Values Point of View«, *Sustainability*, Vol. 15, No. 11, S. 8916.

Altenburg, S., Kienzler, H.-P. & Maur, A. auf der (2018) »Einführung von Automatisierungsfunktionen in der Pkw-Flotte: Auswirkungen auf Bestand und Sicherheit«, *Prognos AG*, 2018 [Online]. Verfügbar unter https://www.prognos.com/de/projekt/einfuehrung-von-automatisierungsfunktionen-der-pkw-flotte (Abgerufen am 09. Juli 2023).

Anderson, E. S. (2000) »Warum eigentlich Gleichheit?«, in Krebs, A. (Hg.) *Gleichheit oder Gerechtigkeit: Texte der neuen Egalitarismuskritik,* Frankfurt/Main, Suhrkamp, S. 117–171.

Anderson, J. M., Kalra, N., Stanley, K. D., Sorensen, P., Samaras, C. & Oluwatola, O. A. (2016) »Autonomous Vehicle Technology: A Guide for Policymakers«, *Santa Monica, CA: RAND Corporation*, 2016 [Online]. Verfügbar unter https://www.rand.org/pubs/research_reports/RR443-2.html (Abgerufen am 16. Juli 2023).

Anderson, M. & Anderson, S. L. (Hg.) (2011) *Machine Ethics*, Cambridge, Cambridge University Press.

Anderson, M., Anderson, S. L. & Armen, C. (2005) »Towards Machine Ethics: Implementing Two Action-Based Ethical Theories«, *Proceedings of the AAAI 2005 Fall Symposium on Machine Ethics*, S. 1–7.

Anderson, M., Anderson, S. L. & Armen, C. (2006) »MedEthEx: A Prototype Medical Ethics Advisor«, *Proceedings of the 18th conference on innovative applications of artificial intelligence, Boston, Massachusetts*, S. 1759–1765.

ARD-Tagesschau (2023) *Umstrittener Autopilot: Tesla gewinnt Prozess nach tödlichem Unfall* [Online]. Verfügbar unter https://www.tagesschau.de/wirtschaft/unternehmen/tesla-prozess-autonomes-fahren-100.html (Abgerufen am 03. November 2023).

Arfini, S., Spinelli, D. & Chiffi, D. (2022) »Ethics of Self-driving Cars: A Naturalistic Approach«, *Minds and Machines*, Vol. 32, No. 4, S. 717–734.

Arkin, R. C. (2010) »The Case for Ethical Autonomy in Unmanned Systems«, *Journal of Military Ethics*, Vol. 9, No. 4, S. 332–341.

Arkin, R. C. (2018) »Lethal Autonomous Systems and the Plight of the Non-combatant«, in Kiggins, R. (Hg.) *The Political Economy of Robots: Prospects for Prosperity and Peace in the Automated 21st Century,* Cham, Palgrave Macmillan; Springer International Publishing, S. 317–326.

Arneson, R. J. (1989) »Equality and Equal Opportunity for Welfare«, *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, Vol. 56, No. 1, S. 77–93.

Arneson, R. J. (1999) »Egalitarianism and Responsibility«, *The Journal of Ethics*, Vol. 3, No. 3, S. 225–247.

Asimov, I. (2004 [1942]) »Runaround«, in Asimov, I. (Hg.) *I, Robot,* New York, Bantam Books, S. 25–45.

AUTOSAR GbR (2024) »About AUTOSAR« [Online]. Verfügbar unter https://www.autosar.org/about (Abgerufen am 24. August 2024).

Awad, E., Anderson, M., Anderson, S. L. & Liao, B. (2020) »An approach for combining ethical principles with public opinion to guide public policy«, *Artificial Intelligence*, Vol. 287, S. 103349.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F. & Rahwan, I. (2018) »The Moral Machine experiment«, *Nature*, Vol. 563, No. 7729, S. 59–64.

Awad, E., Levine, S., Kleiman-Weiner, M., Dsouza, S., Tenenbaum, J. B., Shariff, A., Bonnefon, J.-F. & Rahwan, I. (2019) »Drivers are blamed more than their automated cars when both make mistakes«, *Nature Human Behaviour*, Vol. 4, S. 134–143.

Bagloee, S. A., Tavana, M., Asadi, M. & Oliver, T. (2016) »Autonomous vehicles: challenges, opportunities, and future implications for transportation policies«, *Journal of Modern Transportation*, Vol. 24, No. 4, S. 284–303.

Baker, S., Abbany, Z., Freund, A., Dillon, C. & Schmidt, F. (2018) »Can autonomous cars have a moral conscience? Views from DW's science desk«, *Deutsche Welle*, 26. Oktober [Online]. Verfügbar unter https://www.dw.com/en/can-autonomous-cars-have-a-moral-conscience-views-from-dws-science-desk/a-46056690 (Abgerufen am 16. April 2023).

Banse, G. (1996) »Herkunft und Anspruch der Risikoforschung«, in Banse, G. (Hg.) *Risikoforschung zwischen Disziplinarität und Interdisziplinarität: Von der Illusion der Sicherheit zum Umgang mit Unsicherheit,* Berlin, edition sigma, S. 15–72.

Bartneck, C., Lütge, C., Wagner, A. R. & Welsh, S. (2019) *Ethik in KI und Robotik*, München, Carl Hanser Verlag.

Beauchamp, T. L. & Childress, J. (1994) *Principles of Biomedical Ethics*, New Jersey, Oxford University Press.

Bechmann, G. (1997 [1993]) »Risiko als Schlüsselkategorie der Gesellschaftstheorie«, in Bechmann, G. (Hg.) *Risiko und Gesellschaft: Grundlagen und Ergebnisse interdisziplinärer Risikoforschung,* 2. Aufl., Opladen, Westdeutscher Verlag, S. 237–276.

Beck, D. (2024) »Autonome Autos steuern meist sicherer als Menschen«, *ARD-aktuell/tagesschau.de*, 22. Juni [Online]. Verfügbar unter https://www.tagesschau.de/wissen/technologie/studie-selbstfahrende-autos-100.html (Abgerufen am 12. August 2024).

Beck, U. (2016 [1986]) *Risikogesellschaft. Auf dem Weg in eine andere Moderne*, Berlin, Suhrkamp Verlag.

Becker, F. & Axhausen, K. W. (2017) »Literature review on surveys investigating the acceptance of automated vehicles«, *Transportation*, Vol. 44, No. 6, S. 1293–1306.

Behrends, J. & Basl, J. (2022) »Trolleys and Autonomous Vehicles: New Foundations for the Ethics of Machine Learning«, in Jenkins, R., Černý, D. & Hríbek, T. (Hg.) *Autonomous Vehicle Ethics: The Trolley Problem and Beyond,* New York, Oxford University Press, S. 58–79.

Beiker, S. A. (2012) »Legal Aspects of Autonomous Driving«, *Santa Clara Law Review*, Vol. 52, No. 4, S. 1145–1156.

Beiker, S. A. (2015) »Einführungsszenarien für höhergradig automatisierte Straßenfahrzeuge«, in Maurer, M., Gerdes, J. C., Lenz, B. & Winner, H. (Hg.) *Autonomes Fahren: Technische, rechtliche und gesellschaftliche Aspekte,* Berlin/Heidelberg, Springer Vieweg, S. 197–217.

Bendel, O. (2016) *Die Moral in der Maschine: Beiträge zu Roboter- und Maschinenethik*, Hannover, Heise Medien GmbH & Co. KG.

Bendel, O. (2018) »Überlegungen zur Disziplin der Maschinenethik«, *Aus Politik und Zeitgeschichte*, Vol. 68, 06–08, S. 34–38.

Bendel, O. (2019) »Wozu brauchen wir die Maschinenethik?«, in Bendel, O. (Hg.) *Handbuch Maschinenethik,* Wiesbaden, Springer VS, S. 13–32.

Bengler, K., Dietmayer, K., Färber, B., Maurer, M., Stiller, C. & Winner, H. (2014) *Die Zukunft der Fahrerassistenz: Ein Strategiepapier der Uni-DAS,* Open Access Repositorium der Universität Ulm und Technischen Hochschule Ulm.

Bennett, S. (2022) »Algorithms of Life and Death: A Utilitarian Approach to the Ethics of Self-Driving Cars«, in Jenkins, R., Černý, D. & Hríbek, T. (Hg.) *Autonomous Vehicle Ethics: The Trolley Problem and Beyond,* New York, Oxford University Press, S. 191–209.

Bentham, J. (2005 [1970]) *The Collected Works of Jeremy Bentham: An Introduction to the Principles of Morals and Legislation*, New York, Oxford University Press.

Bentham, J. (2007 [1789]) *Introduction to the Principles of Morals and Legislation*, Oxford, Clarendon Press.

Bergmann, L. T., Schlicht, L., Meixner, C., König, P., Pipa, G., Boshammer, S. & Stephan, A. (2018) »Autonomous Vehicles Require Socio-Political Acceptance-An Empirical and Philosophical Perspective on the Problem of Moral Decision Making«, *Frontiers in Behavioral Neuroscience*, Vol. 12, S. 1–12.

Berkey, B. (2022) »Autonomous Vehicles, Business Ethics, and Risk Distribution in Hybrid Traffic«, in Jenkins, R., Černý, D. & Hríbek, T. (Hg.) *Autonomous Vehicle Ethics: The Trolley Problem and Beyond,* New York, Oxford University Press, S. 210–228.

Beutnagel, W. (2018) »Softwarefehler führte zu Uber-Unfall«, *Automotive IT*, 8. Mai [Online]. Verfügbar unter https://www.automotiveit.eu/technology/softwarefehler-fuehrte-zu-uber-unfall-146.html (Abgerufen am 07. Oktober 2023).

Bhargava, V. & Kim, T. W. (2017) »Autonomous Vehicles and Moral Uncertainty«, in Lin, P., Jenkins, R. & Abney, K. (Hg.) *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence,* Oxford, Oxford University Press, S. 5–19.

Birnbacher, D. (1996) »Risiko und Sicherheit – philosophische Aspekte«, in Banse, G. (Hg.) *Risikoforschung zwischen Disziplinarität und Interdisziplinarität: Von der Illusion der Sicherheit zum Umgang mit Unsicherheit,* Berlin, edition sigma, S. 193–210.

Birnbacher, D. & Birnbacher, W. (2016) »Automatisiertes Fahren. Ethische Fragen an der Schnittstelle von Technik und Gesellschaft«, *Information Philosophie*, No. 4, S. 8–15.

Bjorndahl, A., London, A. J. & Zollman, K. J. (2017) »Kantian Decision Making Under Uncertainty: Dignity, Price, and Consistency«, *Philosophers' Imprint*, Vol. 17, No. 7, S. 1–22.

Blanco, M., Atwood, J., Russell, S., Trimble, T., McClafferty, J. & Perez, M. (2016) »Automated Vehicle Crash Rate Comparison Using Naturalistic Data«, *Virginia Tech Transportation Institute* [Online]. Verfügbar unter https://www.vtti.vt.edu/PDFs/Automated%20Vehicle%20Crash%20Rate%20Comparison%20Using%20Naturalistic%20Data_Final%20Report_20160107.pdf (Abgerufen am 04. Februar 2023).

Blyth, P.-L., Mladenovic, M. N., Nardi, B. A., Ekbia, H. R. & Su, N. M. (2016) »Expanding the Design Horizon for Self-Driving Vehicles: Distributing Benefits and Burdens«, *IEEE Technology and Society Magazine*, Vol. 35, No. 3, S. 44–49.

Bodenschatz, A., Uhl, M. & Walkowitz, G. (2021) »Autonomous systems in ethical dilemmas: Attitudes toward randomization«, *Computers in Human Behavior Reports*, Vol. 4, S. 100145.

Boeglin, J. (2015) »The costs of self-driving cars: reconciling freedom and privacy with tort liability in autonomous vehicle regulation«, *Yale Journal of Law & Technology*, Vol. 17, S. 171–203.

Bohle, H.-G. & Pohl, J. (2014) »Risikoforschung als Grenzwissenschaft«, *Nova Acta Leopoldina (NAL NF)*, Vol. 117, No. 397, S. 155–162.

Bonnefon, J.-F., Shariff, A. & Rahwan, I. (2015) »Autonomous Vehicles Need Experimental Ethics: Are We Ready for Utilitarian Cars?«, *arXiv preprint arXiv:1510.03346*, S. 1–15.

Bonnefon, J.-F., Shariff, A. & Rahwan, I. (2016) »The social dilemma of autonomous vehicles«, *Science*, Vol. 352, No. 6293, S. 1573–1576.

Bonnefon, J.-F., Shariff, A. & Rahwan, I. (2019) »The Trolley, The Bull Bar, and Why Engineers Should Care About The Ethics of Autonomous Cars«, *Proceedings of the IEEE*, Vol. 107, No. 3, S. 502–504.

Bonnefon, J.-F., Shariff, A. & Rahwan, I. (2020) »The moral psychology of AI and the ethical opt-out problem«, in Liao, S. M. (Hg.) *Ethics of Artificial Intelligence,* Oxford, Oxford University Press, S. 109–126.

Bonß, W. (1995). *Vom Risiko. Unsicherheit und Ungewißheit in der Moderne*, Hamburg, Hamburger Edition.

Bonß, W. (1996) »Die Rückkehr der Unsicherheit. Zur gesellschaftstheoretischen Bedeutung des Risikobegriffs«, in Banse, G. (Hg.) *Risikoforschung zwischen Disziplinarität und Interdisziplinarität: Von der Illusion der Sicherheit zum Umgang mit Unsicherheit,* Berlin, edition sigma, S. 165–184.

Bonß, W. (2013 [1997]) »Die gesellschaftliche Konstruktion von Sicherheit«, in Lippert, E., Wachtler, G. & Prüfert, A. (Hg.) *Sicherheit in der unsicheren Gesellschaft,* Springer, S. 21–41.

Borenstein, J., Herkert, J. & Miller, K. (2017) »Self-Driving Cars: Ethical Responsibilities of Design Engineers«, *IEEE Technology and Society Magazine*, Vol. 36, No. 2, S. 67–75.

Borenstein, J., Herkert, J. R. & Miller, K. W. (2019) »Self-Driving Cars and Engineering Ethics: The Need for a System Level Analysis«, *Science and Engineering Ethics*, Vol. 25, No. 2, S. 383–398.

Bosch Mobility Solutions (Hg.) (2021) *Automated Valet Parking: Der fahrerlose Parkservice* [Online]. Verfügbar unter https://www.bosch-mobility-solutions.com/de/loesungen/parken/automated-valet-parking/ (Abgerufen am 01. Januar 2022).

Bradshaw, J. M., Hoffman, R. R., Johnson, M. & Woods, D. D. (2013) »The Seven Deadly Myths of ›Autonomous Systems‹", *IEEE Intelligent Systems*, Vol. 28, No. 3, S. 54–61.

Bradshaw-Martin, Heather & Easton, C. (2014) »Autonomous or ›driverless‹ cars and disability: a legal and ethical analysis«, *European Journal of Current Legal Issues*, Vol. 20, No. 3.

Brändle, C. & Grunwald, A. (2019) »Autonomes Fahren aus Sicht der Maschinenethik«, in Bendel, O. (Hg.) *Handbuch Maschinenethik,* Wiesbaden, Springer VS, S. 281–300.

Brändle, C. & Schmidt, M. W. (2021) »Autonomous Driving and Public Reason: a Rawlsian Approach«, *Philosophy & Technology*, Vol. 34, No. 4, S. 1475–1499.

Brell, T., Philipsen, R. & Ziefle, M. (2019) »sCARy! Risk Perceptions in Autonomous Driving: The Influence of Experience on Perceived Benefits and Barriers«, *Risk Analysis: An Official Publication of the Society for Risk Analysis*, Vol. 39, No. 2, S. 342–357.

Brey, P. (2010) »Values in technology and disclosive computer ethics«, in Floridi, L. (Hg.) *The Cambridge Handbook of Information and Computer Ethics,* Cambridge, Cambridge University Press, S. 41–58.

Brink, D. O. (1994) »Moral Conflict and Its Structure«, *The Philosophical Review*, Vol. 103, No. 2, S. 215–247.

Brooks, R. (2017) *Unexpected Consequences of Self Driving Cars* [Online]. Verfügbar unter http://rodneybrooks.com/unexpected-consequences-of-self-driving-cars/ (Abgerufen am 08. April 2023).

Broome, J. (1984) »Selecting People Randomly«, *Ethics*, Vol. 95, No. 1, S. 38–55.

Bruers, S. & Braeckman, J. (2014) »A Review and Systematization of the Trolley Problem«, *Philosophia*, Vol. 42, No. 2, S. 251–269.

Bruno, G., Spoto, A., Lotto, L., Cellini, N., Cutini, S. & Sarlo, M. (2023a) »Framing self-sacrifice in the investigation of moral judgment and moral emotions in human and autonomous driving dilemmas«, *Motivation and Emotion*, S. 1–14.

Bruno, G., Spoto, A., Sarlo, M., Lotto, L., Marson, A., Cellini, N. & Cutini, S. (2023b) »Moral reasoning behind the veil of ignorance: An investigation into perspective-taking accessibility in the context of autonomous vehicles«, *British Journal of Psychology*, S. 1–25.

Bundesministerium für Digitales und Verkehr (Hg.) (2021) *Gesetz zum autonomen Fahren tritt in Kraft* [Online]. Verfügbar unter https://www.bmvi.de/SharedDocs/DE/Artikel/DG/gesetz-zum-autonomen-fahren.html (Abgerufen am 04. Januar 2022).

Bundesministerium für Digitales und Verkehr (Hg.) (2022a) *AVF-Projekte* [Online]. Verfügbar unter https://www.bmvi.de/DE/Themen/Digitales/Automatisiertes-und-vernetztes-Fahren/AVF-Forschungsprogramm/Projekte/avf-projekte.html (Abgerufen am 04. Januar 2022).

Bundesministerium für Digitales und Verkehr (2022b) »Verordnung zur Regelung des Betriebs von Kraftfahrzeugen mit automatisierter und autonomer Fahrfunktion und zur Änderung straßenverkehrsrechtlicher Vorschriften«, *Bundesgesetzblatt*, Vol. 2022, Teil I Nr. 22.

Bundesministerium für Digitales und Verkehr (2023) »Digitalstrategie der Bundesregierung: Gemeinsam digitale Werte schöpfen«, 25. April [Online]. Verfügbar unter https://digitalstrategie-deutschland.de/static/fcf23bbf9736d543d02b79ccad34b729/Digitalstrategie_Aktualisierung_25.04.2023.pdf (Abgerufen am 04. Juli 2023).

Bundeszentrale für politische Bildung (2016) »Allokation«, in Bundeszentrale für politische Bildung (Hg.) *Duden Wirtschaft von A bis Z: Grundlagenwissen für Schule und Studium, Beruf und Alltag,* 6. Aufl., Mannheim, Bibliographisches Institut.

BverfG Bundesverfassungsgericht (2006) *Urteil des Ersten Senats vom 15. Februar 2006, 1 BvR 357/05 – Rn. (1 – 156)* [Online]. Verfügbar unter http://www.bverfg.de/e/rs200602151bvr035705.html (Abgerufen am 03. September 2022).

Cao, F., Zhang, J., Song, L., Wang, S., Miao, D. & Peng, J. (2017) »Framing Effect in the Trolley Problem and Footbridge Dilemma«, *Psychological Reports*, Vol. 120, No. 1, S. 88–101.

Cecchini, D., Brantley, S. & Dubljević, V. (2023) »Moral judgment in realistic traffic scenarios: moving beyond the trolley paradigm for ethics of autonomous vehicles«, *AI & Society*, S. 1–12.

Černý, D. (2022) »Autonomous Vehicles, the Badness of Death, and Discrimination«, in Jenkins, R., Černý, D. & Hríbek, T. (Hg.) *Autonomous Vehicle Ethics: The Trolley Problem and Beyond,* New York, Oxford University Press, S. 20–40.

Chasins, K. (2024) *Overview of Select NHTSA Activities Automated Driving Systems* [Online]. National Highway Traffic Safety Administration (NHTSA). Verfügbar unter https://www.nhtsa.gov/sites/nhtsa.gov/files/2024-02/16180-NSR-231211-007_SAE_Overview%20of%20Select%20NHTSA%20Activities%20-%20Automated%20Driving%20Systems-tag.pdf (Abgerufen am 20. August 2024).

Chilson, K. (2022) »An Epistemic Approach to Cultivating Appropriate Trust in Autonomous Vehicles«, in Jenkins, R., Černý, D. & Hríbek, T. (Hg.) *Autonomous Vehicle Ethics: The Trolley Problem and Beyond,* New York, Oxford University Press, S. 229–242.

Choi, J. K. & Ji, Y. G. (2015) »Investigating the Importance of Trust on Adopting an Autonomous Vehicle«, *International Journal of Human-Computer Interaction*, Vol. 31, No. 10, S. 692–702.

Coca-Vila, I. (2018) »Self-driving Cars in Dilemmatic Situations: An Approach Based on the Theory of Justification in Criminal Law«, *Criminal Law and Philosophy*, Vol. 12, No. 1, S. 59–82.

Coeckelbergh, M. (2016) »Responsibility and the Moral Phenomenology of Using Self-Driving Cars«, *Applied Artificial Intelligence*, Vol. 30, No. 8, S. 748–757.

Commission of the European Communites (2000) *Communication from the Commission on the Precautionary Principle* [Online]. Verfügbar unter https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2000:0001:FIN:en:PDF (Abgerufen am 19. Oktober 2023).

Conee, E. (1982) »Against Moral Dilemmas«, *The Philosophical Review*, Vol. 91, No. 1, S. 87–97.

Contissa, G., Lagioia, F. & Sartor, G. (2017) »The Ethical Knob: ethically-customisable automated vehicles and the law«, *Artificial Intelligence and Law*, Vol. 25, No. 3, S. 365–378.

Crew, B. (2015) »Driverless Cars Could Reduce Traffic Fatalities by Up to 90 %, Says Report«, *ScienceAlert*, 1. Oktober [Online]. Verfügbar unter chttps://www.sciencealert.com/driverless-cars-could-reduce-traffic-fatalities-by-up-to-90-says-report (Abgerufen am 09. Februar 2023).

Danaher, J. (2016) »Robots, law and the retribution gap«, *Ethics and Information Technology*, Vol. 18, No. 4, S. 299–309.

Dancy, J. (2017) *Moral Particularism* [Online], The Stanford Encyclopedia of Philosophy. Verfügbar unter https://plato.stanford.edu/archives/win2017/entries/moral-particularism/ (Abgerufen am 04. Dezember 2022).

Daniels, N. (1979) »Wide Reflective Equilibrium and Theory Acceptance in Ethics«, *The Journal of Philosophy*, Vol. 76, No. 5, S. 256–282.

Davnall, R. (2020) »Solving the Single-Vehicle Self-Driving Car Trolley Problem Using Risk Theory and Vehicle Dynamics«, *Science and Engineering Ethics*, Vol. 26, No. 1, S. 431–449.

Decker, M. (2013) »Technikfolgen«, in Grunwald, A. (Hg.) *Handbuch Technikethik,* Stuttgart, J.B. Metzler, S. 33–38.

Degenhart, C. (2023) »Grundrechte – Eine Einführung (Teil 1)«, *Iurratio Media GmbH*, 2023 [Online]. Verfügbar unter https://iurratio.de/journal/grundrechte-einfuehrung-degenhart (Abgerufen am 09. August 2023).

Deutscher Bundestag (Hg.) (2021) *Experten: Gesetz zum autonomen Fahren geht in die richtige Richtung* [Online]. Verfügbar unter https://www.bundestag.de/dokumente/textarchiv/2021/kw18-pa-verkehr-autonomes-fahren-835640 (Abgerufen am 04. Januar 2022).

Deutscher Bundestag (2020) *Technikfolgenabschätzung (TA) Autonome Waffensysteme: Bericht des Ausschusses für Bildung, Forschung und Technikfolgenabschätzung (18. Ausschuss) gemäß § 56a der Geschäftsordnung* [Online]. Verfügbar unter https://dserver.bundestag.de/btd/19/236/1923672.pdf (Abgerufen am 11. Juli 2023).

Dewitt, B., Fischhoff, B. & Sahlin, N.-E. (2019) „›Moral machine‹ experiment is no basis for policymaking«, *Nature*, Vol. 567, No. 7747, S. 31.

Diekmann, A. & Voss, T. (2004) »Die Theorie rationalen Handelns: Stand und Perspektiven«, in Diekmann, A. & Voss, T. (Hg.) *Rational-Choice-Theorie in den Sozialwissenschaften: Anwendungen und Probleme,* München, Oldenbourg, S. 13–29.

Dietmayer, K. (2015) »Prädiktion von maschineller Wahrnehmungsleistung beim automatisierten Fahren«, in Maurer, M., Gerdes, J. C., Lenz, B. & Winner, H. (Hg.) *Autonomes Fahren: Technische, rechtliche und gesellschaftliche Aspekte,* Berlin/Heidelberg, Springer Vieweg, S. 419–438.

Dietrich, M. (2020) »Understanding Autonomous Driving as Institutional Activity: Opening New Ways to React to Discriminatory Concerns in Autonomous Driving«, in Nørskov, M., Seibt, J. & Quick, O. S. (Hg.) *Culturally sustainable social robotics: Proceedings of Robophilosophy,* Amsterdam, IOS Press, S. 335–373.

Dietrich, M. (2021) »Addressing inequal risk exposure in the development of automated vehicles«, *Ethics and Information Technology*, Vol. 23, No. 4, S. 727–738.

Dietrich, M. & Weisswange, T. H. (2019) »Distributive justice as an ethical principle for autonomous vehicle behavior beyond hazard scenarios«, *Ethics and Information Technology*, Vol. 21, No. 3, S. 227–239.

Dietvorst, B. J., Simmons, J. P. & Massey, C. (2015) »Algorithm aversion: people erroneously avoid algorithms after seeing them err«, *Journal of Experimental Psychology: General*, Vol. 144, No. 1, S. 114–126.

Di Fabio, U., Broy, M., Jungo Brüngger, R., Eichhorn, U., Grunwald, A., Heckmann, D., Hilgendorf, E., Kagermann, H., Losinger, A., Lutz-Bachmann, M., Lütge, C., Markl, A., Müller, K. & Nehm, K. (2017) *Ethik-Kommission Automatisiertes und vernetztes Fahren, Bericht Juni 2017* [Online]. Bundesministerium für Verkehr und Digitale Infrastruktur. Verfügbar unter https://www.bmvi.de/SharedDocs/DE/Publikationen/DG/bericht-der-ethik-kommission.pdf?__blob=publicationFile (Abgerufen am 14. November 2023).

Dilich, M. A., Kopernik, D. & Goebelbecker, J. (2002) »Evaluating Driver Response to a Sudden Emergency: Issues of Expectancy, Emotional Arousal and Uncertainty«, *SAE Transactions*, Vol. 111, S. 238–248.

Dogan, E., Costantini, F. & Le Boennec, R. (2020) »Ethical issues concerning automated vehicles and their implications for transport«, in Milakis, D., Thomopoulos, N. & van Wee, B. (Hg.) *Advances in Transport Policy and Planning*, S. 215–233.

Donagan, A. (1984) »Consistency in Rationalist Moral Systems«, *The Journal of Philosophy*, Vol. 81, No. 6, S. 291–309.

Douglas, M. & Wildavsky, A. (1983) *Risk and Culture: An Essay on the Selection of Technological and Environmental Dangers*, Berkeley and Los Angeles, University of California Press.

Douma, F. & Palodichuk, S. A. (2012) »Criminal liability issues created by autonomous vehicles«, *Santa Clara Law Review*, Vol. 52, No. 4, S. 1157–1170.

Dubljevic, V. & Bauer, W. A. (2022) »Autonomous Vehicles and the Basic Structure of Society«, in Jenkins, R., Černý, D. & Hríbek, T. (Hg.) *Autonomous Vehicle Ethics: The Trolley Problem and Beyond,* New York, Oxford University Press, S. 295–314.

Dunn, J. (2012) »Virtual worlds and moral evaluation«, *Ethics and Information Technology*, Vol. 14, No. 4, S. 255–265.

Dworkin, R. (1981a) »What is Equality? Part 1: Equality of Welfare«, *Philosophy & Public Affairs*, Vol. 10, No. 3, S. 185–246.

Dworkin, R. (1981b) »What is Equality? Part 2: Equality of Resources«, *Philosophy & Public Affairs*, Vol. 10, No. 4, S. 283–345.

Dworkin, R. (1987a) »What is Equality? Part 3: The Place of Liberty«, *Iowa Law Review*, Vol. 73, No. 1, S. 1–54.

Dworkin, R. (1987b) »What is Equality? Part 4: Political Equality«, *San Francisco Law Review*, Vol. 22, S. 1–30.

Dworkin, R. (2000) *Sovereign Virtue: The Theory and Practice of Equality*, Cambridge (Mass.), Harvard University Press.

Edmonds, D. (2018) »Cars without drivers still need a moral compass. But what kind?«, *The Guardian*, 14. November [Online]. Verfügbar unter https://www.theguardian.com/commentisfree/2018/nov/14/cars-drivers-ethical-dilemmas-machines (Abgerufen am 23. Februar 2022).

Edmonds, E. (2019) »Three in Four Americans Remain Afraid of Fully Self-Driving Vehicles«, *AAA NewsRoom*, 14. März [Online]. Verfügbar unter https://newsroom.aaa.com/2019/03/americans-fear-self-driving-cars-survey/ (Abgerufen am 22. Februar 2022).

Ess, C. (2019 [2009]) *Digital Media Ethics*, 3. Aufl., Cambridge, Polity Press.

Etzioni, A. & Etzioni, O. (2017) »Incorporating Ethics into Artificial Intelligence«, *The Journal of Ethics*, Vol. 21, No. 4, S. 403–418.

Europäische Gruppe für Ethik der Naturwissenschaften und der neuen Technologien (2018) *Erklärung zu künstlicher Intelligenz, Robotik und »autonomen« Systemen,* Publications Office [Online]. Verfügbar unter https://data.europa.eu/doi/10.2777/611386 (Abgerufen am 19. Februar 2023).

Europäische Kommission, Generaldirektion Forschung und Innovation (2020) *Ethics of Connected and Automated Vehicles: recommendations on road safety, privacy, fairness, explainability and responsibility*, Publications Office [Online]. Verfügbar unter https://data.europa.eu/doi/10.2777/035239 (Abgerufen am 13. Juni 2023).

Europäische Kommission (2024) *Künstliche Intelligenz – Fragen und Antworten* [Online]. Verfügbar unter https://ec.europa.eu/commission/presscorner/detail/de/QANDA_21_1683 (Abgerufen am 03. August 2024).

European Commission Mobility und Transport (2023) *ITS & Vulnerable Road Users* [Online]. Verfügbar unter https://transport.ec.europa.eu/transport-themes/intelligent-transport-systems/road/action-plan-and-directive/its-vulnerable-road-users_en (Abgerufen am 10. Oktober 2023).

Evans, K., Moura, N. de, Chauvier, S., Chatila, R. & Dogan, E. (2020) »Ethical Decision Making in Autonomous Vehicles: The AV Ethics Project«, *Science and Engineering Ethics*, Vol. 26, No. 6, S. 3285–3312.

Evans, L. (1996) »The dominant role of driver behavior in traffic safety«, *American Journal of Public Health*, Vol. 86, No. 6, S. 784–786.

Evans, L. (2008) »Death in Traffic: Why Are the Ethical Issues Ignored?«, *Studies in Ethics, Law, and Technology*, Vol. 2, No. 1.

Evans, N. G. (2022) »Ethics and Risk Distribution for Autonomous Vehicles«, in Jenkins, R., Černý, D. & Hríbek, T. (Hg.) *Autonomous Vehicle Ethics: The Trolley Problem and Beyond,* New York, Oxford University Press, S. 7–19.

Fagnant, D. J. & Kockelman, K. (2015) »Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations«, *Transportation Research Part A: Policy and Practice*, Vol. 77, S. 167–181.

Fagnant, D. J. & Kockelman, K. M. (2018) »Dynamic ride-sharing and fleet sizing for a system of shared autonomous vehicles in Austin, Texas«, *Transportation*, Vol. 45, No. 1, S. 143–158.

Fahrenholz, P. (2021) »Das ›Ja, aber‹-Gesetz von Andreas Scheuer«, *Süddeutsche Zeitung*, 23. Juni [Online]. Verfügbar unter https://www.sueddeutsche.de/auto/autonomes-fahren-mobilitaet-zukunft-verkehrspolitik-1.5325240 (Abgerufen am 04. November 2022).

Färber, B. (2015) »Kommunikationsprobleme zwischen autonomen Fahrzeugen und menschlichen Fahrern«, in Maurer, M., Gerdes, J. C., Lenz, B. & Winner, H. (Hg.) *Autonomes Fahren: Technische, rechtliche und gesellschaftliche Aspekte,* Berlin/Heidelberg, Springer Vieweg, S. 127–146.

Faulhaber, A. K., Dittmer, A., Blind, F., Wächter, M. A., Timm, S., Sütfeld, L. R., Stephan, A., Pipa, G. & König, P. (2019) »Human Decisions in Moral Dilemmas are Largely Described by Utilitarianism: Virtual Car Driving Study Provides Guidelines for Autonomous Driving Vehicles«, *Science and Engineering Ethics*, Vol. 25, No. 2, S. 399–418.

Ferretti, M. P. (2010) »Risk and distributive justice: the case of regulating new technologies«, *Science and Engineering Ethics*, Vol. 16, No. 3, S. 501–515.

Filipović, A. (2016) »Angewandte Ethik«, in Heesen, J. (Hg.) *Handbuch Medien- und Informationsethik,* Stuttgart, J.B. Metzler, S. 41–49.

Fleetwood, J. (2017) »Public Health, Ethics, and Autonomous Vehicles«, *American Journal of Public Health*, Vol. 107, No. 4, S. 532–537.

Floridi, L. (2019) »Autonomous Vehicles: from Whether and When to Where and How«, *Philosophy & Technology*, Vol. 32, No. 4, S. 569–573.

Floridi, L. & Sanders, J. W. (2004) »On the Morality of Artificial Agents«, *Minds and Machines*, Vol. 14, No. 3, S. 349–379.

Foot, P. (1978) »The Problem of Abortion and the Doctrine of the Double Effect«, in Foot, P. (Hg.) *Virtues and Vices and Other Essays in Moral Philosophy,* Berkeley and Los Angeles, University of California Press, S. 19–32.

Foot, P. (1987) »Moral Realism and Moral Dilemma«, in Gowans, C. W. (Hg.) *Moral Dilemmas,* New York, Oxford University Press, S. 250–270.

Formosa, P. (2022) »Autonomous Vehicles and Ethical Settings: Wo Should Decide?«, in Jenkins, R., Černý, D. & Hríbek, T. (Hg.) *Autonomous Vehicle Ethics: The Trolley Problem and Beyond,* New York, Oxford University Press, S. 176–190.

Fossa, F. (2023) »Unavoidable Collisions. The Automation of Moral Judgment«, in Fossa, F. (Hg.) *Ethics of Driving Automation,* Cham, Springer Nature Switzerland, S. 65–94.

Fossa, F. (2024) »Artificial intelligence and human autonomy: the case of driving automation«, *AI & Society*, S. 1–12.

Fournier, T. (2016) »Will My Next Car Be a Libertarian or a Utilitarian? Who Will Decide?«, *IEEE Technology and Society Magazine*, Vol. 35, No. 2, S. 40–45.

Fraedrich, E. & Lenz, B. (2014) »Automated driving: Individual and societal aspects«, *Transportation Research Record*, Vol. 2416, No. 1, S. 64–72.

Frank, D.-A., Chrysochou, P., Mitkidis, P. & Ariely, D. (2019) »Human decision-making biases in the moral dilemmas of autonomous vehicles«, *Scientific Reports*, Vol. 9, No. 1, S. 13080.

Frankfurt, H. G. (1997) »Equality and respect«, *Social Research*, Vol. 64, S. 3–15.

Freitas, J. de, Anthony, S. E., Censi, A. & Alvarez, G. A. (2020) »Doubting Driverless Dilemmas«, *Perspectives on Psychological Science*, Vol. 15, No. 5, S. 1284–1288.

Freitas, J. de, Censi, A., Walker Smith, B., Di Lillo, L., Anthony, S. E. & Frazzoli, E. (2021) »From driverless dilemmas to more practical commonsense tests for automated vehicles«, *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 118, No. 11, e2010202118.

Fried, B. H. (2012) »What Does Matter? The Case for Killing the Trolley Problem (Or Letting It Die)«, *The Philosophical Quarterly*, Vol. 62, No. 248, S. 505–529.

Friedrich, B. (2015) »Verkehrliche Wirkung autonomer Fahrzeuge«, in Maurer, M., Gerdes, J. C., Lenz, B. & Winner, H. (Hg.) *Autonomes Fahren: Technische, rechtliche und gesellschaftliche Aspekte,* Berlin/Heidelberg, Springer Vieweg, S. 331–350.

Frison, A.-K., Wintersberger, P. & Riener, A. (2016) »First Person Trolley Problem: Evaluation of Drivers' Ethical Decisions in a Driving Simulator«, *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications Adjunct*, ACM, S. 117–122.

Future of Life Institute (FLI) (2024) *Zusammenfassung des AI-Gesetzes auf hoher Ebene* [Online]. Verfügbar unter https://artificialintelligenceact.eu/de/high-level-summary/ (Abgerufen am 30. August 2024).

Gantsho, L. (2022) »God does not play dice but self-driving cars should«, *AI and Ethics*, Vol. 2, S. 177–184.

Garza, A. P. (2011) »Look ma, no hands: Wrinkles and wrecks in the age of autonomous vehicles«, *New Eng. L. Rev.*, Vol. 46, S. 581–616.

Gasser, T. M. (2015) »Grundlegende und spezielle Rechtsfragen für autonome Fahrzeuge«, in Maurer, M., Gerdes, J. C., Lenz, B. & Winner, H. (Hg.) *Autonomes Fahren: Technische, rechtliche und gesellschaftliche Aspekte,* Berlin/Heidelberg, Springer Vieweg, S. 543–574.

Geiger, T., Wieler, J. & Rudschies, W. (2024) »Test BMW i5: Was den elektrischen 5er so gut macht«, *ADAC Online*, 10. Juni [Online]. Verfügbar unter https://www.adac.de/rund-ums-fahrzeug/autokatalog/marken-modelle/bmw/bmw-i5/#freihaendig-auf-der-autobahn-fahren (Abgerufen am 28. Juli 2024).

Geisslinger, M., Poszler, F., Betz, J., Lütge, C. & Lienkamp, M. (2021) »Autonomous Driving Ethics: from Trolley Problem to Ethics of Risk«, *Philosophy & Technology*, Vol. 34, No. 4, S. 1033–1055.

Geisslinger, M., Poszler, F. & Lienkamp, M. (2023a) »An ethical trajectory planning algorithm for autonomous vehicles«, *Nature Machine Intelligence*, Vol. 5, No. 2, S. 137–144.

Geisslinger, M., Trauth, R., Kaljavesi, G. & Lienkamp, M. (2023b) »Maximum Acceptable Risk as Criterion for Decision-Making in Autonomous Vehicle Trajectory Planning«, *IEEE Open Journal of Intelligent Transportation Systems*, Vol. 4, S. 570–579.

Gerdes, J. C. (2020) »The Virtues of Automated Vehicle Safety – Mapping Vehicle Safety Approaches to Their Underlying Ethical Frameworks«, *IEEE Intelligent Vehicles Symposium (IV)*, Las Vegas, NV, USA, S. 107–113.

Gerdes, J. C. & Thornton, S. M. (2015) »Implementable Ethics for Autonomous Vehicles«, in Maurer, M., Gerdes, J. C., Lenz, B. & Winner, H. (Hg.) *Autonomes Fahren: Technische, rechtliche und gesellschaftliche Aspekte,* Berlin/Heidelberg, Springer Vieweg, S. 87–102.

Gethmann, C. F. (2000) »Ethische Probleme der Verteilungsgerechtigkeit beim Handeln unter Risiko«, in Gethmann-Siefert, A. & Gethmann, C. F. (Hg.) *Philosophie und Technik,* München, Wilhelm Fink Verlag, S. 61–73.

Geyer, S., Baltzer, M., Franz, B., Hakuli, S., Kauer, M., Kienle, M., Meier, S., Weißgerber, T., Bengler, K., Bruder, R., Flemisch, F. & Winner, H. (2014) »Concept and development of a unified ontology for generating test and use-case catalogues for assisted and automated vehicle guidance«, *IET Intelligent Transport Systems*, Vol. 8, No. 3, S. 183–189.

Gibson, M. (1985) »Consent and Autonomy«, in Gibson, M. (Hg.) *To Breathe Freely: Risk, Consent, and Air,* Totowa, Rowman & Littlefield, S. 141–168.

Gill, T. (2021) »Ethical dilemmas are really important to potential adopters of autonomous vehicles«, *Ethics and Information Technology*, Vol. 23, No. 4, S. 657–673.

Glancy, D. J. (2012) »Privacy in autonomous vehicles«, *Santa Clara Law Review*, Vol. 52, No. 4, S. 1171–1240.

Goër Herve, M. de, Schinko, T. & Handmer, J. (2023) »Risk justice: Boosting the contribution of risk management to sustainable development«, *Risk Analysis: An Official Publication of the Society for Risk Analysis*, S. 1–15.

Göpfert, A. (2024) »Hat Tesla noch eine Chance gegen Waymo?«, *ARD-Tagesschau/tagesschau.de*, 22. Juni [Online]. Verfügbar unter https://www.tagesschau.de/wirtschaft/technologie/tesla-waymo-robotaxi-100.html (Abgerufen am 24. August 2024).

Gogoll, J. & Müller, J. F. (2017) »Autonomous Cars: In Favor of a Mandatory Ethics Setting«, *Science and Engineering Ethics*, Vol. 23, No. 3, S. 681–700.

Goodall, N. J. (2014a) »Ethical Decision Making During Automated Vehicle Crashes«, *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2424, S. 58–65.

Goodall, N. J. (2014b) »Machine Ethics and Automated Vehicles«, in Meyer, G. & Beiker, S. (Hg.) *Road Vehicle Automation: Lecture Notes in Mobility,* Cham, Springer International Publishing, S. 93–102.

Goodall, N. J. (2016a) »Away from Trolley Problems and Toward Risk Management«, *Applied Artificial Intelligence*, Vol. 30, No. 8, S. 810–821.

Goodall, N. J. (2016b) »Can you program ethics into a self-driving car?«, *IEEE Spectrum*, Vol. 53, No. 6, S. 28–58.

Goodall, N. J. (2017) »From Trolleys to Risk: Models for Ethical Autonomous Driving«, *American Journal of Public Health*, Vol. 107, No. 4, S. 496.

Goodall, N. J. (2020) »Vehicle Automation and the Duty to Act«, *Proceedings of the 21st World Congress on Intelligent Transport Systems, 7–11 September 2014, Detroit*, S. 1–8.

Goodall, N. J. (2021) »Comparison of automated vehicle struck-from-behind crash rates with national rates using naturalistic data«, *Accident Analysis and Prevention*, Vol. 154, S. 106056.

Gowans, C. W. (1987) »Introduction: The Debate on Moral Dilemmas«, in Gowans, C. W. (Hg.) *Moral Dilemmas,* New York, Oxford University Press, S. 3–33.

Gowans, C. W. (1994) *Innocence Lost: An Examination of Inescapable Moral Wrongdoing*, New York, Oxford University Press.

Graham, J. D. & Hsia, S. (2002) »Europe's precautionary principle: promise and pitfalls«, *Journal of Risk Research*, Vol. 5, No. 4, S. 371–390.

Grau, C. (2006) »There Is No ›I‹ in ›Robot‹: Robots and Utilitarianism«, *IEEE Intelligent Systems*, Vol. 21, No. 4, S. 52–55.

Greene, J. (2013) *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them*, New York, Penguin.

Greene, J. & Haidt, J. (2002) »How (and where) does moral judgment work?«, *Trends in Cognitive Sciences*, Vol. 6, No. 12, S. 517–523.

Greenspan, P. S. (1995) *Practical Guilt: Moral Dilemmas, Emotions, and Social Norms*, New York, Oxford University Press.

Grunwald, A. (2005) »Zur Rolle von Akzeptanz und Akzeptabilität von Technik bei der Bewältigung von Technikkonflikten«, *ATuP-Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis*, Vol. 14, No. 3, S. 54–60.

Grunwald, A. (2010 [2002]) *Technikfolgenabschätzung: Eine Einführung*, 2. Aufl., Berlin, edition sigma.

Grunwald, A. (2013) »Einleitung und Überblick«, in Grunwald, A. (Hg.) *Handbuch Technikethik,* Stuttgart, J.B. Metzler, S. 1–11.

Grunwald, A. (2015) »Gesellschaftliche Risikokonstellation für autonomes Fahren – Analyse, Einordnung und Bewertung«, in Maurer, M., Gerdes, J. C., Lenz, B. & Winner, H. (Hg.) *Autonomes Fahren: Technische, rechtliche und gesellschaftliche Aspekte,* Berlin/Heidelberg, Springer Vieweg, S. 661–685.

Grunwald, A. (2016) »Technikethik«, in Heesen, J. (Hg.) *Handbuch Medien- und Informationsethik,* Stuttgart, J.B. Metzler, S. 25–33.

Gunkel, D. J. (2020) »Mind the gap: responsible robotics and the problem of responsibility«, *Ethics and Information Technology*, Vol. 22, S. 307–320.

Gurney, J. K. (2013) »Sue my car, not me: products liability and accidents involving autonomous vehicles«, *Journal of Law, Technology and Policy*, Vol. 2, S. 247–277.

Gurney, J. K. (2015) »Crashing Into the Unknown: An Examination of Crash Optimization Algorithms Through the Two Lanes of Ethics and Law«, *Albany Law Review*, Vol. 79, S. 183–267.

Gurney, J. K. (2017) »Imputing Driverhood: Applying a Reasonable Driver Standard to Accidents Caused by Autonomous Vehicles«, in Lin, P., Jenkins, R. & Abney, K. (Hg.) *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence,* Oxford, Oxford University Press, S. 51–65.

Gurney, J. K. (2022) »Unintended Externalities of Highly Automated Vehicles«, in Jenkins, R., Černý, D. & Hríbek, T. (Hg.) *Autonomous Vehicle Ethics: The Trolley Problem and Beyond,* New York, Oxford University Press, S. 147–158.

Habermas, J (Hg.) (1991a) *Erläuterungen zur Diskursethik*, Frankfurt/Main, Suhrkamp Verlag.

Habermas, J. (1991b) »Vom pragmatischen, ethischen und moralischen Gebrauch der praktischen Vernunft«, in Habermas, J. (Hg.) *Erläuterungen zur Diskursethik,* Frankfurt/Main, Suhrkamp Verlag, S. 100–118.

Habermas, J. (2000 [1968]) *Technik und Wissenschaft als ›Ideologie‹*, 17. Aufl., Frankfurt/Main, Suhrkamp.

Hacking, I. (1986) »Culpable Ignorance of Interferences Effects«, in MacLean, D. (Hg.) *Values at Risk,* Savage, Rowman & Littlefield, S. 136–154.

Hansson, S. O. (1993) »The False Promises of Risk Analysis«, *Ratio*, Vol. 6, S. 16–26.

Hansson, S. O. (2003) »Ethical Criteria of Risk Acceptance«, *Erkenntnis*, Vol. 59, No. 3, S. 291–309.

Hansson, S. O. (2005) »Seven Myths of Risk«, *Risk Management*, Vol. 7, S. 7–17.

Hansson, S. O. (2007a) »Ethics and radiation protection«, *Journal of Radiological Protection*, Vol. 27, No. 2, S. 147–156.

Hansson, S. O. (2007b) »Risk and Ethics: Three Approaches«, in Lewens, T. (Hg.) *Risk: Philosophical Perspectives,* London, Routledge, S. 21–35.

Hansson, S. O. (2009) »From the Casino to the Jungle: Dealing with Uncertainty in Technological Risk Management«, *Synthese*, Vol. 168, No. 3, S. 423–432.

Hansson, S. O. (2013) *The Ethics of Risk: Ethical Analysis in an Uncertain World*, Basingstoke, Hampshire, Palgrave Macmillan.

Hansson, S. O., Belin, M.-Å. & Lundgren, B. (2021) »Self-Driving Vehicles—an Ethical Overview«, *Philosophy & Technology*, Vol. 34, S. 1383–1408.

Hare, R. M. (1987) »Moral Conflicts«, in Gowans, C. W. (Hg.) *Moral Dilemmas,* New York, Oxford University Press, S. 205–238.

Harris, J. (1975) »The Survival Lottery«, *Philosophy*, Vol. 50, No. 191, S. 81–87.

Harris, J. (2020) »The Immoral Machine«, *Cambridge Quarterly of Healthcare Ethics*, Vol. 29, No. 1, S. 71–79.

Harsanyi, J. C. (1977a) »Advances in Understanding Rational Behaviour«, in Butts, R. E. & Hintikka, J. (Hg.) *Foundational Problems in Special Science,* Dordrecht, Reidel, S. 315–343.

Harsanyi, J. C. (1977b) »Morality and the Theory of Rational Behavior«, *Social Research*, Vol. 44, No. 4, S. 623–656.

Hayenhjelm, M. & Wolff, J. (2012) »The Moral Problem of Risk Impositions: A Survey of the Literature«, *European Journal of Philosophy*, Vol. 20, S. e26–e51.

Heinrichs, D. (2015) »Autonomes Fahren und Stadtstruktur«, in Maurer, M., Gerdes, J. C., Lenz, B. & Winner, H. (Hg.) *Autonomes Fahren: Technische, rechtliche und gesellschaftliche Aspekte,* Berlin/Heidelberg, Springer Vieweg, S. 219–239.

Hengl, H-T. (2024) »EU beschließt KI-Etikette«, *Magazin des Fraunhofer Instituts für Kognitive Systeme IKS*, 27. Mai [Online]. Verfügbar unter https://safe-intelligence.fraunhofer.de/artikel/eu-ai-act---eu-beschliesst-ki-etikette (Abgerufen am 03. August 2024).

Hevelke, A. & Nida-Rümelin, J. (2015a) »Ethische Fragen zum Verhalten selbstfahrender Autos bei unausweichlichen Unfällen: Der Schutz von Unbeteiligten«, *Zeitschrift für philosophische Forschung*, Vol. 69, No. 2, S. 217–224.

Hevelke, A. & Nida-Rümelin, J. (2015b) »Responsibility for crashes of autonomous vehicles: An ethical analysis«, *Science and Engineering Ethics*, Vol. 21, No. 3, S. 619–630.

Hevelke, A. & Nida-Rümelin, J. (2015c) »Selbstfahrende Autos und Trolley-Probleme: Zum Aufrechnen von Menschenleben im Falle unausweichlicher Unfälle«, *Jahrbuch für Wissenschaft und Ethik*, Vol. 19, No. 1, S. 5–24.

Hevelke, A. & Nida-Rümelin, J. (2017) »Intelligente Autos im Dilemma«, in Könneker, C. (Hg.) *Unsere digitale Zukunft: In welcher Welt wollen wir leben?,* Heidelberg, Springer, S. 195–204.

Hilgendorf, E. (2017a) »Auf dem Weg zu einer Regulierung des automatisierten Fahrens: Anmerkungen zur jüngsten Reform des StVG«, *Kriminalpolitische Zeitschrift (KriPoZ)*, Vol. 2, No. 4, S. 225–229.

Hilgendorf, E. (2017b) »Dilemma-Problem gelöst. Ergebnisse der Ethikkommission zum automatisierten Fahren«, *ATZelektronik*, Vol. 12, No. 4, S. 46–49.

Hilgendorf, E. (2018a) »Dilemma-Probleme beim automatisierten Fahren: Ein Beitrag zum Problem des Verrechnungsverbots im Zeitalter der Digitalisierung«, *Zeitschrift für die gesamte Strafrechtswissenschaft (ZStW)*, Vol. 130, No. 3, S. 674–703.

Hilgendorf, E. (2018b) »Offene Fragen der neuen Mobilität: Problemfelder im Kontext von automatisiertem Fahren und Recht«, *Recht – Automobil – Wirtschaft (RAW)*, Vol. 2, S. 85–93.

Hilgendorf, E. (2019) »Automatisiertes Fahren als Herausforderung für Ethik und Rechtswissenschaft«, in Bendel, O. (Hg.) *Handbuch Maschinenethik,* Wiesbaden, Springer VS, S. 355–372.

Himmelreich, J. (2018) »Never Mind the Trolley: The Ethics of Autonomous Vehicles in Mundane Situations«, *Ethical Theory and Moral Practice*, Vol. 21, No. 3, S. 669–684.

Himmelreich, J. (2019) »Ethics of technology needs more political philosophy«, *Communications of the ACM*, Vol. 63, No. 1, S. 33–35.

Höffe, O. (1993) *Moral als Preis der Moderne: Ein Versuch über Wissenschaft, Technik und Umwelt*, Frankfurt/Main, Suhrkamp.

Höhn, H.-J. (1996) »Technikethik als Risikoethik. Ansätze einer sozialethischen Risikobeurteilung«, *Jahrbuch für christliche Sozialwissenschaften*, Vol. 37, S. 29–50.

Holbo, J. (2002) »Moral Dilemmas and the Logic of Obligation«, *American Philosophical Quarterly*, Vol. 39, No. 3, S. 259–274.

Holzbock, A., Kern, N., Waldschmidt, C., Dietmayer, K. & Belagiannis, V. (2023a) »Gesture Recognition with Keypoint and Radar Stream Fusion for Automated Vehicles«, in Karlinsky, M., Michaeli, T. & Nishino, K. (Hg.) *Computer Vision – ECCV 2022 Workshops*, Cham, Springer, S. 570–584.

Holzbock, A., Tsaregorodtsev, A. & Belagiannis, V. (2023b) »Pedestrian Environment Model for Automated Driving«, *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, S. 534–540.

Horn, C. (2003) »Zum Begriff der Gerechtigkeit«, *Die Politische Meinung*, Vol. 409, S. 25–36.

Howard, D. (2013) *Robots on the Road: The Moral Imperative of the Driverless Car* [Online]. Verfügbar unter https://donhoward-blog.nd.edu/2013/11/07/robots-on-the-road-the-moral-imperative-of-the-driverless-car/#.YUxkf2JBxPYU1oq-1ffKZ1 (Abgerufen am 11. Juni 2023).

Hubig, C. (1993) *Technik- und Wissenschaftsethik: Ein Leitfaden*, Berlin/Heidelberg, Springer.

Hubig, C. (1999) »Pragmatische Entscheidungslegitimation angesichts von Expertendilemmata«, in Grunwald, A. (Hg.) *Ethik in der Technikgestaltung: Praktische Relevanz und Legitimation,* Berlin, Springer, S. 197–209.

Hubmann, C., Becker, M., Althoff, D., Lenz, D. & Stiller, C. (2017) »Decision Making for Autonomous Driving considering Interaction and Uncertain Prediction of Surrounding Vehicles«, *IEEE Intelligent Vehicles Symposium (IV)*, S. 1671–1678.

Hübner, D. & White, L. (2018) »Crash Algorithms for Autonomous Cars: How the Trolley Problem Can Move Us Beyond Harm Minimisation«, *Ethical Theory and Moral Practice*, Vol. 21, No. 3, S. 685–698.

Hula, A., Snapp, L., Alson, J. & Simon, K. (2018) »The Environmental Potential of Autonomous Vehicles«, in Meyer, G. & Beiker, S. (Hg.) *Road Vehicle Automation 4: Lecture Notes in Mobility,* Cham, Springer International Publishing, S. 89–95.

Hulverscheidt, C. (2015) »Crash-Kurs mit Google«, *Süddeutsche Zeitung*, 10. Oktober [Online]. Verfügbar unter https://www.sueddeutsche.de/auto/autonomes-fahren-crash-kurs-mit-google-1.2684782 (Abgerufen am 17. Mai 2023).

Hursthouse, R. & Pettigrove, G. (2023) *Virtue Ethics* [Online], The Stanford Encyclopedia of Philosophy Archive. Verfügbar unter https://plato.stanford.edu/archives/fall2023/entries/ethics-virtue/ (Abgerufen am 02. Dezember 2023).

Hütt, M.-T. & Schubert, C. (2020) »Fairness von KI-Algorithmen«, in Mainzer, K. (Hg.) *Philosophisches Handbuch Künstliche Intelligenz,* Wiesbaden, Springer VS, S. 1–22.

Institute of Electrical and Electronics Engineers (IEEE) (2021) »IEEE Standard Model Process for Addressing Ethical Concerns during System Design«, *IEEE Std 7000–2021* [Online]. Verfügbar unter https://ieeexplore.ieee.org/servlet/opac?punumber=9536677 (Abgerufen am 29. August 2024).

Inoue, A., Shimizu, K., Udagawa, D. & Wakamatsu, Y. (2022) »The Trolley Problem and the Ethics of Autonomous Vehicles in the Eyes of the Public: Experimental Evidence«, in Jenkins, R., Černý, D. & Hríbek, T. (Hg.) *Autonomous Vehicle Ethics: The Trolley Problem and Beyond,* New York, Oxford University Press, S. 80–98.

International Organisation for Standardisation (2018) *ISO 26262–1:2018: Road vehicles – Functional Safety* [Online]. Verfügbar unter https://www.iso.org/standard/68383.html (Abgerufen am 21. August 2023).

Jackson, T. (2009) »Prosperity without growth? The transition to a sustainable economy«, *Sustainable Development Commission*, 2009 [Online]. Verfügbar unter https://www.sd-commission.org.uk/data/files/publications/prosperity_without_growth_report.pdf (Abgerufen am 18. Januar 2022).

Jacquette, D. (1991) »Moral Dilemmas, Disjunctive Obligations, and Kant's Principle That ›Ought‹ Implies ›Can‹", *Synthese*, Vol. 88, No. 1, S. 43–55.

JafariNaimi, N. (2018) »Our Bodies in the Trolley's Path, or Why Self-driving Cars Must *Not* Be Programmed to Kill«, *Science, Technology, & Human Values*, Vol. 43, No. 2, S. 302–323.

Jenkins, R. (2022) »Introduction to: Part II, Ethical Issues Beyond the Trolley Problem«, in Jenkins, R., Černý, D. & Hríbek, T. (Hg.) *Autonomous Vehicle Ethics: The Trolley Problem and Beyond,* New York, Oxford University Press, S. 143–146.

Jonas, H. (1979) *Das Prinzip Verantwortung: Versuch einer Ethik für die technologische Zivilisation*, Frankfurt/Main, Insel Verlag.

Jonas, H. (1993) »Warum die Technik ein Gegenstand für die Technik ist. Fünf Gründe«, in Lenk, H. & Ropohl, G. (Hg.) *Technik und Ethik,* 2. Aufl., Stuttgart, Reclam, Philipp, jun. GmbH, Verlag, S. 81–91.

Jong, R. de (2020) »The Retribution-Gap and Responsibility-Loci Related to Robots and Automated Technologies: A Reply to Nyholm«, *Science and Engineering Ethics*, Vol. 26, No. 2, S. 727–735.

Jungermann, H. & Slovic, P. (1997) »Die Psychologie der Kognition und Evaluation von Risiko«, in Bechmann, G. (Hg.) *Risiko und Gesellschaft: Grundlagen und Ergebnisse interdisziplinärer Risikoforschung,* 2. Aufl., Opladen, Westdeutscher Verlag, S. 167–207.

Kahn, L. (2022) »How Soon Is Now?: On the Timing and Conditions for Adopting Widespread Use of Autonomous Vehicles«, in Jenkins, R., Černý, D. & Hríbek, T. (Hg.) *Autonomous Vehicle Ethics: The Trolley Problem and Beyond,* New York, Oxford University Press, S. 243–256.

Kahneman, D. & Tversky, A. (1979) »Prospect Theory: An Analysis of Decision under Risk«, *Econometrica*, Vol. 47, No. 2, S. 263–291.

Kallioinen, N., Pershina, M., Zeiser, J., Nosrat Nezami, F., Pipa, G., Stephan, A. & König, P. (2019) »Moral Judgements on the Actions of Self-Driving Cars and Human Drivers in Dilemma Situations From Different Perspectives«, *Frontiers in Psychology*, Vol. 10, S. 2415.

Kamm, F. M. (2020) »The Use and Abuse of the Trolley Problem: Self-Driving Cars, Medical Treatments, and the Distribution of Harm«, in Liao, S. M. (Hg.) *Ethics of Artificial Intelligence,* Oxford, Oxford University Press, S. 79–108.

Kant, I. (1900ff.) *Gesammelte Schriften*, Berlin, Preußische Akademie der Wissenschaften.

Karner, A., London, J., Rowangould, D. & Manaugh, K. (2020) »From Transportation Equity to Transportation Justice: Within, Through, and Beyond the State«, *Journal of Planning Literature*, Vol. 35, No. 4, S. 440–459.

Karnouskos, S. (2020) »Self-Driving Car Acceptance and the Role of Ethics«, *IEEE Transactions on Engineering Management*, Vol. 67, No. 2, S. 252–265.

Kauppinen, A. (2021) »Who Should Bear the Risk When Self-Driving Vehicles Crash?«, *Journal of Applied Philosophy*, Vol. 38, No. 4, S. 630–645.

Keeling, G. (2018a) »Against Leben's Rawlsian Collision Algorithm for Autonomous Vehicles«, in Müller, V. C. (Hg.) *Philosophy and Theory of Artificial Intelligence 2017,* Cham, Springer, S. 259–272.

Keeling, G. (2018b) »Legal Necessity, Pareto Efficiency & Justified Killing in Autonomous Vehicle Collisions«, *Ethical Theory and Moral Practice*, Vol. 21, No. 2, S. 413–427.

Keeling, G. (2020) »Why Trolley Problems Matter for the Ethics of Automated Vehicles«, *Science and Engineering Ethics*, Vol. 26, S. 293–307.

Keeling, G. (2022) »Automated Vehicles and the Ethics of Classification«, in Jenkins, R., Černý, D. & Hríbek, T. (Hg.) *Autonomous Vehicle Ethics: The Trolley Problem and Beyond,* New York, Oxford University Press, S. 41–57.

Keeling, G., Evans, K., Thornton, S. M., Mecacci, G. & Santoni de Sio, F. (2019) »Four Perspectives on What Matters for the Ethics of Automated Vehicles«, in Meyer, G. & Beiker, S. (Hg.) *Road Vehicle Automation 6: Lecture Notes in Mobility,* Cham, Springer International Publishing, S. 49–60.

Kennemer, Q. (2024) »Cruise will resume robotaxi tests after one of its cars ran someone over«, *The Verge* [Online]. Verfügbar unter https://www.theverge.com/2024/4/9/24125618/cruise-resume-robotaxi-testing-self-driving (Abgerufen am 11. August 2024).

Kenwright, B. (2018) »Virtual Reality: Ethical Challenges and Dangers [Opinion]«, *IEEE Technology and Society Magazine*, Vol. 37, No. 4, S. 20–25.

Kinjo, K. & Ebina, T. (2017) »Optimal program for autonomous driving under Bentham- and Nash-type social welfare functions«, *Procedia Computer Science*, Vol. 112, S. 61–70.

Kirkpatrick, K. (2015) »The moral challenges of driverless cars«, *Communications of the ACM*, Vol. 58, No. 8, S. 19–20.

Klima-Allianz Deutschland (2020) »Klimafreundliche Mobilität für alle: Positionspapier für Politik und Entscheider*innen«, Mai 2020 [Online]. Verfügbar unter https://www.klima-allianz.de/fileadmin/user_upload/Dateien/Daten/Publikationen/Positionen/Klimafreundliche_Mobilit%C3%A4t_f%C3%BCr_alle.pdf (Abgerufen am 06. Dezember 2021).

Klincewicz, M. (2017) »Challenges to Engineering Moral Reasoners: Time and Context«, in Lin, P., Jenkins, R. & Abney, K. (Hg.) *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence,* Oxford, Oxford University Press, S. 244–257.

Knight, F. (1921) *Risk, Uncertainty and Profit*, Boston, Houghton Mifflin Company.

Koch, B. & Rinke, B. (2018) *Ethische Fragestellungen im Kontext autonomer Waffensysteme,* Institut für Theologie und Frieden, Hamburg.

Koch, B. A. (2022) »Produkthaftung für autonome Fahrzeuge«, in Laimer, S. & Perathoner, C. (Hg.) *Mobilitäts- und Transportrecht in Europa: Bestandsaufnahme und Zukunftsperspektiven,* Berlin/Heidelberg, Springer, S. 113–129.

Köllner, C. (2017) »Verkehrstote gibt es trotz automatisiertem Straßenverkehr«, *Springer Professional* [Online]. Verfügbar unter https://www.springerprofessional.de/fahrzeugsicherheit/automatisiertes-fahren/verkehrstote-gibt-es-trotz-automatisiertem-strassenverkehr/15292594 (Abgerufen am 15. Januar 2022).

Köllner, C. (2018) »Vision Zero ist nur eine Vision«, *Springer Professional* [Online]. Verfügbar unter https://www.springerprofessional.de/automatisiertes-fahren/fahrzeugsicherheit/vision-zero-ist-nur-eine-vision/15548402 (Abgerufen am 15. Januar 2022).

Köllner, C. (2021) »Teilautomatisierung macht unaufmerksam«, *Springer Professional* [Online]. Verfügbar unter https://www.springerprofessional.de/fahrerassistenz/verkehrssicherheit/teilautomatisierung-macht-unaufmerksam/18749652 (Abgerufen am 02. Februar 2023).

Köllner, C. (2024a) »Was bedeutet das KI-Gesetz für die Autoindustrie?«, *Springer Professional* [Online]. Verfügbar unter https://www.springerprofessional.de/kuenstliche-intelligenz/automatisiertes-fahren/was-bedeutet-das-ki-gesetz-fuer-die-autoindustrie-/26700940 (Abgerufen am 03. August 2024).

Köllner, C. (2024b) »Vector bringt zertifizierte Basissoftware für autonomes Fahren«, *Springer Professional* [Online]. Verfügbar https://www.springerprofessional.de/automobilelektronik---software/automatisiertes-fahren/vector-bringt-zertifizierte-basissoftware-fuer-autonomes-fahren/27440072 (Abgerufen am 09. August 2024).

Köllner, C. (2024c) »Mercedes-Benz hat SAE-Level-4-Freigabe für Peking«, *Springer Professional* [Online]. Verfügbar unter h https://www.springerprofessional.de/automatisiertes-fahren/unternehmen---institutionen/mercedes-benz-hat-sae-level-4-freigabe-fuer-peking/27457604 (Abgerufen am 19. August 2024).

KPMG LLP, Center for Automotive Research (2012) *Self-driving cars: The next revolution* [Online]. Verfügbar unter https://www.cargroup.org/wp-content/uploads/2017/02/Self_driving-cars-The-next-revolution.pdf (Abgerufen am 19. Februar 2022).

Kriebitz, A., Max, R. & Lütge, C. (2022) »The German Act on Autonomous Driving: Why Ethics Still Matters«, *Philosophy & Technology*, Vol. 35, No. 2, S. 29.

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

Krohn, W. & Krücken, G. (1993) »Risiko als Konstruktion und Wirklichkeit«, in Krohn, W. & Krücken, G. (Hg.) *Riskante Technologien: Reflexion und Regulation: Einführung in die sozialwissenschaftliche Risikoforschung,* Frankfurt/Main, Suhrkamp, S. 9–44.

Krügel, S. & Uhl, M. (2022) »Autonomous vehicles and moral judgments under risk«, *Transportation Research Part A: Policy and Practice*, Vol. 155, S. 1–10.

Krügel, S., Uhl, M. & Balcombe, B. (2021) »Automated vehicles and the morality of post-collision behavior«, *Ethics and Information Technology*, Vol. 23, No. 4, S. 691–701.

Kumfer, W. & Burgess, R. (2015) »Investigation into the Role of Rational Ethics in Crashes of Automated Vehicles«, *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2489, No. 1, S. 130–136.

Lacroix, A. (2018) »Sind Maschinen moralischer als wir?«, *Philosophie Magazin*, No. 4, S. 26–31.

LaCroix, T. (2022) »Moral dilemmas for moral machines«, *AI and Ethics*, Vol. 2, S. 737–746.

LaFrance, A. (2016) »How Self-Driving Cars Will Threaten Privacy«, *The Atlantic*, 21. März [Online]. Verfügbar unter https://www.theatlantic.com/technology/archive/2016/03/self-driving-cars-and-the-looming-privacy-apocalypse/474600/ (Abgerufen am 17. Februar 2022).

Lawlor, R. (2022) »The Ethics of Automated Vehicles: Why Self-driving Cars Should not Swerve in Dilemma Cases«, *Res Publica*, Vol. 28, No. 1, S. 193–216.

Leben, D. (2017) »A Rawlsian algorithm for autonomous vehicles«, *Ethics and Information Technology*, Vol. 19, No. 2, S. 107–115.

Leben, D. (2022) »Discrimination in Algorithmic Trolley Problems«, in Jenkins, R., Černý, D. & Hríbek, T. (Hg.) *Autonomous Vehicle Ethics: The Trolley Problem and Beyond,* New York, Oxford University Press, S. 130–142.

Lechner, D. & Malaterre, G. (1991) »Emergency Manuever Experimentation Using a Driving Simulator«, *SAE Technical Paper*, No. 910016.

Lee, D. (2019) »Uber self-driving crash ›mostly caused by human error‹“, *BBC News Online*, 20. November [Online]. Verfügbar unter https://www.bbc.com/news/technology-50484172 (Abgerufen am 20. September 2023).

Lemmon, E. J. (1962) »Moral Dilemmas«, *The Philosophical Review*, Vol. 71, No. 2, S. 139–158.

Lemmon, E. J. (1965) »Deontic Logic and the Logic of Imperatives«, *Logique et Analyse*, Vol. 8, No. 29, S. 39–71.

Lenk, H. & Maring, M. (Hg.) (1998) *Technikethik und Wirtschaftsethik: Fragen der praktischen Philosophie*, Opladen, Leske + Budrich.

Lenk, H. & Ropohl, G. (Hg.) (1993 [1987]) *Technik und Ethik*, 2. Aufl., Stuttgart, Reclam, Philipp, jun. GmbH, Verlag.

Lenz, B. & Fraedrich, E. (2015) »Neue Mobilitätskonzepte und autonomes Fahren: Potenziale der Veränderung«, in Maurer, M., Gerdes, J. C., Lenz, B. & Winner, H. (Hg.) *Autonomes Fahren: Technische, rechtliche und gesellschaftliche Aspekte,* Berlin/Heidelberg, Springer Vieweg, S. 175–195.

Leonard, H. B. & Zeckhauser, R. J. (1986) »Cost-Benefit Analysis Applied to Risks: Its Philosophy and Legitimacy«, in MacLean, D. (Hg.) *Values at Risk,* Savage, Rowman & Littlefield, S. 31–48.

Levin, S. & Wong, J. C. (2018) »Self-driving Uber kills Arizona woman in first fatal crash involving pedestrian«, *The Guardian*, 19. März [Online]. Verfügbar unter https://www.theguardian.com/technology/2018/mar/19/uber-self-driving-car-kills-woman-arizona-tempe (Abgerufen am 10. Februar 2022).

Lim, H. & Taeihagh, A. (2018) »Autonomous Vehicles for Smart and Sustainable Cities: An In-Depth Exploration of Privacy and Cybersecurity Implications«, *Energies*, Vol. 11, No. 5, S. 1062.

Lin, P., Jenkins, R. & Abney, K. (Hg.) (2017) *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*, Oxford, Oxford University Press.

Lin, P. (2013a) »The ethics of saving lives with autonomous cars is far murkier than you think«, *WIRED* [Online]. Verfügbar unter https://www.wired.com/2013/07/the-surprising-ethics-of-robot-cars (Abgerufen am 04. April 2022).

Lin, P. (2013b) »The Ethics of Autonomous Cars«, *The Atlantic* [Online]. Verfügbar unter https://www.theatlantic.com/technology/archive/2013/10/the-ethics-of-autonomous-cars/280360/ (Abgerufen am 03. Dezember 2023).

Lin, P. (2014a) »The Robot Car of Tomorrow May Just Be Programmed to Hit You«, *WIRED* [Online]. Verfügbar unter https://www.wired.com/2014/05/the-robot-car-of-tomorrow-might-just-be-programmed-to-hit-you/ (Abgerufen am 13. April 2022).

Lin, P. (2014b) »Here's a terrible idea: robot cars with adjustable ethics settings«, *WIRED* [Online]. Verfügbar unter https://www.wired.com/2014/08/heres-a-terrible-idea-robot-cars-with-adjustable-ethics-settings/ (Abgerufen am 03. Mai 2022).

Lin, P. (2015) »Why Ethics Matters for Autonomous Cars«, in Maurer, M., Gerdes, J. C., Lenz, B. & Winner, H. (Hg.) *Autonomes Fahren: Technische, rechtliche und gesellschaftliche Aspekte,* Berlin/Heidelberg, Springer Vieweg, S. 69–85.

Lin, P. (2017) »Robot Cars And Fake Ethical Dilemmas«, *Forbes Media LLC*, 3. April [Online]. Verfügbar unter https://www.forbes.com/sites/patricklin/2017/04/03/robot-cars-and-fake-ethical-dilemmas/?sh=8c31b4213a26 (Abgerufen am 25. Januar 2022).

Lippert, E., Prüfert, A. & Wachtler, G. (2013) »Einleitung«, in Lippert, E., Wachtler, G. & Prüfert, A. (Hg.) *Sicherheit in der unsicheren Gesellschaft,* Springer, S. 7–20.

Liu, H.-Y. (2016) »Structural Discrimination and Autonomous Vehicles: Immunity Devices, Trump Cards and Crash Optimisation«, in Seibt, J., Nørskov, M. & Andersen, S. S. (Hg.) *What Social Robots Can and Should Do: Proceedings of Robophilosophy 2016/TRANSOR 2016,* Amsterdam, IOS Press, S. 164–173.

Liu, H.-Y. (2017) »Irresponsibilities, inequalities and injustice for autonomous vehicles«, *Ethics and Information Technology*, Vol. 19, No. 3, S. 193–207.

Liu, H.-Y. (2018) »Three Types of Structural Discrimination Introduced by Autonomous Vehicles«, *UC Davis Law Review Online*, Vol. 51, S. 149–180.

Liu, P., Du, M. & Li, T. (2021) »Psychological consequences of legal responsibility misattribution associated with automated vehicles«, *Ethics and Information Technology*, Vol. 23, No. 4, S. 763–776.

Loh, W. & Loh, J. (2017) »Autonomy and Responsibility in Hybrid Systems: The Example of Autonomous Cars«, in Lin, P., Jenkins, R. & Abney, K. (Hg.) *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence,* Oxford, Oxford University Press, S. 35–50.

Loh, W. & Misselhorn, C. (2019) »Autonomous Driving and Perverse Incentives«, *Philosophy & Technology*, Vol. 32, No. 4, S. 575–590.

Lucas Jr., G. R. (2015) »Ethics and UAVs«, in Valavanis, K. P. & Vachtsevanos, G. J. (Hg.) *Handbook of Unmanned Aerial Vehicles,* Dordrecht, Springer, S. 2865–2878.

Lucifora, C., Grasso, G. M., Perconti, P. & Plebe, A. (2021) »Moral reasoning and automatic risk reaction during driving«, *Cognition, Technology & Work*, Vol. 23, No. 4, S. 705–713.

Luetge, C. (2017) »The German Ethics Code for Automated and Connected Driving«, *Philosophy & Technology*, Vol. 30, No. 4, S. 547–558.

Luhmann, N. (1991a) »Verständigung über Risiken und Gefahren«, *Die Politische Meinung*, Vol. 36, S. 86–95.

Luhmann, N. (1997) »Die Moral des Risikos und das Risiko der Moral«, in Bechmann, G. (Hg.) *Risiko und Gesellschaft: Grundlagen und Ergebnisse interdisziplinärer Risikoforschung,* 2. Aufl., Opladen, Westdeutscher Verlag, S. 327–338.

Luhmann, N. (2003 [1991b]) *Soziologie des Risikos*, Berlin/New York, de Gruyter.

Lundgren, B. (2021) »Safety requirements vs. crashing ethically: what matters most for policies on autonomous vehicles«, *AI & Society*, Vol. 36, S. 405–415.

Lütge, C. (2011) »Das Gefangenendilemma und seine ethischen Implikationen bei Aristoteles, Locke und Hume«, in Nguyen, T. (Hg.) *Mensch und Markt: Die ethische Dimension wirtschaftlichen Handelns,* Wiesbaden, Gabler Verlag Springer Fachmedien, S. 17–40.

Lütge, C., Kriebitz, A. & Max, R. (2020) »Ethische und rechtliche Herausforderungen des autonomen Fahrens«, in Mainzer, K. (Hg.) *Philosophisches Handbuch Künstliche Intelligenz,* Wiesbaden, Springer VS, S. 1–18.

Lyons, D. (1965) *Forms and Limits of Utilitarianism*, New York, Oxford University Press.

Marchant, G. E. & Lindor, R. A. (2012) »The Coming Collision Between Autonomous Vehicles and the Liability System«, *Santa Clara Law Review*, Vol. 52, No. 4, S. 1321–1340.

Marcus, G. (2012) »Moral Machines«, *The New Yorker*, 2012 [Online]. Verfügbar unter https://www.newyorker.com/news/news-desk/moral-machines (Abgerufen am 17. Mai 2023).

Marcus, R. B. (1980) »Moral Dilemmas and Consistency«, *The Journal of Philosophy*, Vol. 77, No. 3, S. 121–136.

Margalit, A. (1996) *The Decent Society*, Cambridge (Mass.), Harvard University Press.

Margalit, A. (1997) »Decent Equality and Freedom: A Postscript«, *Social Research*, Vol. 64, No. 1, S. 147–160.

Martens, K. (2016) *Transport Justice: Designing Fair Transportation Systems*, New York, Routledge.

Martin, D. (2017) »Who Should Decide How Machines Make Morally Laden Decisions?«, *Science and engineering ethics*, Vol. 23, No. 4, S. 951–967.

Martínez-Buelvas, L., Rakotonirainy, A., Grant-Smith, D. & Oviedo-Trespalacios, O. (2022) »A transport justice approach to integrating vulnerable road users with automated vehicles«, *Transportation Research Part D: Transport and Environment*, Vol. 113, S. 103499.

Matthias, A. (2004) »The responsibility gap: Ascribing responsibility for the actions of learning automata«, *Ethics and Information Technology*, Vol. 6, No. 3, S. 175–183.

Maurer, M., Gerdes, J. C., Lenz, B. & Winner, H. (Hg.) (2015) *Autonomes Fahren: Technische, rechtliche und gesellschaftliche Aspekte*, Berlin/Heidelberg, Springer Vieweg.

Maxmen, A. (2018) »Self-driving car dilemmas reveal that moral choices are not universal«, *Nature*, Vol. 562, No. 7728, S. 469.

Mayer, M. M., Bell, R. & Buchner, A. (2021) »Self-protective and self-sacrificing preferences of pedestrians and passengers in moral dilemmas involving autonomous vehicles«, *PloS one*, Vol. 16, No. 12, S. e0261673.

McCarthy, D. (1997) »Rights, Explanation, and Risks«, *Ethics*, Vol. 107, No. 2, S. 205–225.

McConnell, T. C. (1976) »Moral Dilemmas and Requiring the Impossible«, *Philosophical Studies*, Vol. 29, No. 6, S. 409–413.

McConnell, T. C. (1978) »Moral Dilemmas and Consistency in Ethics«, *Canadian Journal of Philosophy*, Vol. 8, No. 2, S. 269–287.

McConnell, T. C. (1996) »Moral Residue and Dilemmas«, in Mason, H. E. (Hg.) *Moral Dilemmas and Moral Theory,* New York, Oxford University Press, S. 36–47.

McConnell, T. C. (2022) *Moral Dilemmas* [Online], The Stanford Encyclopedia of Philosophy. Verfügbar unter https://plato.stanford.edu/archives/fall2022/entries/moral-dilemmas/ (Abgerufen am 15. April 2023).

McElligott, S. (2023) »NHTSA Announces New Autonomous Driving Regulations«, *U.S. News & World Report, L.P. (»U.S. News«)* [Online]. Verfügbar unter https://cars.usnews.com/cars-trucks/features/nhtsa-announces-new-autonomous-driving-regulations (Abgerufen am 09. August 2024).

McMillan, J. & King, M. (2017) »Why Be Moral in a Virtual World?«, *Journal of Practical Ethics*, Vol. 5, No. 2, S. 30–48.

Mecacci, G. & Santoni de Sio, F. (2020) »Meaningful human control as reason-responsiveness: the case of dual-mode vehicles«, *Ethics and Information Technology*, Vol. 22, No. 2, S. 103–115.

Meder, B., Fleischhut, N., Krumnau, N.-C. & Waldmann, M. R. (2019) »How Should Autonomous Cars Drive? A Preference for Defaults in Moral Judgments Under Risk and Uncertainty«, *Risk Analysis: An Official Publication of the Society for Risk Analysis*, Vol. 39, No. 2, S. 295–314.

Melo, C. M. de, Marsella, S. & Gratch, J. (2020) »Risk of Injury in Moral Dilemmas With Autonomous Vehicles«, *Frontiers in Robotics and AI*, Vol. 7, S. 572529.

Mercedes-Benz Group AG (2020) *Fahrerloses Parken. Automated Valet Parking* [Online]. Verfügbar unter https://group.mercedes-benz.com/innovation/case/autonomous/fahrerlos-geparkt.html (Abgerufen am 02. Dezember 2023).

Mercedes-Benz Group AG (2022) *Genehmigung für Serieneinsatz: Fahrerloses Parksystem von Mercedes-Benz und Bosch* [Online]. Verfügbar unter https://group.mercedes-benz.com/innovation/produktinnovation/autonomes-fahren/intelligent-park-pilot.html (Abgerufen am 14. August 2024).

Metz, C. (2016) »Self-Driving Cars Will Teach Themselves to Save Lives – But Also Take Them«, *WIRED* [Online]. Verfügbar unter https://www.wired.com/2016/06/self-driving-cars-will-power-kill-wont-conscience/ (Abgerufen am 01. Dezember 2022).

Mill, J. S. (1863 [1861]) *Utilitarianism*, London, Parker, Son and Bourn, West Strand.

Mill, J. S. (1963–91) *The Collected Works of John Stuart Mill*, Toronto, University of Toronto Press.

Millar, J. (2014a) »Proxy Prudence: Rethinking Models of Responsibility for Semi-Autonomous Robots«, *SSRN Electronic Journal*.

Millar, J. (2014b) »You Should Have a Say in Your Robot Car's Code of Ethics«, *WIRED* [Online]. Verfügbar unter https://www.wired.com/2014/09/set-the-ethics-robot-car/ (Abgerufen am 07. Juni 2023).

Millar, J. (2014c) »An ethical dilemma: When robot cars must kill, who should pick the victim?«, *Robohub, ROBOTS Association*, 11. Juni [Online]. Verfügbar unter https://robohub.org/an-ethical-dilemma-when-robot-cars-must-kill-who-should-pick-the-victim/ (Abgerufen am 06. März 2022).

Millar, J. (2015) »Technology as moral proxy: autonomy and paternalism by design«, *IEEE Technology and Society Magazine*, Vol. 34, No. 2, S. 47–55.

Millar, J., Paz, D., Thornton, S. M., Parisi, C. & Gerdes, J. C. (2020) »A Framework for Addressing Ethical Considerations in the Engineering of Automated Vehicles (and other technologies)«, *Proceedings of the Design Society: DESIGN Conference*, Vol. 1, S. 1485–1494.

Millard-Ball, A. (2018) »Pedestrians, Autonomous Vehicles, and Cities«, *Journal of Planning Education and Research*, Vol. 38, No. 1, S. 6–12.

Miller, K. W., Wolf, M. J. & Grodzinsky, F. (2017) »This ›Ethical Trap‹ Is for Roboticists, Not Robots: On the Issue of Artificial Agent Ethical Decision-Making«, *Science and Engineering Ethics*, Vol. 23, No. 2, S. 389–401.

Minx, E. & Dietrich, R. (2015) »Geleitwort«, in Maurer, M., Gerdes, J. C., Lenz, B. & Winner, H. (Hg.) *Autonomes Fahren: Technische, rechtliche und gesellschaftliche Aspekte,* Berlin/Heidelberg, Springer Vieweg, S. V–VII.

Misselhorn, C. (Hg.) (2015a) *Collective Agency and Cooperation in Natural and Artificial Systems*, Cham, Springer International Publishing.

Misselhorn, C. (2015b) »Collective Agency and Cooperation in Natural and Artificial Systems«, in Misselhorn, C. (Hg.) *Collective Agency and Cooperation in Natural and Artificial Systems,* Cham, Springer International Publishing, S. 3–24.

Misselhorn, C. (2018a) »Artificial Morality. Concepts, Issues and Challenges«, *Society*, Vol. 55, S. 161–169.

Misselhorn, C. (2018b) *Grundfragen der Maschinenethik*, Stuttgart, Reclam.

Misselhorn, C. (2019) »Maschinenethik und Philosophie«, in Bendel, O. (Hg.) *Handbuch Maschinenethik,* Wiesbaden, Springer VS, S. 33–56.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S. & Floridi, L. (2016) »The ethics of algorithms: Mapping the debate«, *Big Data & Society*, Vol. 3, No. 2, S. 1–21.

Mittelstraß, J. (1991) »Auf dem Wege zu einer Reparaturethik?«, in Wils, Jean-Pierre: Mieth, Dietmar (Hg.) *Ethik ohne Chance? Erkundungen im technologischen Zeitalter,* Tübingen, Attempto Verlag, S. 89–108.

Mladenovic, M. N. & McPherson, T. (2016) »Engineering Social Justice into Traffic Control for Self-Driving Vehicles?«, *Science and Engineering Ethics*, Vol. 22, No. 4, S. 1131–1149.

Moor, J. H. (2005) »Why we need better ethics for emerging technologies«, *Ethics and Information Technology*, Vol. 7, No. 3, S. 111–119.

Moor, J. H. (2006) »The Nature, Importance, and Difficulty of Machine Ethics«, *IEEE Intelligent Systems*, Vol. 21, No. 4, S. 18–21.

Moor, R. (2016) *What happens to American myth when you take the driver out of it? The self-driving car and the future of the self* [Online]. Verfügbar unter https://nymag.com/intelligencer/2016/10/is-the-self-driving-car-un-american.html (Abgerufen am 08. März 2022).

Moravec, H. (1988) *Mind Children: The Future of Robot and Human Intelligence*, Cambridge (Mass.), London, Harvard University Press.

Motwani, S., Sharma, T. & Gupta, A. (2021) »Ethics in Autonomous Vehicle Software: The Dilemmas«, *Computer*, Vol. 54, No. 8, S. 46–55.

Mullen, C., Tight, M., Whiteing, A. & Jopson, A. (2014) »Knowing their place on the roads: What would equality mean for walking and cycling?«, *Transportation Research Part A: Policy and Practice*, Vol. 61, S. 238–248.

Müller, J. F. & Gogoll, J. (2020) »Should Manual Driving be (Eventually) Outlawed?«, *Science and Engineering Ethics*, Vol. 26, S. 1549–1567.

Murray, C. J. (1994) »Quantifying the burden of disease: the technical basis for disability-adjusted life years«, *Bulletin of the World Health Organization*, Vol. 72, No. 3, S. 429–445.

Nagel, T. (1972) »War and Massacre«, *Philosophy & Public Affairs*, Vol. 1, No. 2, S. 123–144.

Nagel, T. (1979a) *Mortal Questions*, Cambridge, Cambridge University Press.

Nagel, T. (1987b [1979b]) »The Fragmentation of Value«, in Gowans, C. W. (Hg.) *Moral Dilemmas,* New York, Oxford University Press, S. 174–187.

Nagel, T. (1995 [1991]) *Equality and Partiality*, New York, Oxford University Press.

Nagel, T. (2013 [1979c]) »Moral Luck«, in Shafer-Landau, R. (Hg.) *Ethical Theory: An Anthology,* 2. Aufl., Malden, Wiley-Blackwell, S. 355–362.

Nassehi, A. (1997a) »Das Problem der Optionssteigerung. Überlegungen zur Risikokultur der Moderne«, *Berliner Journal für Soziologie*, Vol. 7, S. 21–36.

Nassehi, A. (1997b) »Risiko — Zeit — Gesellschaft Gefahren und Risiken der anderen Moderne«, in Hijikata, T. & Nassehi, A. (Hg.) *Riskante Strategien: Beiträge zur Soziologie des Risikos,* Opladen, Westdeutscher Verlag, S. 37–64.

Nassehi, A. (1997c) »Risikogesellschaft«, in Kneer, G., Nassehi, A. & Schroer, M. (Hg.) *Soziologische Gesellschaftsbegriffe. Konzepte moderner Zeitdiagnosen,* München, Fink.

Nassehi, A. (2019) *Muster: Theorie der digitalen Gesellschaft*, München, C.H. Beck.

National Highway Traffic Safety Administration (2008) »National Motor Vehicle Crash Causation Survey (NMVCCS)«, *U.S. Department of Transportation, National Technical Information Service*, 07.2008 [Online]. Verfügbar unter https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/811059 (Abgerufen am 16. März 2023).

National Highway Traffic Safety Administration (2013) »Preliminary Statement of Policy Concerning Automated Vehicles«, 2013 [Online]. Verfügbar unter https://www.nhtsa.gov/sites/nhtsa.gov/files/documents/automated_vehicles_policy.pdf (Abgerufen am 16. März 2023).

Nello-Deakin, S. (2019) »Is there such a thing as a ›fair‹ distribution of road space?«, *Journal of Urban Design*, Vol. 24, No. 5, S. 698–714.

Németh, B. (2022) »Route selection method with ethical considerations for automated vehicles under critical situations«, *IEEE 20th Jubilee World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, Poprad, Slovakia, S. 419–424.

Németh, B. (2023) »Coordinated Control Design for Ethical Maneuvering of Autonomous Vehicles«, *Energies*, Vol. 16, No. 10, S. 4254.

Neuhäuser, C. (2015) »Some Sceptical Remarks Regarding Robot Responsibility and a Way Forward«, in Misselhorn, C. (Hg.) *Collective Agency and Cooperation in Natural and Artificial Systems,* Cham, Springer International Publishing, S. 131–146.

Nida-Rümelin, J. (Hg.) (2005 [1996]) *Angewandte Ethik. Die Bereichsethiken und ihre theoretische Fundierung: Ein Handbuch*, 2. Aufl., Stuttgart, Alfred Kröner Verlag.

Nida-Rümelin, J. (1993) *Kritik des Konsequentialismus*, München, Oldenbourg.

Nida-Rümelin, J. (2005) »Ethik des Risikos«, in Nida-Rümelin, J. (Hg.) *Angewandte Ethik. Die Bereichsethiken und ihre theoretische Fundierung: Ein Handbuch,* 2. Aufl., Stuttgart, Alfred Kröner Verlag, S. 862–885.

Nida-Rümelin, J. (2006) »Eine Verteidigung von Freiheit und Gleichheit«, *Zeitschrift für Politik*, Vol. 53, No. 1, S. 3–25.

Nida-Rümelin, J. & Rechenauer, M. (2009) »Internationale Gerechtigkeit«, in Ferdowski, M. A. (Hg.) *Internationale Politik als Überlebensstrategie,* München, Bayerische Landeszentrale für politische Bildungsarbeit, S. 297–321.

Nida-Rümelin, J. & Schulenburg, J. (2013) »Risiko«, in Grunwald, A. (Hg.) *Handbuch Technikethik,* Stuttgart, J.B. Metzler, S. 18–22.

Nida-Rümelin, J., Schulenburg, J. & Rath, B. (2012) *Risikoethik*, Berlin, de Gruyter.

Nozick, R. (1974) *Anarchy, State and Utopia*, Oxford, Basil Blackwell.

Nussbaum, M. C. (1999) *Gerechtigkeit oder das gute Leben*, Frankfurt/Main, Suhrkamp Verlag.

Nussbaum, M. C. (2000) »The costs of tragedy: Some moral limits of cost-benefit analysis«, *The Journal of Legal Studies*, Vol. 29, S2, S. 1005–1036.

Nussbaum, M. C. (2006) *Frontiers of Justice: Disability, Nationality, Species Membership*, Cambridge, Harvard University Press.

Nussbaum, M. C. (2009) »Creating Capabilities: The Human Development Approach and Its Implementation«, *Hypatia*, Vol. 24, No. 3, S. 211–215.

Nyholm, S. (2018a) »Attributing Agency to Automated Systems: Reflections on Human-Robot Collaborations and Responsibility-Loci«, *Science and Engineering Ethics*, Vol. 24, No. 4, S. 1201–1219.

Nyholm, S. (2018b) »The ethics of crashes with self-driving cars: A roadmap, I«, *Philosophy Compass*, Vol. 13, No. 7, S. e12507.

Nyholm, S. (2018c) »The ethics of crashes with self-driving cars: A roadmap, II«, *Philosophy Compass*, Vol. 13, No. 7, S. e12506.

Nyholm, S. (2020a) *Humans and Robots: Ethics, Agency, and Anthropomorphism*, Lanham, Rowman & Littlefield Publishers.

Nyholm, S. (2020b) »In Evaluating Technological Risks, When and Why Should We Consult Our Emotions?«, *Science and engineering ethics*, Vol. 26, S. 1903–1912.

Nyholm, S. (2023a) »Ethical and Legal Issues Related to Autonomous Vehicles«, in Gordon, J.-S. (Hg.) *Future Law, Ethics, and Smart Technologies: The Future of Legal Education,* Brill, S. 190–204.

Nyholm, S. (2023b) »Minding the Gap(s): Different Kinds of Responsibility Gaps Related to Autonomous Vehicles and How to Fill Them«, in Fossa, F. & Cheli, F. (Hg.) *Connected and Automated Vehicles: Integrating Engineering and Ethics,* Cham, Springer, S. 1–18.

Nyholm, S. & Smids, J. (2016) »The Ethics of Accident-Algorithms for Self-Driving Cars: an Applied Trolley Problem?«, *Ethical Theory and Moral Practice*, Vol. 19, No. 5, S. 1275–1289.

Nyholm, S. & Smids, J. (2020) »Automated Cars Meet Human Drivers: Responsible Human-Robot Coordination and The Ethics of Mixed Traffic«, *Ethics and Information Technology*, Vol. 22, S. 335–344.

Ohnsman, A. (2024) »Waymo: Robotaxi-Durchbruch und Umsatzsprung«, *Forbes Media LLC* [Online]. Verfügbar unter https://www.forbes.at/artikel/waymo-robotaxi-durchbruch-und-umsatzsprung.html (Abgerufen am 11. August 2024).

Osório, A. & Pinto, A. (2019) »Information, uncertainty and the manipulability of artificial intelligence autonomous vehicles systems«, *International Journal of Human-Computer Studies*, Vol. 130, S. 40–46.

Othman, K. (2021) »Public acceptance and perception of autonomous vehicles: a comprehensive review«, *AI and Ethics*, Vol. 1, No. 3, S. 355–387.

Othman, K. (2023) »Understanding how moral decisions are affected by accidents of autonomous vehicles, prior knowledge, and perspective-taking: a continental analysis of a global survey«, *AI and Ethics*, S. 1–18.

Ott, K. (1998) »Ökonomische und moralische Risikoargumente in der Technikbewertung«, in Lenk, H. & Maring, M. (Hg.) *Technikethik und Wirtschaftsethik: Fragen der praktischen Philosophie,* Opladen, Leske + Budrich, S. 123–151.

Ott, K. (2005) »Technikethik«, in Nida-Rümelin, J. (Hg.) *Angewandte Ethik. Die Bereichsethiken und ihre theoretische Fundierung: Ein Handbuch,* 2. Aufl., Stuttgart, Alfred Kröner Verlag, S. 568–647.

Owens, J. M., Sandt, L., Habibovic, A., Rebolloso McCullough, S., Snyder, R., Wall Emerson, R., Varaiya, P., Combs, T., Feng, F., Yousuf, M. & Soriano, B. (2019) »Automated Vehicles and Vulnerable Road Users: Envisioning a Healthy, Safe and Equitable Future«, in Meyer, G. & Beiker, S. (Hg.) *Road Vehicle Automation 6: Lecture Notes in Mobility,* Cham, Springer International Publishing, S. 61–71.

Pagallo, U. (2022) »The Politics of Self-Driving Cars: Soft Ethics, Hard Law, Big Business, Social Norms«, in Jenkins, R., Černý, D. & Hríbek, T. (Hg.) *Autonomous Vehicle Ethics: The Trolley Problem and Beyond,* New York, Oxford University Press, S. 159–175.

Pan, S., Thornton, S. M. & Gerdes, J. C. (2016) »Prescriptive and proscriptive moral regulation for autonomous vehicles in approach and avoidance«, *IEEE International Symposium on Ethics in Engineering, Science and Technology (ETHICS)*, Vancouver, BC, Canada, S. 1–6.

Papadimitriou, Eleonora, Farah, H., van de Kaa, G., Santoni de Sio, F., Hagenzieker, Marjan, van Gelder, P., Papadimitriou, E. & Hagenzieker, M. (2022) »Towards common ethical and safe ›behaviour‹ standards for automated vehicles«, *Accident Analysis and Prevention*, Vol. 174, S. 106724.

Parfit, D. (2003) »Equality and Priority«, in Matravers, D. & Pike, J. (Hg.) *Debates in Contemporary Political Philosophy: An Anthology,* London, Routledge, S. 115–132.

Pastötter, B., Gleixner, S., Neuhauser, T. & Bäuml, K.-H. T. (2013) »To push or not to push? Affective influences on moral judgment depend on decision frame«, *Cognition*, Vol. 126, No. 3, S. 373–377.

Paulo, N. (2023) »The Trolley Problem in the Ethics of Autonomous Vehicles«, *The Philosophical Quarterly*, Vol. 73, No. 4, S. 1046–1066.

Paulo, N., Möck, L. A. & Kirchmair, L. (2023) »The Use and Abuse of Moral Preferences in the Ethics of Self-Driving Cars«, in Viciana, H., Gaitán, A. & Aguiar, F. (Hg.) *Experiments in Moral and Political Philosophy,* New York, Routledge, S. 290–309.

Pavone, M. (2015) »Autonomous Mobility-on-Demand Systems for Future Urban Mobility«, in Maurer, M., Gerdes, J. C., Lenz, B. & Winner, H. (Hg.) *Autonomes Fahren: Technische, rechtliche und gesellschaftliche Aspekte,* Berlin/Heidelberg, Springer Vieweg, S. 399–416.

Pereira, R. H. M., Schwanen, T. & Banister, D. (2017) »Distributive justice and equity in transportation«, *Transport Reviews*, Vol. 37, No. 2, S. 170–191.

Platon (2015 [ca. 375 v. Chr.]) *Der Staat (Politeia)*, Stuttgart, Reclam.

Pojman, L. P. (2016 [2005]) *Justice: An Anthology*, New York, Routledge.

Pölzler, T. (2021) »The Relativistic Car: Applying Metaethics to the Debate about Self-Driving Vehicles«, *Ethical Theory and Moral Practice*, Vol. 24, No. 3, S. 833–850.

Posner, R. A. (2004) *Catastrophe. Risk and Response*, Oxford, Oxford University Press.

Poszler, F., Geisslinger, M., Betz, J. & Lütge, C. (2023) »Applying ethical theories to the decision-making of self-driving vehicles: A systematic review and integration of the literature«, *Technology in Society*, Vol. 75, S. 102350.

Powers, T. M. (2006) »Prospects for a Kantian Machine«, *IEEE Intelligent Systems*, Vol. 21, No. 4, S. 46–51.

Prakken, H. (2017) »On the problem of making autonomous vehicles conform to traffic law«, *Artificial Intelligence and Law*, Vol. 25, No. 3, S. 341–363.

Quinn, W. S. (1989) »Actions, Intentions, and Consequences: The Doctrine of Double Effect«, *Philosophy and Public Affairs*, Vol. 18, No. 4, S. 334–351.

Ramirez, E. J. & LaBarge, S. (2018) »Real moral problems in the use of virtual reality«, *Ethics and Information Technology*, Vol. 20, No. 4, S. 249–263.

Rannenberg, K. (2015) »Erhebung und Nutzbarmachung zusätzlicher Daten – Möglichkeiten und Risiken«, in Maurer, M., Gerdes, J. C., Lenz, B. & Winner, H. (Hg.) *Autonomes Fahren: Technische, rechtliche und gesellschaftliche Aspekte,* Berlin/Heidelberg, Springer Vieweg, S. 515–538.

Raters, M.-L. (2016 [2013]) *Das moralische Dilemma: Antinomie der praktischen Vernunft?*, 2. Aufl., Freiburg/München, Verlag Karl Alber.

Rath, B. (2011) *Entscheidungstheorien der Risikoethik. Eine Diskussion etablierter Entscheidungstheorien und Grundzüge eines prozeduralen libertären Risikoethischen Kontraktualismus*, Marburg, Tectum Wissenschaftsverlag.

Rath, M., Krotz, F. & Karmasin, M. (Hg.) (2019) *Maschinenethik: Normative Grenzen autonomer Systeme*, Wiesbaden, Springer VS.

Raue, M., D'Ambrosio, L. A., Ward, C., Lee, C., Jacquillat, C. & Coughlin, J. F. (2019) »The Influence of Feelings While Driving Regular Cars on the Perception and Acceptance of Self-Driving Cars«, *Risk Analysis: An Official Publication of the Society for Risk Analysis*, Vol. 39, No. 2, S. 358–374.

Rawls, J. (1997 [1971]) *A Theory of Justice*, 22. Aufl., Cambridge (Mass.), Harvard University Press.

Rawls, J. (2005 [1993]) *Political Liberalism*, 2. Aufl., New York, Columbia University Press.

Raz, J. (1986) *The Morality of Freedom*, New York, Oxford University Press.

Reed, N., Leiman, T., Palade, P., Martens, M. & Kester, L. (2021) »Ethics of automated vehicles: breaking traffic rules for road safety«, *Ethics and Information Technology*, Vol. 23, No. 4, S. 777–789.

Regenbogen, A. & Meyer, U. (Hg.) (2013) *Wörterbuch der philosophischen Begriffe*, Hamburg, Felix Meiner Verlag.

Rehmann-Sutter, C. (1998) »Ethik«, in Rehmann-Sutter, C., Vatter, A. & Seiler, H. (Hg.) *Partizipative Risikopolitik,* Opladen und Wiesbaden, Westdeutscher Verlag, S. 29–166.

Reschka, A. (2015) »Sicherheitskonzept für autonome Fahrzeuge«, in Maurer, M., Gerdes, J. C., Lenz, B. & Winner, H. (Hg.) *Autonomes Fahren: Technische, rechtliche und gesellschaftliche Aspekte,* Berlin/Heidelberg, Springer Vieweg, S. 489–513.

Rippe, K. P. (2013) »Risiko, Ethik und die Frage des Zumutbaren«, *Zeitschrift für philosophische Forschung*, No. 4, S. 517–537.

Robinson, J., Smyth, J., Woodman, R. & Donzella, V. (2022) »Ethical considerations and moral implications of autonomous vehicles and unavoidable collisions«, *Theoretical Issues in Ergonomics Science*, Vol. 23, No. 4, S. 435–452.

Robinson, P., Sun, L., Furey, H., Jenkins, R., Phillips, C. R., Powers, T. M., Ritterson, R. S., Xie, Y., Casagrande, R. & Evans, N. G. (2021) »Modelling Ethical Algorithms in Autonomous Vehicles Using Crash Data«, *IEEE Transactions on Intelligent Transportation Systems*, Vol. 23, No. 7, S. 7775–7784.

Rodríguez-Alcázar, J., Bermejo-Luque, L. & Molina-Pérez, A. (2021) »Do Automated Vehicles Face Moral Dilemmas? A Plea for a Political Approach«, *Philosophy & Technology*, Vol. 34, S. 811–832.

Roeser, S. (2018) *Risk, technology, and moral emotions*, London, Routledge.

Roeser, S. (2020) »Risk, Technology, and Moral Emotions: Reply to Critics«, *Science and engineering ethics*, Vol. 26, No. 4, S. 1921–1934.

Rohrmann, B. & Renn, O. (2000) »Risk Perception Research: An Introduction«, in Renn, O. & Rohrmann, B. (Hg.) *Cross-Cultural Risk Perception: A Survey of Empirical Studies,* Dordrecht, Springer Science+Business Media, S. 11–53.

Ropohl, G. (1979) *Eine Systemtheorie der Technik*, München, Hauser.

Ropohl, G. (1996) *Ethik und Technikbewertung*, Frankfurt/Main, Suhrkamp.

Ropohl, G. (2017) »Verantwortung und Risiko«, in Heidbrink, L., Langbehn, C. & Loh, J. (Hg.) *Handbuch Verantwortung,* Wiesbaden, Springer VS, S. 887–908.

Ross, W. D. (2002 [1930]) *The Right and the Good*, 2. Aufl., Oxford, Oxford University Press.

Ross, W. D. (2000 [1939]) *Foundations of Ethics*, illustrierte Ausgabe, Oxford, Oxford University Press.

Rottenstreich, Y. & Hsee, C. K. (2001) »Money, Kisses, and Electric Shocks: On the Affective Psychology of Risk«, *Psychological Science*, Vol. 12, No. 3, S. 185–190.

Roy, A. (2016) »Autonomous Cars Don't Have a ›Trolley Problem‹ Problem«, *The Drive*, 19. Oktober [Online]. Verfügbar unter https://www.thedrive.com/tech/5620/autonomous-cars-dont-have-a-trolley-problem-problem (Abgerufen am 05. Juli 2023).

Rudschies, W. &. Kroher, T. (2024) »Autonomes Fahren: So fahren wir in Zukunft«, *ADAC Online*, 03. Mai [Online]. Verfügbar unter https://www.adac.de/rund-ums-fahrzeug/ausstattung-technik-zubehoer/autonomes-fahren/technik-vernetzung/aktuelle-technik/ (Abgerufen am 12. August 2024).

Ryan, M. (2020) »The Future of Transportation: Ethical, Legal, Social and Economic Impact pf Self-driving Vehicles in the Year 2025«, *Science and engineering ethics*, Vol. 26, S. 1185–1208.

Ryazanov, A. A., Wang, S. T., Nelkin, D. K., McKenzie, C. R. & Rickless, S. C. (2023) »Beyond killing one to save five: Sensitivity to ratio and probability in moral judgment«, *Journal of Experimental Social Psychology*, Vol. 108, S. 104499.

SAE On-Road Automated Vehicle Standards Committee (2014) »SAE Standard J 3016: Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems«, *SAE International*, 2014 [Online]. Verfügbar unter https://www.sae.org/news/2019/01/sae-updates-j3016-automated-driving-graphic (Abgerufen am 04. Februar 2023).

Sahlin, N.-E. & Persson, J. (1994) »Epistemic Risk: The Significance of Knowing What One Does Not Know«, in Brehmer, B. & Sahlin, N.-E. (Hg.) *Future Risks and Risk Management,* Dordrecht, Springer Science+Business Media, S. 37–62.

Samuel, S., Yahoodik, S., Yamani, Y., Valluru, K. & Fisher, D. L. (2020) »Ethical decision making behind the wheel – A driving simulator study«, *Transportation Research Interdisciplinary Perspectives*, Vol. 5, S. 100147.

Santoni de Sio, F. (2017) »Killing by Autonomous Vehicles and the Legal Doctrine of Necessity«, *Ethical Theory and Moral Practice*, Vol. 20, No. 2, S. 411–429.

Santoni de Sio, F. (2021) »The European Commission report on ethics of connected and automated vehicles and the future of ethics of transportation«, *Ethics and Information Technology*, Vol. 23, No. 4, S. 713–726.

Santoni de Sio, F. & Mecacci, G. (2021) »Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them«, *Philosophy & Technology*, Vol. 34, No. 4, S. 1057–1084.

Santoni de Sio, F. & van den Hoven, J. (2018) »Meaningful Human Control over Autonomous Systems: A Philosophical Account«, *Frontiers in Robotics and AI*, Vol. 5, S. 15.

Sartre, J.-P. (2014 [1946]) *Der Existentialismus ist ein Humanismus und andere philosophische Essays*, 7. Aufl., Reinbek bei Hamburg, Rowohlt Taschenbuch Verlag.

Savulescu, J., Gyngell, C. & Kahane, G. (2021) »Collective Reflective Equilibrium in Practice (CREP) and controversial novel technologies«, *Bioethics*, Vol. 35, No. 7, S. 652–663.

Scanlon, T. (1982) »Contractualism and Utilitarianism«, in Sen, A. & Williams, B. (Hg.) *Utilitarianism and beyond,* Cambridge, Cambridge University Press, S. 102–128.

Scanlon, T. (1998) *What We Owe to Each Other*, Cambridge, Harvard University Press.

Schäfer, P. (2018) »Der lange Kampf um Vision Zero«, *Springer Professional*, 23. Mai [Online]. Verfügbar unter https://www.springerprofessional.de/fahrzeugsicherheit/automatisiertes-fahren/der-lange-kampf-um-vision-zero/15771248 (Abgerufen am 12. März 2022).

Schäffner, V. (2018) »Caught Up in Ethical Dilemmas: An Adapted Consequentialist Perspective on Self-Driving Vehicles«, in Coeckelbergh, M., Loh, J., Funk, M., Seibt, J. & Nørskov, M. (Hg.) *Envisioning Robots in Society – Power, Politics, and Public Space: Proceedings of Robophilosophy 2018/TRANSOR 2018,* Amsterdam, IOS Press, S. 327–335.

Schäffner, V. (2020a) »Is Utilitarianism Entirely Useless for Self-Driving Car Ethics? A Critical Reflection on the Rationale for Rule Utilitarianism«, in Goecke, B. P. & Rosenthal-von der Pütten, Astrid Marieke (Hg.) *Artificial Intelligence: Reflections in Philosophy, Theology, and the Social Sciences,* Paderborn, Brill mentis, S. 173–187.

Schäffner, V. (2020b) »Wenn Ethik zum Programm wird: Eine risikoethische Analyse moralischer Dilemmata des autonomen Fahrens«, *Zeitschrift für Ethik und Moralphilosophie (ZEMO)*, Vol. 3, No. 1, S. 27–49.

Schäffner, V. (2021) »Between Real World and Thought Experiment: Framing Moral Decision-Making in Self-Driving Car Dilemmas«, *Humanistic Management Journal (HMAJ)*, Vol. 6, No. 2, S. 249–272.

Schäffner, V. (2022) »Die Algorithmisierung der Moral: Über die (Un-)Möglichkeit moralischer Maschinen und die Grenzen maschineller Moral«, in Endres, E.-M., Puzio, A. & Rutzmoser, C. (Hg.) *Menschsein in einer technisierten Welt: Interdisziplinäre Perspektiven auf den Menschen im Zeichen der digitalen Transformation,* Wiesbaden, Springer VS, S. 75–90.

Schäffner, V. (2024) »Crash dilemmas and the ethical design of self-driving vehicles: implications from metaethics and pragmatic road marks«, *AI and Ethics.*

Scheffler, S. (1985) »The Role of Consent in the Legitimation of Risky Activity«, in Gibson, M. (Hg.) *To Breathe Freely: Risk, Consent, and Air,* Totowa, Rowman & Littlefield, S. 75–88.

Schelsky, H. (1961) »Der Mensch in der wissenschaftlichen Zivilisation«, in Schelsky, H. (Hg.) *Der Mensch in der wissenschaftlichen Zivilisation,* Wiesbaden, VS Verlag für Sozialwissenschaften, S. 5–46.

Schmidt, H. (2021) »Selbstfahrende Autos müssen in den USA nicht alle Sicherheitsnormen erfüllen«, *Neue Zürcher Zeitung*, 15. Januar [Online]. Verfügbar unter https://www.nzz.ch/mobilitaet/auto-mobil/autonome-autos-duerfen-in-usa-weniger-crashsicher-sein-ld.1596584 (Abgerufen am 19. September 2023).

Schuessler, D. (2024) »The probability problems of the Moral Machine Experiment«, *AI and Ethics*, Vol. 4, S. 501–510.

Schumann, O., Buchholz, M. & Dietmayer, K. (2023) „ Efficient Path Planning in Large Unknown Environments with Switchable System Models for Automated Vehicles ", *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, S. 2466–2472.

Science Media Center Germany (2018) *Welcher Ethik sollten autonome Autos folgen?* [Online]. Verfügbar unter https://www.sciencemediacenter.de/alle-angebote/research-in-context/details/news/welcher-ethik-sollten-autonome-autos-folgen/ (Abgerufen am 28. März 2023).

Searle, J. (1980) »Minds, brains, and programs«, *The Behavioral and Brain Sciences*, Vol. 3, No. 3, S. 417–457.

Seeger, S. A. (2010) *Verantwortung. Tradition und Dekonstruktion*, Würzburg, Königshausen und Neumann.

Sen, A. (1976) »Welfare inequalities and Rawlsian axiomatics«, *Theory and Decision*, Vol. 7, S. 243–262.

Sen, A. (1980) »Equality of What?«, in McMurrin, S. M. (Hg.) *The Tanner Lectures on Human Values,* Cambridge, Cambridge University Press, S. 196–220.

Shariff, A., Bonnefon, J.-F. & Rahwan, I. (2017) »Psychological roadblocks to the adoption of self-driving vehicles«, *Nature Human Behaviour*, Vol. 1, No. 10, S. 694–696.

Shaw, D. M. & Schneble, C. O. (2021) »Advance Car-Crash Planning: Shared Decision Making between Humans and Autonomous Vehicles«, *Science and Engineering Ethics*, Vol. 27, No. 6, S. 75.

Shrader-Frechette, K. S. (1991) *Risk and Rationality: Philosophical Foundations for Populist Reforms*, Berkeley and Los Angeles, University of California Press.

Siegel, J. & Pappas, G. (2023) »Morals, ethics, and the technology capabilities and limitations of automated and self-driving vehicles«, *AI & Society*, Vol. 38, S. 213–226.

Simon, J. (2016) »Values in Design«, in Heesen, J. (Hg.) *Handbuch Medien- und Informationsethik,* Stuttgart, J.B. Metzler, S. 357–364.

Singer, P. (1977) »Utility and the Survival Lottery«, *Philosophy*, Vol. 52, No. 200, S. 218–222.

Sinnott-Armstrong, W. (1988) *Moral Dilemmas*, Oxford, Basil Blackwell.

Skorupinski, B. & Ott, K. (2000) *Technikfolgenabschätzung und Ethik: Eine Verhältnisbestimmung in Theorie und Praxis*, Zürich, vdf Hochschulverlag AG an der ETH Zürich.

Slovic, P. (1992) »Perceptions of risk: Reflections on the psychometric paradigm«, in Krimsky, S. & Golding, D. (Hg.) *Social Theories of Risk,* New York, Praeger Publisher, S. 117–152.

Slovic, P., Finucane, M. L., Peters, E. & MacGregor, D. G. (2007) »The affect heuristic«, *European Journal of Operational Research*, Vol. 177, No. 3, S. 1333–1352.

Smilansky, S. (2022) »Autonomous Vehicles and Normative Pluralism«, in Jenkins, R., Černý, D. & Hríbek, T. (Hg.) *Autonomous Vehicle Ethics: The Trolley Problem and Beyond,* New York, Oxford University Press, S. 114–129.

Smith, B. (2019) »Personality facets and ethics positions as directives for self-driving vehicles«, *Technology in Society*, Vol. 57, S. 115–124.

Smith, P. T. (2022) »Distributive Justice, Institutionalism, and Autonomous Vehicles«, in Jenkins, R., Černý, D. & Hríbek, T. (Hg.) *Autonomous Vehicle Ethics: The Trolley Problem and Beyond,* New York, Oxford University Press, S. 279–294.

Sparrow, R. (2007) »Killer robots«, *Journal of Applied Philosophy*, Vol. 24, No. 1, S. 62–77.

Sparrow, R. (2016) »Robots and Respect: Assessing the Case Against Autonomous Weapon Systems«, *Ethics & International Affairs*, Vol. 30, No. 1, S. 93–116.

Sparrow, R. & Howard, M. (2017) »When human beings are like drunk robots: Driverless vehicles, ethics, and the future of transport«, *Transportation Research Part C: Emerging Technologies*, Vol. 80, S. 206–215.

Spiekermann, S. (2015) *Ethical IT Innovation: A Value-Based System Design Approach*, Boca Raton, CRC Press.

Spiekermann, S. & Winkler, T. (2022) „ Value-Based Engineering With IEEE 7000«, *IEEE Technology and Society Magazine*, Vol. 41, No. 3, S. 71–80.

Starr, C. (1969) »Social Benefit versus Technological Risk: What is our society willing to pay for safety?«, *Science*, Vol. 165, No. 3899, S. 1232–1238.

Statistisches Bundesamt (2022) *Verkehrsunfälle*, Fachserie 8 Reihe 7 [Online]. Verfügbar unter https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Verkehrsunfaelle/Publikationen/_publikationen-verkehrsunfaelle.html#_hzn1vjya3 (Abgerufen am 05. Dezember 2023).

Statistisches Bundesamt (2023) *Fehlverhalten der Fahrer bei Unfällen mit Personenschaden* [Online]. Verfügbar unter https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Verkehrsunfaelle/Tabellen/fehlverhalten-fahrzeugfuehrer.html (Abgerufen am 04. Dezember 2023).

Statman, D. (1990) »The Debate over the So-called Reality of Moral Dilemmas«, *Philosophical Papers*, Vol. 19, No. 3, S. 191–211.

Statman, D. (1995) *Moral Dilemmas*, Amsterdam, Editions Rodopi B.V.

Steigleder, K. (2016a) »Climate risks, climate economics, and the foundations of rights-based risk ethics«, *Journal of Human Rights*, Vol. 15, No. 2, S. 251–271.

Steigleder, K. (2016b) »Risiko«, in Goppel, A., Mieth, C. & Neuhäuser, C. (Hg.) *Handbuch Gerechtigkeit,* Stuttgart, J.B. Metzler, S. 438–443.

Styron, W. (1980) *Sophie's Choice*, New York, Bantam Books.

Sundararajan, A. (2017) *The sharing economy: The end of employment and the rise of crowd-based capitalism*, Cambridge (Mass.), MIT Press.

Sunstein, C. R. (2002) *Risk and Reason. Safety, Law, and the Environment*, Cambridge, Cambridge University Press.

Sunstein, C. R. (2003) »Terrorism and Probability Neglect«, *The Journal of Risk and Uncertainty*, Vol. 26, 2–3, S. 121–136.

Sunstein, C. R. (2005) *Laws of Fear. Beyond the Precautionary Principle*, Cambridge, Cambridge University Press.

Sütfeld, L., Gast, R., König, P. & Pipa, G. (2017) »Using Virtual Reality to Assess Ethical Decisions in Road Traffic Scenarios: Applicability of Value-of-Life-Based Models and Influences of Time Pressure«, *Frontiers in Behavioral Neuroscience*, Vol. 11, S. 122.

Sütfeld, L., König, P. & Pipa, G. (2019) »Towards a Framework for Ethical Decision Making in Automated Vehicles«, *PsyArXiv preprint* [Online]. Verfügbar unter https://doi.org/10.31234/osf.io/4duca (Abgerufen am 10. Oktober 2023).

Talbot, B., Jenkins, R. & Purves, D. (2017) »When robots should do the wrong thing«, in Lin, P., Jenkins, R. & Abney, K. (Hg.) *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence,* Oxford, Oxford University Press, S. 258–273.

Tessman, L. (2015) *Moral Failure: On the Impossible Demands of Morality*, New York, Oxford University Press.

Thomas von Aquin (1933ff. [1916]) *Summa theologica. Die deutsche Thomas-Ausgabe: Vollständige ungekürzte Deutsch-Lateinisch Ausgabe*, Graz, Styria.

Thompson, D. F. (1980) »Moral responsibility of public officials: The problem of many hands«, *The American Political Science Review*, Vol. 74, No. 4, S. 905–916.

Thomsen, F. K. (2023) »Algorithmic indirect discrimination, fairness and harm«, *AI and Ethics*, S. 1–15.

Thomson, J. J. (1976) »Killing, letting die, and the trolley problem«, *The Monist*, Vol. 59, No. 2, S. 204–217.

Thomson, J. J. (1985a) »Imposing Risks«, in Gibson, M. (Hg.) *To Breathe Freely: Risk, Consent, and Air,* Totowa, Rowman & Littlefield, S. 124–140.

Thomson, J. J. (1985b) »The trolley problem«, *The Yale Law Journal*, Vol. 94, No. 6, S. 1395–1415.

Thomson, J. J. (1986a) *Rights, Restitution, & Risk: Essays in Moral Theory*, Cambridge (Mass.), Harvard University Press.

Thomson, J. J. (1986b) »Some questions about government regulation of behavior«, in Parent, W. (Hg.) *Rights, Restitution, & Risk: Essays in Moral Theory,* Cambridge (Mass.), Harvard University Press, S. 154–172.

Thomson, J. J. (1990) *The Realm of Rights*, Cambridge (Mass.), Harvard University Press.

Thomson, J. J. (2008) »Turning the trolley«, *Philosophy & Public Affairs*, Vol. 36, No. 4, S. 359–374.

Thornton, S. M., Pan, S., Erlien, S. M. & Gerdes, J. C. (2017) »Incorporating Ethical Considerations Into Automated Vehicle Control«, *IEEE Transactions on Intelligent Transportation Systems*, Vol. 18, No. 6, S. 1429–1439.

Totschnig, W. (2020) »Fully Autonomous AI«, *Science and Engineering Ethics*, Vol. 26, S. 2473–2485.

Trappl, R. (2015) »Robots' Ethical Systems: From Asimov's Laws to Principlism, from Assistive Robots to Self-Driving Cars«, in Trappl, R. (Hg.) *A Construction Manual for Robots' Ethical Systems: Requirements, Methods, Implementations,* Cham, Springer, S. 1–8.

Trappl, R. (2016) »Ethical Systems for Self-Driving Cars: An Introduction«, *Applied Artificial Intelligence*, Vol. 30, No. 8, S. 745–747.

Trigg, R. (1971) »Moral Conflict«, *Mind*, Vol. 80, No. 317, S. 41–55.

Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M. & Floridi, L. (2022) »The ethics of algorithms: key problems and solutions«, *AI & Society*, Vol. 37, S. 215–230.

TÜV-Verband e.V. (2024) »Sichere KI beim hochautomatisierten & autonomen Fahren. Unabhängige Drittprüfung von KI als wichtiger Beitrag zur Vision Zero« (Policy Sheet Europawahl 2024) [Online]. Verfügbar unter https://www.tuev-verband.de/policy-sheets/europawahl-autonomes-fahren (Abgerufen am 01. August 2024).

Vallor, S. (2016) *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*, New York, Oxford University Press.

Vallor, S. & Bekey, G. A. (2017) »Artificial Intelligence and the Ethics of Self-Learning Robots«, in Lin, P., Jenkins, R. & Abney, K. (Hg.) *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence,* Oxford, Oxford University Press, S. 338–353.

van de Poel, I, Royakkers, L & Zwart, SD (Hg.) (2015) *Moral Responsibility and the Problem of Many Hands*, New York, Routledge.

van de Poel, I. & Nihlén Fahlquist, J. (2013) »Risk and Responsibility«, in Roeser, S., Hillerbrand, R., Sandin, P. & Peterson, M. (Hg.) *Essentials of Risk Theory,* Dordrecht, Springer, S. 107–143.

van Fraassen, B. C. (1973) »Values and the Heart's Command«, *The Journal of Philosophy*, Vol. 70, No. 1, S. 5–19.

Véliz, C. (2021) »Moral zombies: why algorithms are not moral agents«, *AI & Society*, Vol. 36, S. 487–497.

Verband der Automobilindustrie (VDA) e.V. (Hg.) (2022a) *Deutsche Poleposition* [Online]. Verfügbar unter https://www.vda.de/de/themen/digitalisierung/autonomes-fahren (Abgerufen am 04. Januar 2022).

Verband der Automobilindustrie (VDA) e.V. (Hg.) (2022b) *Von Fahrerassistenzsystemen zum autonomen Fahren* [Online]. Verfügbar unter https://www.vda.de/de/themen/digitalisierung/automatisiertes-fahren (Abgerufen am 04. Januar 2022).

Verband der Automobilindustrie (VDA) e.V. (2023) »Assistenzsysteme im Nutzfahrzeug«, *VDA Online*, 2023 [Online]. Verfügbar unter https://www.vda.de/de/themen/automobilindustrie/nutzfahrzeuge/assistenzsysteme-im-nutzfahrzeug (Abgerufen am 05. Dezember 2023).

Wachenfeld, W. & Winner, H. (2015) »Lernen autonome Fahrzeuge?«, in Maurer, M., Gerdes, J. C., Lenz, B. & Winner, H. (Hg.) *Autonomes Fahren: Technische, rechtliche und gesellschaftliche Aspekte,* Berlin/Heidelberg, Springer Vieweg, S. 465–488.

Wallach, W. & Allen, C. (2009 [2008]) *Moral Machines: Teaching Robots Right From Wrong*, New York, Oxford University Press.

Wang, H., Huang, Y., Khajepour, A., Cao, D. & Lv, C. (2020) »Ethical Decision-Making Platform in Autonomous Vehicles With Lexicographic Optimization Based Model Predictive Controller«, *IEEE Transactions on Vehicular Technology*, Vol. 69, No. 8, S. 8164–8175.

Wang, H., Khajepour, A., Cao, D. & Liu, T. (2022) »Ethical Decision Making in Autonomous Vehicles: Challenges and Research Progress«, *IEEE Intelligent Transportation Systems Magazine*, Vol. 14, No. 1, S. 6–17.

Weber, K. & Zoglauer, T. (2019) »Maschinenethik und Technikethik«, in Bendel, O. (Hg.) *Handbuch Maschinenethik,* Wiesbaden, Springer VS, S. 145–163.

Weydner-Volkmann, S. (2021) »Technikvertrauen. Beiträge zur Technikfolgenabschätzung jenseits von Akzeptanz und Akzeptabilität?«, *Technikfolgenabschätzung – Theorie und Praxis*, Vol. 30, No. 2, S. 53–59.

Whetstone, J. T. (2001) »How Virtue Fits Within Business Ethics«, *Journal of Business Ethics*, Vol. 33, No. 2, S. 101–114.

Williams, B. (1987) »Ethical Consistency«, in Gowans, C. W. (Hg.) *Moral Dilemmas,* New York, Oxford University Press, S. 115–137.

Winkle, T. (2015) »Sicherheitspotenzial automatisierter Fahrzeuge: Erkenntnisse aus der Unfallforschung«, in Maurer, M., Gerdes, J. C., Lenz, B. & Winner, H. (Hg.) *Autonomes Fahren: Technische, rechtliche und gesellschaftliche Aspekte,* Berlin/Heidelberg, Springer Vieweg, S. 351–376.

Wintersberger, P., Prison, A.-K., Riener, A. & Hasirlioglu, S. (2017) »The experience of ethics: Evaluation of self harm risks in automated vehicles«, *IEEE Intelligent Vehicles Symposium (IV)*, Los Angeles, CA, USA, S. 385–391.

Woisetschläger, D. M. (2015) »Marktauswirkungen des automatisierten Fahrens«, in Maurer, M., Gerdes, J. C., Lenz, B. & Winner, H. (Hg.) *Autonomes Fahren: Technische, rechtliche und gesellschaftliche Aspekte,* Berlin/Heidelberg, Springer Vieweg, S. 709–732.

Wolkenstein, A. (2018) »What has the Trolley Dilemma ever done for us (and what will it do in the future)? On some recent debates about the ethics of self-driving cars«, *Ethics and Information Technology*, Vol. 20, No. 3, S. 163–173.

Wölm, E. (2019) »Warum mein Auto nie allein schuld sein wird«, in Rath, M., Krotz, F. & Karmasin, M. (Hg.) *Maschinenethik: Normative Grenzen autonomer Systeme,* Wiesbaden, Springer VS, S. 173–191.

Wolmar, C. (2018) *Driverless Cars: On a Road to Nowhere*, London, London Publishing Partnership.

Woollard, F. (2022) »The New Trolley Problem: Driverless Cars and Deontological Distinctions«, *Journal of Applied Philosophy*, Vol. 40, No. 1, S. 49–64.

Wu, S. S. (2015) »Product Liability Issues in the U.S. and Associated Risk Management«, in Maurer, M., Gerdes, J. C., Lenz, B. & Winner, H. (Hg.) *Autonomes Fahren: Technische, rechtliche und gesellschaftliche Aspekte,* Berlin/Heidelberg, Springer Vieweg, S. 575–592.

Wu, S. S. (2020) »Autonomous vehicles, trolley problems, and the law«, *Ethics and Information Technology*, Vol. 22, S. 1–13.

Zeckhauser, R. J. (1996) »The Economics of Catastrophes«, *Journal of Risk and Uncertainty*, Vol. 12, S. 113–140.

Zerilli, J., Knott, A., Maclaurin, J. & Gavaghan, C. (2019) »Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?«, *Philosophy & Technology*, Vol. 32, No. 4, S. 661–683.

Zhang, Y., Wu, J., Yu, F. & Xu, L. (2023) »Moral Judgments of Human vs. AI agents in Moral Dilemmas«, *Behavioral Sciences*, Vol. 13, No. 2, S. 181.

Zhao, L. & Li, W. (2020) „›Choose for No Choose‹—Random-Selecting Option for the Trolley Problem in Autonomous Driving«, in Zhang, J., Dresner, M., Zhang, R., Hua, G. & Shang, X. (Hg.) *LISS2019: Proceedings of the 9th International Conference on Logistics, Informatics and Service Sciences,* Singapore, Springer, S. 665–672.

Zweig, K. (2019) *Ein Algorithmus hat kein Taktgefühl: Wo künstliche Intelligenz sich irrt, warum uns das betrifft und was wir dagegen tun können*, München, Heyne Verlag.