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.